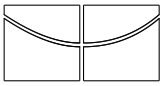

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Regressão Binomial Negativa Geograficamente Ponderada: Modelando Superdispersão Espacial

por

Thais Carvalho Valadares Rodrigues

Orientador: Prof. Dr. Alan Ricardo da Silva

Fevereiro de 2012

Thais Carvalho Valadares Rodrigues

Regressão Binomial Negativa Geograficamente Ponderada: Modelando Superdispersão Espacial

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Fevereiro de 2012

Ao meu admirável esposo.

Agradecimentos

Agradeço a Deus por estar ao meu lado sempre.

Ao meu esposo, por suas grandes contribuições nesta dissertação e pelo apoio incondicional, sempre com muita paciência, amor e dedicação.

Aos meus pais, pelo carinho com que cuidam de mim e por se dedicarem de forma excepcional à minha formação. Às minhas irmãs, por serem verdadeiras amigas. E a toda minha família, em especial, à minha vó, por ser uma pessoa admirável.

Ao meu orientador, Professor Alan, pela sua paciência e por estar sempre disponível a ajudar.

Aos amigos do mestrado, que tornaram o curso mais alegre e prazeroso.

E à CAPES, pelo apoio financeiro, que possibilitou minha dedicação exclusiva aos estudos.

Sumário

Lista de Figuras **4**

Lista de Tabelas **5**

Resumo **6**

Abstract **7**

Introdução **8**

1 Modelos Lineares Generalizados **11**

1.1 Introdução 11

1.2 Família Exponencial de Distribuições 11

1.3 Modelo Linear Generalizado 12

1.3.1 Casos especiais 13

1.3.1.1 Regressão clássica 14

1.3.1.2 Regressão de Poisson 15

1.3.1.3 Regressão Binomial Negativa 16

1.3.2 Algoritmos de estimação 18

1.3.2.1 Newton Raphson 19

1.3.2.2 Mínimos Quadrados Reponderados Iterativo 20

2 Regressão Geograficamente Ponderada **24**

2.1 Introdução 24

2.2 Indicadores de autocorrelação espacial 25

2.2.1 Matriz de proximidade espacial 25

2.2.2 Indicadores globais 27

2.2.3 Indicadores locais 29
2.2.4 Diagrama de espalhamento de Moran 29
2.3 Regressão espacial global 31
2.4 Regressão com regimes espaciais 32
2.5 Regressão Geograficamente Ponderada 32
2.5.1 Modelo RGP 33
2.5.2 Função de ponderação espacial 37
2.5.3 Determinação do parâmetro de suavização 39
2.5.4 Testes de não estacionariedade 41
2.6 Regressão de Poisson Geograficamente Ponderada 43
2.6.1 Modelo RPGP 43

3 Regressão Binomial Negativa Geograficamente Ponderada 49
3.1 Introdução 49
3.2 Modelo RBNGP 50
3.3 Modelo RBNGPg 55

4 Simulações e aplicações 57
4.1 Introdução 57
4.2 Simulação da RBNGP 58
4.2.1 Tamanho de amostra $n = 77$ 59
4.2.2 Tamanho de amostra $n = 504$ 67
4.3 Simulação da RPGP 74
4.4 Simulação da Regressão Global 80
4.5 Aplicação para a cidade de Tóquio 83
4.6 Aplicação para o estado do Espírito Santo 87

5 Conclusões e Trabalhos Futuros 95
5.1 Conclusões 95
5.2 Trabalhos Futuros 96

Referências Bibliográficas 98

Apêndice 101

Lista de Figuras

2.1 Exemplo de configuração espacial 26
2.2 Diagrama de Espalhamento de Moran 30
2.3 Superfície de β_k com a aproximação local no ponto i por um plano . . 34
2.4 Função de ponderação espacial 38

4.1 Diagrama de espalhamento de Moran da variável y 59
4.2 Análise exploratória da estacionariedade espacial da variável y 60
4.3 Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e b_2 62
4.4 Superfícies reais e estimadas, erro padrão e status da estimativa do parâmetro α 64
4.5 Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e b_2 65
4.6 Status das estimativas dos parâmetros b_0, b_1 e b_2 66
4.7 Amostra 68
4.8 Parâmetro de suavização b da RPGP 68
4.9 Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e b_2 69
4.10 Superfícies reais e estimadas, erro padrão e status da estimativa do parâmetro α 71
4.11 Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e b_2 72
4.12 Status das estimativas dos parâmetros b_0, b_1 e b_2 73
4.13 Parâmetro de suavização b da RBNGP que minimiza o CV 75
4.14 Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e α 77
4.15 Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e α 78
4.16 Status das estimativas dos parâmetros b_0, b_1 e α 79
4.17 Parâmetro de suavização b da RBNGP que minimiza o CV 81
4.18 Parâmetro de suavização b da RBNGP que minimiza o CV 85

4.19 Mapa da variável *Frota* e da variável *Indústrias* 88
4.20 Diagrama de espalhamento de Moran 89
4.21 Mapa de espalhamento de Moran e Mapa de Moran 95% 90
4.22 Estimativas pontuais, erros padrão e estatísticas *pseudo* t da RBNGP 92
4.23 Resíduo da RBNGP 93

Lista de Tabelas

1.1 Algoritmo de Newton Raphson 19
1.2 Algoritmo MQRI com matriz de informação observada 23

4.1 Comparação entre modelos 61
4.2 Comparação das médias das estimativas dos parâmetros 67
4.3 Comparação entre modelos 68
4.4 Comparação entre modelos 76
4.5 Sumário das estimativas dos parâmetros da RBNGP e da RBN 82
4.6 Sumário das estimativas dos parâmetros utilizando %gwglm (Chen e Yang, 2011), GWR3.0 e %gwnbr 84
4.7 Medianas dos erros padrão dos parâmetros utilizando %gwglm (Chen e Yang, 2011), GWR3.0 e %gwnbr 85
4.8 Comparação entre modelos 90

Resumo

A regressão global pressupõe que um modelo único é adequado para descrever todas as partes de uma região de estudo. No entanto, a força dos relacionamentos entre as variáveis pode não ser espacialmente constante. Além disso, os fatores envolvidos são geralmente tão complexos, que é difícil identificá-los na forma de variáveis explanatórias. Muitas vezes, ainda tem-se o problema de tamanho de amostra reduzido.

Neste contexto, surge a Regressão Geograficamente Ponderada (RGP), a fim de modelar dados espaciais não estacionários. Utilizando funções *kernel*, a RGP permite que os parâmetros do modelo variem espacialmente, produzindo superfícies não paramétricas das suas estimativas.

Considerando dados de contagem com superdispersão, o mais adequado é utilizar a distribuição Binomial Negativa. Por isso, o presente trabalho propõe dois modelos de Regressão Geograficamente Ponderada utilizando esta distribuição, a saber, RBNGPg e RBNGP. Estes modelos diferem-se na forma de estimação do parâmetro de superdispersão e, consequentemente, em termos de complexidade.

Neste trabalho, os modelos propostos são aplicados a 5 estudos de caso, envolvendo dados reais e simulados. Os resultados obtidos mostram a superioridade deles no ajuste de dados de contagem não estacionários e com superdispersão com respeito aos modelos concorrentes, a saber, regressão global - Poisson e Binomial Negativa - e Regressão de Poisson Geograficamente Ponderada. Além disso, verifica-se que estes modelos concorrentes são casos especiais do modelo mais robusto RBNGP.

Palavras chave: *Modelos Lineares Generalizados, distribuição de Poisson, distribuição Binomial Negativa, superdispersão, análise espacial e Regressão de Poisson Geograficamente Ponderada.*

Abstract

The global regression assumes that a single model is adequate to describe all parts of a study region. However, the strength of relationships between variables may not be spatially constant. In addition, the factors involved are often so complex that it is difficult to identify them in the form of explanatory variables. Many times, we also have the problem of small sample size.

In this context Geographically Weighted Regression (GWR) is introduced in order to model non-stationary spatial data. Using kernel functions, GWR allows the model parameters to vary spatially, producing non-parametric surfaces of their estimates.

To model count data with overdispersion, the most appropriate is to use the Negative Binomial distribution. Therefore, we propose two models of Geographically Weighted Regression using this distribution, namely GWNBRg and GWNBR. These models differ in the way the overdispersion parameter is estimated and, consequently, in terms of complexity.

In this dissertation, the proposed models are applied to 5 case studies involving real and simulated data. The results show their superiority in modelling non-stationary count data with overdispersion with respect to competing models, namely, global regression - Poisson and Negative Binomial - and Geographically Weighted Poisson Regression. Moreover, we demonstrate that these competing models are special cases of the more robust model GWNBR.

Key words: *Generalized Linear Models, Poisson distribution, Negative Binomial distribution, overdispersion, spatial analysis, Geographically Weighted Poisson Regression.*

Introdução

A Regressão Geograficamente Ponderada - RGP (ou do inglês, *Geographically Weighted Regression- GWR*) possibilita a modelagem espacial de dados não estacionários. Um processo espacial é dito estacionário se sua distribuição de probabilidade é invariante no espaço. Esta hipótese, que está presente no modelo de regressão global, é muito restritiva, pois somente em contextos muito particulares pode-se afirmar que um modelo único global representa adequadamente todas as partes da região de estudo. Processos sociais, por exemplo, são tipicamente não estacionários, pois a medida de uma relação depende em parte de onde esta medida é mensurada (Fotheringham et al., 2002).

Suponha, por exemplo, que uma corretora deseje modelar o preço de um imóvel no Distrito Federal em função da sua área útil, em m^2, e de uma variável indicadora, que assume 1 caso o imóvel tenha garagem. No entanto, o acréscimo no preço do imóvel decorrente do aumento de 1 m^2 em sua área, ou da presença de uma garagem, dependerá da localidade do mesmo. Portanto, a RGP é mais adequada para modelar este processo não estacionário.

No modelo RGP, a visualização das relações existentes entre a variável dependente e as variáveis independentes pode ser feita por meio de um mapa com as superfícies das estimativas locais dos parâmetros e dos erros padrão associados. Assim, a determinação de padrões espaciais e o entendimento de suas possíveis causas tornam-se facilitados.

A extensão do modelo de regressão global para o RGP é feita permitindo que os parâmetros $\boldsymbol{\beta}$ variem no espaço, conforme a equação

$$y_j = \beta_0(u_j, v_j) + \sum_k \beta_k(u_j, v_j) x_{jk} + \varepsilon_j \ , \tag{1}$$

onde (u_j, v_j) é a coordenada do j-ésimo ponto no espaço, $\beta_k(u_j, v_j)$ é a realização da função contínua $\beta_k(u, v)$ no j-ésimo ponto e ε_j são erros independentes e identicamente distribuídos N(0, σ^2) (Fotheringham et al., 2002).

Uma restrição limitante do modelo básico RGP dado pela Equação (1) é que a distribuição dos erros ε_j deve ser Gaussiana e, consequentemente, a variável dependente y também. No entanto, em muitas aplicações o termo dependente não é uma variável contínua capaz de assumir valores negativos e positivos, como por exemplo a quantidade de veículos utilizada no transporte rodoviário de cargas, assim como dados de contagem em geral. Neste caso, o modelo gaussiano é claramente inapropriado.

Distribuições mais adequadas para estas situações são a de Poisson e a Binomial Negativa. O modelo RGP para a Poisson foi desenvolvido por Nakaya et al. (2005), no entanto, nada ainda foi feito considerando a distribuição Binomial Negativa. Kobayashi e Lane (2007) e Hadayeghi et al. (2010) comentam em seus artigos a falta de um modelo de regressão geograficamente ponderada para a Binomial Negativa. A vantagem desta última é a possibilidade de modelar dados com superdispersão, visto que ela apresenta um parâmetro adicional α, chamado de parâmetro de heterogeneidade ou superdispersão (Hilbe, 2011). Além disso, a distribuição Binomial Negativa generaliza as distribuições Geométrica e Poisson, quando $\alpha = 1$ e $\alpha \to 0$, respectivamente, como será demonstrado no Capítulo 1.

A distribuição Binomial Negativa com α conhecido pertence à família exponencial de distribuições de Nelder e Wedderburn (1972), assim como a de Poisson. O algoritmo Score de Fisher dos Modelos Lineares Generalizados pode ser estendido para o caso espacial Binomial Negativo considerando a modificação no procedimento de Mínimos Quadrados Reponderados Iterativo - MQRI (ou do inglês, *Iteratively Reweighted Least Squares - IRLS*) proposta para a Regressão de Poisson Geograficamente Ponderada. O parâmetro α da Binomial Negativa, suposto conhecido pelo MQRI, pode ser estimado em uma subrotina pelo método da máxima verossimilhança utilizando o algoritmo de Newton Raphson (NR), conforme sugerido por Hilbe (2011).

Sendo assim, o intuito desta dissertação é estender o MQRI e o NR para o modelo de regressão geograficamente ponderada considerando a distribuição Binomial Negativa, a fim de modelar dados espaciais de contagem não estacionários e com superdispersão.

Esta dissertação está organizada da seguinte forma: no Capítulo 1 é feita uma revisão dos Modelos Lineares Generalizados (MLG) e dos algoritmos de estimação de Newton Raphson e de Mínimos Quadrados Reponderados Iterativo. O Capítulo 2 apresenta técnicas exploratórias de análise de dependência e não estacionariedade espacial. Além disso, o modelo de Regressão Geograficamente Ponderada está detalhado neste capítulo, tanto para a distribuição Normal quanto para a de Poisson. No Capítulo 3, dois modelos inéditos de Regressão Geograficamente Ponderada são propostos utilizando a distribuição Binomial Negativa, a saber, RBNGPg e RBNGP. O Capítulo 4 apresenta cinco estudos de caso, entre simulações e ajustes a dados reais. As principais conclusões e sugestões para trabalhos futuros encontram-se no Capítulo 5. No Apêndice, estão os principais códigos elaborados nesta dissertação.

Capítulo 1

Modelos Lineares Generalizados

1.1 Introdução

A distribuição Binomial Negativa pertence à família exponencial, que fornece a base probabilística para a classe dos Modelos Lineares Generalizados (MLG). Sendo assim, a estimação da sua média pode ser feita utilizando o algoritmo unificado desenvolvido por Nelder e Wedderburn (1972) para estes modelos. Com base nisso, este capítulo pretende descrever os componentes do MLG, apresentando com mais detalhes os modelos Normal, Poisson e Binomial Negativo. Além disso, são explorados os algoritmos de Newton Raphson e de Mínimos Quadrados Reponderados Iterativo, que são os principais métodos utilizados na estimação de modelos de contagem.

1.2 Família Exponencial de Distribuições

A família exponencial uniparamétrica engloba distribuições de probabilidade que podem ser escritas de acordo com a equação

$$f(y;\theta) = h(y)\exp\left[\eta(\theta)t(y) - b(\theta)\right], \tag{1.1}$$

cujo suporte não depende do parâmetro θ e $\eta(\theta)$, $b(\theta)$, $t(y)$ e $h(y)$ são funções que assumem valores reais (Cordeiro e Demétrio, 2010).

Um caso particular ocorre quando $\eta(\theta)$ e $t(y)$ são iguais à função identidade. Assim, a família exponencial apresenta-se na forma canônica e θ é chamado parâmetro

canônico, ou seja,

$$f(y;\theta) = h(y)\exp\left[\theta y - b(\theta)\right] . \tag{1.2}$$

A família exponencial uniparamétrica na forma canônica foi estendida por Nelder e Wedderburn (1972) por meio da inclusão do parâmetro de dispersão ϕ,

$$f(y;\theta,\phi) = \exp\left\{\phi^{-1}[\theta y - b(\theta)] + c(y,\phi)\right\} , \tag{1.3}$$

onde $b(.)$ e $c(.)$ são funções conhecidas. Com isso, distribuições biparamétricas foram incorporadas ao componente aleatório do modelo linear generalizado, conforme será visto na Seção 1.3. Quando ϕ é conhecido, a família exponencial de Nelder e Wedderburn é idêntica à família exponencial na forma canônica (1.2) (Cordeiro e Demétrio, 2010).

Muitas distribuições podem ser escritas conforme a Equação (1.3). Alguns exemplos são a Normal, Binomial, Binomial Negativa, Poisson e Gama. No entanto, para as distribuições biparamétricas, é necessário supor que um dos parâmetros é conhecido. Sendo assim, a família (1.3) é muito abrangente, engloba distribuições discretas e contínuas, com assimetria e com suportes restritos a intervalos do conjunto dos reais.

O valor esperado e a variância da variável aleatória Y pertencente à família (1.3) podem ser calculados a partir da função geradora de cumulantes $b(\theta)$, como

$$E(Y) = \mu = b'(\theta) \quad \text{e} \quad V(Y) = \phi b''(\theta) = \phi V(\mu) . \tag{1.4}$$

Outra propriedade importante das distribuições pertencentes à família (1.3) é que existe uma estatística suficiente minimal, dada por

$$T = \sum_{i=1}^{n} Y_i . \tag{1.5}$$

1.3 Modelo Linear Generalizado

Nelder e Wedderburn (1972), além de estenderem a família exponencial unipa-

ramétrica canônica, unificaram um conjunto de técnicas da modelagem estatística e nomearam de Modelos Lineares Generalizados - MLG (ou do inglês, *Generalized Linear Models - GLM*). O MLG é constituído de três componentes:

i) Componente aleatório: Conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ provenientes da família exponencial (1.3).

ii) Componente sistemático: Conjunto de parâmetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ e variáveis explicativas $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, constituindo o preditor linear $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, cujos elementos são dados por

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r \ . \tag{1.6}$$

iii) Função de ligação: Função monótona e diferenciável $g(.)$ que relaciona a média ao preditor linear, ou seja,

$$\eta_i = g(\mu_i) \ . \tag{1.7}$$

O modelo canônico é obtido escolhendo-se a função de ligação de forma que o preditor linear modele diretamente o parâmetro canônico, isto é, $g(\mu_i) = \theta_i = \eta_i$. A função de ligação canônica apresenta vantagens de simplificação no algoritmo de estimação e de interpretação dos parâmetros. No entanto, não há nenhuma razão, a priori, para que os efeitos sistemáticos do modelo tornem-se aditivos na escala dada por tais funções (Cordeiro e Demétrio, 2010).

1.3.1 Casos especiais

Os modelos Gaussiano, de Poisson e Binomial Negativo serão detalhados a seguir de acordo com a teoria de MLG.

1.3.1.1 Regressão clássica

O modelo clássico de regressão é o caso mais simples dos MLG. A função de densidade da distribuição Normal é dada por

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \tag{1.8}$$

A fim de classificá-la como membro da família exponencial de Nelder e Wedderburn, é necessário escrever (1.8) conforme sugerido na Equação (1.3). Assim, por meio de operações algébricas simples, obtém-se que

$$f(y;\mu,\sigma^2) = \exp\left\{(\sigma^2)^{-1}\left[y\mu - \frac{\mu^2}{2}\right] - \frac{1}{2}\left[\log(2\pi\sigma^2) + (\sigma^2)^{-1}y^2\right]\right\}. \tag{1.9}$$

Comparando (1.9) com (1.3), conclui-se que:

- $\phi = \sigma^2$,
- $\theta = \mu$,
- $b(\theta) = \frac{\theta^2}{2}$,
- $\mu = b'(\theta) = \theta$,
- $V(\mu) = b''(\theta) = 1$,
- $c(y,\phi) = -\frac{1}{2}\left[\log(2\pi\phi) + (\phi)^{-1}y^2\right]$.

Assim, os componentes do Modelo Linear Generalizado para a distribuição gaussiana são:

- Componente aleatório: $Y_i \sim N(\mu_i, \sigma^2)$.
- Componente sistemático: Dada a matriz do modelo $\mathbf{X}$ e o vetor de parâmetros $\boldsymbol{\beta}$, tem-se que $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- Função de ligação canônica: Função identidade, pois $\eta = g(\mu) = \theta = \mu$. Portanto, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

Sendo assim, chega-se ao modelo clássico de regressão linear:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \ . \tag{1.10}$$

Apesar da distribuição Normal ter uma ampla gama de aplicações, ela não é adequada para modelar dados de contagem, especialmente quando os mesmos assumem valores baixos.

1.3.1.2 Regressão de Poisson

A regressão de Poisson é o modelo básico para descrever dados de contagem. A função de probabilidade da distribuição de Poisson, parametrizada em termos da sua média μ, é dada por

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \ . \tag{1.11}$$

Reescrevendo (1.11) de acordo com a família (1.3), tem-se que

$$f(y;\mu) = \exp\left\{[y\log\mu - \mu] - \log(y!)\right\} \ . \tag{1.12}$$

Portanto, os parâmetros relevantes na teoria de MLG para a Poisson são:

- $\phi = 1$,
- $\theta = \log\mu$,
- $b(\theta) = \mu = \exp\{\theta\}$,
- $\mu = b'(\theta) = \exp\{\theta\}$,
- $V(\mu) = b''(\theta) = \exp\{\theta\} = \mu$,
- $c(y) = -\log(y!)$.

Com base no que foi desenvolvido, os componentes do Modelo Linear Generalizado da Poisson são:

- Componente aleatório: $Y_i \sim \text{Poisson}(\mu_i)$.

- Componente sistemático: Dada a matriz do modelo $\mathbf{X}$ e o vetor de parâmetros $\boldsymbol{\eta}$, tem-se que $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.

- Função de ligação canônica: $g(\mu) = \theta = \log(\mu)$. Sendo assim, $\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$.

Uma hipótese muito forte da regressão de Poisson é a equidispersão, isto é, a igualdade entre a média e a variância da variável aleatória Y. Dados de contagem que violam esta afirmação podem ser modelados utilizando a distribuição Binomial Negativa.

1.3.1.3 Regressão Binomial Negativa

A distribuição Binomial Negativa pode ser derivada de diferentes maneiras. Dentro do contexto dos MLG, ela é apresentada como o número de fracassos y antes da ocorrência do r-ésimo sucesso em uma sequência de ensaios de Bernoulli independentes e identicamente distribuídos. Parametrizada desta maneira, y e r devem ser inteiros positivos e a distribuição é chamada de Pascal. No entanto, não há nenhuma restrição matemática que impeça r de assumir qualquer valor positivo real, neste caso ela também é conhecida como distribuição de Polya. Neste trabalho, a trataremos simplesmente como distribuição Binomial Negativa.

A função densidade de probabilidade (fdp) da distribuição Binomial Negativa é dada por

$$f(y; p, r) = \binom{y + r - 1}{r - 1} p^r (1 - p)^y, \quad y \in \mathbb{Z}^+ \tag{1.13}$$

ou, em termos do parâmetro de superdispersão α,

$$f(y; p, \alpha) = \binom{y + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} p^{\frac{1}{\alpha}} (1 - p)^y \ , \tag{1.14}$$

visto que $\alpha = 1/r$.

Considerando o parâmetro r conhecido, é possível reescrever (1.13) de acordo com a família (1.3), como

$$f(y; p, r) = \exp\left\{ [y \log(1 - p) + r \log(p)] + \log\binom{y + r - 1}{r - 1} \right\} . \tag{1.15}$$

Assim, chega-se nas relações importantes do MLG para a Binomial Negativa:

- $\phi = 1$,
- $\theta = \log(1-p)$,
- $p = 1 - \exp(\theta)$,
- $b(\theta) = -r\log(p) = -r\log(1-\exp(\theta))$,
- $\mu = b'(\theta) = \frac{re^{\theta}}{1-e^{\theta}} = \frac{r(1-p)}{p}$,
- $V(\mu) = b''(\theta) = \frac{re^{\theta}(1-e^{\theta})+re^{2\theta}}{(1-e^{\theta})^2} = \frac{r(1-p)}{p^2} = \frac{\mu(\mu+r)}{r} = \mu + \alpha\mu^2$,
- $c(y) = \log\binom{y+r-1}{r-1}$.

Consequentemente, os componentes do modelo linear generalizado da Binomial Negativa são:

- Componente aleatório: $Y_i \sim \text{BN}(p_i, r)$.
- Componente sistemático: Dada a matriz do modelo $\mathbf{X}$ e o vetor de parâmetros $\boldsymbol{\beta}$, tem-se que $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- Função de ligação canônica: $g(\mu) = \theta = \log(1-p) = \log\left(\frac{\mu}{\mu+r}\right)$, pois $p = \frac{r}{\mu+r}$. Sendo assim, $\log\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+r}\right) = \mathbf{X}\boldsymbol{\beta}$.

Diferentemente da regressão clássica e da de Poisson, no modelo Binomial Negativo geralmente não se utiliza a função de ligação canônica. O modelo tradicional de regressão Binomial Negativa, denominado NB-2, utiliza a função de ligação logarítmica $g(\mu) = \theta = \log(\mu)$. Como o modelo Binomial Negativo surge, em geral, quando o de Poisson é insatisfatório, o uso da mesma função de ligação da regressão de Poisson permite uma comparação direta entre eles, tornando-se facilitada a avaliação do ganho na modelagem NB-2 (Hilbe, 2011).

Note que a função de variância da Binomial Negativa é dada por $V(\mu) = \mu + \alpha\mu^2$, onde $\alpha > 0$, então $V(\mu) > \mu$, possibilitando a modelagem de dados com superdispersão. Outra vantagem do modelo Binomial Negativo é que o mesmo engloba o modelo Poisson, visto que este último é o NB-2 com $r \to \infty$ e $\mu = \frac{rp}{1-p}$. A demonstração desta equivalência está apresentada a seguir:

Considerando a fdp da Binomial Negativa com p sendo a probabilidade de fracasso, tem-se que

$$f(y;p,r) = \frac{\Gamma(y+r)}{\Gamma(r)y!}p^y(1-p)^r \ . \tag{1.16}$$

Neste caso, $\mu = \frac{rp}{1-p}$ e, consequentemente, $p = \frac{\mu}{r+\mu}$. Reparametrizando (1.16) em termos de μ, obtém-se que

$$\begin{aligned} f(y;\mu,r) &= \frac{\Gamma(y+r)}{\Gamma(r)y!}\left(\frac{\mu}{r+\mu}\right)^y\left(1-\frac{\mu}{r+\mu}\right)^r \\ &= \frac{\mu^y}{y!}\frac{\Gamma(y+r)}{\Gamma(r)\,(r+\mu)^y}\frac{1}{\left(1+\frac{\mu}{r}\right)^r} \ . \end{aligned} \tag{1.17}$$

Fazendo $r \to \infty$, o segundo termo da Equação (1.17) tende a 1 e o terceiro termo tende a $e^{-\mu}$. Portanto,

$$\begin{aligned} \lim_{r\to\infty} f(y;\mu,r) &= \frac{e^{-\mu}\mu^y}{y!} \ , \\ \text{Poisson}(\mu) &= \lim_{r\to\infty} \text{BN}\left(\frac{\mu}{r+\mu}, r\right) \ . \end{aligned}$$

Sendo assim, a Binomial Negativa é uma alternativa robusta à Poisson, pois se aproxima da Poisson para r grande e tem variância maior do que ela para r pequeno. Além disso, para $r = 1$, a distribuição Binomial Negativa dada em (1.13) é equivalente à Geométrica, modelando o número de fracassos y antes da ocorrência do primeiro sucesso, isto é,

$$f(y;p) = p(1-p)^y, \quad y \in \mathbb{Z}^+ \ . \tag{1.18}$$

1.3.2 Algoritmos de estimação

Dois métodos de máxima verossimilhança são utilizados para estimar os modelos de contagem: o método de Newton Raphson (NR) e o método de Mínimos Quadrados Reponderados Iterativo (MQRI) baseado no escore de Fisher.

1.3.2.1 Newton Raphson

O método de máxima verossimilhança estima os parâmetros do modelo igualando a zero a derivada da função de log-verossimilhança com respeito aos parâmetros $\boldsymbol{\beta}$. O vetor resultante $\mathbf{U}$ é chamado de gradiente ou vetor escore. Se a função de verossimilhança é côncava, os estimadores de máxima verossimilhança são os parâmetros $\boldsymbol{\beta}$ que resolvem a equação (Hilbe, 2011)

$$\mathbf{U} = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \, , \tag{1.19}$$

onde $L(\boldsymbol{\beta})$ é a função de log-verossimilhança.

Quando não há uma solução analítica para o sistema de equações (1.19), o procedimento iterativo de Newton Raphson, baseado na aproximação de Taylor de primeira ordem para o gradiente, pode ser utilizado. A versão multivariada do NR fornece

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left(\mathbf{H}^{(m)}\right)^{-1} \mathbf{U}^{(m)} \, , \tag{1.20}$$

onde $\mathbf{H}$ é a matriz Hessiana, composta pelas derivadas segundas da log-verossimilhança,

$$\mathbf{H} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \, . \tag{1.21}$$

O algoritmo computacional de Newton Raphson está apresentado na Tabela 1.1. Nota-se que valores iniciais para o vetor de parâmetros devem ser fornecidos. Além disso, *tol* é a tolerância desejada no critério de parada, usualmente próxima de 10^{-6}.

Tabela 1.1: Algoritmo de Newton Raphson

Inicializar β
enquanto $(abs(\beta_n - \beta_o) > tol)$ {
$U = \partial L / \partial \beta$
$H = \partial^2 L / \partial \beta^2$
$\beta_o = \beta_n$
$\beta_n = \beta_o - H^{-1} U$
}

Fonte: Hilbe (2011)

1.3.2.2 Mínimos Quadrados Reponderados Iterativo

O método de Mínimos Quadrados Reponderados Iterativo é uma simplificação do método de máxima verossimilhança que é permitida devido a propriedades únicas da família exponencial, da qual os MLG são membros (Hilbe, 2011). Quando as derivadas parciais de segunda ordem da log-verossimilhança não são obtidas facilmente, mas é possível calcular seus valores esperados, a matriz de informação observada $\mathbf{H}$ pode ser substituída pela matriz de informação esperada de Fisher $\mathbf{I}$ (Cordeiro e Demétrio, 2010), ou seja,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left(\mathbf{I}^{(m)}\right)^{-1} \mathbf{U}^{(m)} \,, \tag{1.22}$$

onde a matriz de informação de Fisher $\mathbf{I}$ é dada por

$$\mathbf{I} = -E\left[\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] = \phi^{-1}\mathbf{X}'\mathbf{A}\mathbf{X} \,. \tag{1.23}$$

Por meio de algumas manipulações algébricas da Equação (1.22), detalhadas em Dobson (2002), tem-se que

$$\boldsymbol{\beta}^{(m+1)} = [\mathbf{X}'\mathbf{A}^{(m)}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{A}^{(m)}\mathbf{z}^{(m)} \,, \tag{1.24}$$

onde $\mathbf{z}$ é um vetor, chamado de variável dependente ajustada, cujos elementos são definidos por

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \,, \tag{1.25}$$

e $\mathbf{A}$ é uma matriz diagonal cujos elementos a_i são dados por

$$a_i = \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \,, \tag{1.26}$$

onde $V(\mu_i)$ é dado por (1.4).

A Equação (1.24) é válida para todos os MLG e mostra que solucionar as equações de máxima verossimilhança equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente ajustada $\mathbf{z}$ sobre a matriz $\mathbf{X}$ usando uma

matriz de pesos $\mathbf{A}$ que se modifica no processo iterativo (Cordeiro e Demétrio, 2010).

Note que, devido às propriedades assintóticas do estimador de máxima verossimilhança, a matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$ quando $n \to \infty$ é dada pelo inverso da matriz de informação de Fisher $\mathbf{I}$ (Equação (1.23)), portanto,

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \phi \left(\mathbf{X}' \widehat{\mathbf{A}} \mathbf{X} \right)^{-1} , \tag{1.27}$$

onde $\widehat{\mathbf{A}}$ é a matriz de pesos $\mathbf{A}$ avaliada em $\hat{\boldsymbol{\beta}}$ (Cordeiro e Demétrio, 2010).

No caso particular da regressão clássica, temos que $\mathbf{A}$ é a matriz identidade e a variável dependente ajustada $\mathbf{z}$ é o próprio $\mathbf{y}$. Sendo assim, é possível o cálculo exato da estimativa do vetor de parâmetros $\boldsymbol{\beta}$, sem a necessidade do processo iterativo, ou seja,

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} . \tag{1.28}$$

A matriz de variância e covariância da Equação (1.27) simplifica-se, no caso da regressão clássica, para a forma usual, dada por

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X} \right)^{-1} . \tag{1.29}$$

O critério de parada mais utilizado para o algoritmo MQRI é baseado na estatística desvio, Dev, proposta por Nelder e Wedderburn (1972). O desvio (ou do inglês, *deviance*) é definido por duas vezes a diferença entre a log-verossimilhança do modelo saturado e reduzido (ou corrente), isto é,

$$Dev = 2 \sum_{i=1}^{n} \{ L(y_i; y_i) - L(\mu_i; y_i) \} . \tag{1.30}$$

O desvio também é utilizado para comparar modelos encaixados. Admitindo-se uma combinação satisfatória do componente aleatório e da função de ligação, o desvio auxilia a determinação das variáveis explicativas do modelo, sendo aquele com menor desvio o mais indicado (Cordeiro e Demétrio, 2010).

O algoritmo MQRI, utilizando a matriz de informação esperada, pode ser resumido nos seguintes passos (Hilbe, 2011):

1. Inicializar $\boldsymbol{\mu}$, $g(\boldsymbol{\mu})$ e Dev_0;

2. Calcular os pesos $\mathbf{A}$ utilizando a Equação (1.26);

3. Calcular a variável dependente ajustada $\mathbf{z}$ por meio de (1.25);

4. Regredir $\mathbf{z}$ nas variáveis preditoras $\mathbf{X}$ utilizando a matriz de pesos $\mathbf{A}$ (Equação (1.24));

5. Calcular $\boldsymbol{\eta}$, lembrando que $\boldsymbol{\eta} = \boldsymbol{X\beta}$;

6. Obter $\boldsymbol{\mu}$ fazendo $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$;

7. Calcular o desvio Dev;

8. Se $Dev - Dev_0 > tol$, fazer $Dev_0 = Dev$ e voltar para o passo 2. Caso contrário, parar.

É importante ressaltar que quando a função de ligação utilizada é a canônica, o NR é idêntico ao MQRI. No entanto, para a regressão NB-2, na qual não se utiliza a ligação canônica, diferenças poderão surgir entre os erros padrão das estimativas dos dois métodos. Para contornar esse problema, faz-se uma modificação no MQRI para permitir que os erros padrão sejam calculados com a matriz de informação observada (Hilbe, 2011). Os detalhes teóricos que fundamentam este ajuste estão apresentados em Hardin e Hilbe (2001). Na prática, define-se uma nova matriz diagonal $\boldsymbol{A_0}$, cujos elementos a_{i0} apresentam um termo a mais do que (1.26),

$$a_{i0} = \frac{1}{V(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + (y_i - \mu_i) \frac{V(\mu_i) g''(\mu_i) + V'(\mu_i) g'(\mu_i)}{V(\mu_i)^2 g'(\mu_i)^3} \, . \tag{1.31}$$

O algoritmo MQRI com matriz de informação observada está apresentado na Tabela 1.2. A vantagem da utilização deste algoritmo é que os erros padrão das estimativas serão os mesmos do algoritmo de NR e, adicionalmente, as estatísticas de qualidade do ajuste e os resíduos dos MLG podem ser calculados com facilidade aproveitando-se as técnicas unificadas de modelagem desenvolvida para os MLG (Hilbe, 2011).

Tabela 1.2: Algoritmo MQRI com matriz de informação observada

$Dev_0 = 0$
$\mu = (y + media(y))/2$
$\eta = g(\mu)$
enquanto $(abs(difDev) > tol)$ {
$A = 1/(Vg'^2)$
$z = \eta + (y - \mu)g'$
$A_o = A + (y - \mu)(Vg'' + V'g')/(V^2g'^3)$
$\beta = [X'A_0X]^{-1}X'A_0z$
$\eta = X'\beta$
$\mu = g^{-1}(\eta)$
$difDev = Dev - Dev_0$
$Dev_0 = Dev$
}

Fonte: Hilbe (2011)

No caso da regressão Binomial Negativa, que apresenta um parâmetro adicional, Hilbe (2011) sugere estimar r utilizando o algoritmo de NR (Tabela 1.1), e estimar μ por meio do algoritmo MQRI com matriz de informação observada, Tabela 1.2.

Capítulo 2

Regressão Geograficamente Ponderada

2.1 Introdução

A modelagem de dados espaciais procura mensurar propriedades e relacionamentos entre variáveis levando-se em conta a localização espacial do fenômeno em estudo (Druck et al., 2004). A informação adicional da localização geográfica, presente nos dados espaciais, permite a construção de um modelo que incorpore a estrutura espacial dos dados. Portanto, dados espaciais, em geral, não são um conjunto de observações independentes. Tobler (1979) observa esse fato ao enunciar a Primeira Lei da Geografia, "todas as coisas são parecidas, mas coisas mais próximas se parecem mais que coisas mais distantes".

Este capítulo apresenta técnicas para modelagem de dados espaciais não estacionários, isto é, dados cuja distribuição de probabilidade varia no espaço. Inicialmente, mostra-se como pode ser feita uma análise exploratória para identificar dependência espacial nos dados. Em seguida, o modelo de regressão espacial global é brevemente descrito. Constatando-se que essa técnica é inapropriada para dados não estacionários, a Regressão Geograficamente Ponderada é apresentada como solução alternativa para tratar este problema. Sendo assim, a Regressão Geograficamente Ponderada para dados com distribuição Normal é detalhada, bem como tópicos sobre a escolha da função de ponderação, a determinação do parâmetro de suavização e testes de não estacionariedade espacial. O capítulo finaliza apresentando a Regressão de Poisson Geograficamente Ponderada.

2.2 Indicadores de autocorrelação espacial

Os indicadores de autocorrelação espacial são estatísticas construídas com o objetivo de caracterizar a dependência espacial dos dados. Esta caracterização pode ser resumida em um único índice para toda a região de estudo ou pode ser desagregada localmente dentro dessa região, sendo os indicadores globais e locais, respectivamente. A fim de descrever estes índices, é necessário compreender o conceito de matriz de proximidade espacial.

2.2.1 Matriz de proximidade espacial

A matriz de proximidade espacial, também conhecida por matriz $\mathbf{W}$, é uma ferramenta auxiliar utilizada no cálculo de indicadores de autocorrelação espacial. Seu objetivo é representar, quantitativamente, a estrutura espacial entre as áreas da região de estudo. Sendo assim, dado um conjunto de n áreas, $A_1, \ldots, A_n$, os elementos w_{ij} da matriz $\mathbf{W}$, cuja dimensão é $n \times n$, representam alguma medida de proximidade entre as áreas A_i e A_j (Assunção, 2003). Sendo que, por definição, a diagonal de $\mathbf{W}$ é nula, isto é, $w_{ii} = 0$ para $i = 1, \ldots, n$.

A escolha dessa medida de proximidade é subjetiva e depende tanto do fenômeno em estudo quanto da familiaridade do analista com o assunto. Algumas possibilidades apresentadas por Assunção (2003) estão descritas a seguir:

1. $w_{ij} = 1$, se A_i faz fronteira com A_j, e $w_{ij} = 0$ caso contrário;
2. $w_{ij} = 1$, se o centróide (ou centro político) de A_i está a uma distância menor do que k quilômetros de A_j, e $w_{ij} = 0$ caso contrário;
3. $w_{ij} = 1/(1 + d_{ij})$, onde d_{ij} é a distância entre os centróides das áreas A_i e A_j;
4. $w_{ij} = 1/(1 + t_{ij})$, onde t_{ij} é o tempo necessário para ir de A_i para A_j.
5. $w_{ij} = q_{ij}$, onde q_{ij} é a quantidade de trocas comerciais entre as áreas A_i e A_j (Silva, 2006).

Em geral, trabalha-se com a matriz $\mathbf{W}$ padronizada ($\mathbf{W}_p$), na qual cada elemento w_{ij} é dividido pela soma dos pesos da linha de $\mathbf{W}$ correspondente.

A seguir, tem-se um exemplo de construção de $\mathbf{W}$ e $\mathbf{W}_p$ utilizando a matriz binária do item 1. A configuração espacial utilizada está ilustrada na Figura 2.1.

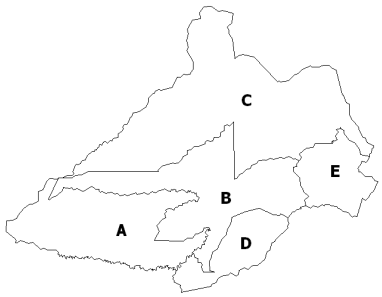

Figura 2.1: Exemplo de configuração espacial

$$\mathbf{W} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array} \qquad \mathbf{W}_p = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ 0.33 & 0.33 & 0 & 0 & 0.33 \\ 0 & 0.5 & 0.5 & 0 & 0 \end{pmatrix} \end{array}$$

É importante ressaltar que a dependência espacial não está restrita somente à dependência geográfica. Da mesma forma, quando nos referimos a "espaço" não estamos restritos ao espaço geográfico. No contexto da regressão espacial, o "espaço" apresenta um conceito mais amplo, e está associado ao tipo de relacionamento e dependência entre as variáveis (Silva, 2006). Por exemplo, no contexto econômico, podemos explicar a produção das indústrias de bens duráveis (automobilística, material elétrico, eletroeletrônicos e outras) utilizando a matriz insumo-produto apresentada no item 5, visto que a dependência entre elas é melhor caracterizada pela quantidade de trocas comerciais do que pela proximidade geográfica.

A matriz $\mathbf{W}$ apresentada no item 1 considera os vizinhos de primeira ordem, $\mathbf{W}^{(1)}$, ou simplesmente $\mathbf{W}$. É possível generalizar este conceito para vizinhos de ordem k (para $k = 2$, por exemplo, tem-se os vizinhos dos vizinhos), a matriz $\mathbf{W}$ correspondente é denotada por $\mathbf{W}^{(k)}$.

2.2.2 Indicadores globais

As estatísticas globais de autocorrelação espacial são úteis na análise exploratória dos dados. O índice mais utilizado é o I de Moran (Moran, 1950), apresentado na Equação (2.1), onde n é o número de áreas, y_i é o valor do atributo na área i e w_{ij} são os elementos da matriz de proximidade espacial $\mathbf{W}$:

$$I = \frac{n}{\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ . \tag{2.1}$$

Caso haja interesse na autocorrelação entre vizinhos de ordem superior a um, basta substituir w_{ij} na Equação (2.1) pelos elementos $w_{ij}^{(k)}$ da matriz $W^{(k)}$.

O índice de Moran está restrito ao intervalo $[-1, 1]$. O valor $I = 0$ indica ausência de autocorrelação entre as observações (considerando a matriz $\mathbf{W}$ utilizada), $I = 1$ é autocorrelação positiva máxima e $I = -1$ representa autocorrelação negativa máxima. Nota-se, por meio da Equação (2.1), que o índice de Moran é uma adaptação do coeficiente de correlação de Pearson para dados espaciais de uma mesma variável aleatória.

Outro índice global bastante utilizado é o C de Geary (Geary, 1954), dado por

$$C = \frac{n-1}{2\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ . \tag{2.2}$$

O índice C de Geary é mais indicado quando a quantidade de vizinhanças é pequena. O intervalo de variação deste índice é de 0 a 2, sendo $C = 1$ ausência de autocorrelação espacial (novamente, com referência a matriz $\mathbf{W}$ utilizada), $C = 0$ autocorrelação positiva máxima e $C = 2$ autocorrelação negativa máxima.

A validade estatística dos índices apresentados pode ser testada por meio de um teste de aleatorização (Druck et al., 2004). Nesse caso, a hipótese nula (H_0) é a independência espacial. Sendo assim, sob H_0, constrói-se a distribuição empírica do estimador gerando-se m permutações aleatórias dos valores dos atributos nas áreas da região de estudo e calcula-se o valor do índice para cada arranjo espacial obtido. Ordenando os índices de forma decrescente, considere que o posto do índice observado na amostra original seja p. Então, o p-valor do teste é obtido pela razão $p/(m+1)$ (Hope, 1968). Por não fazer pressupostos a respeito da distribuição de probabilidade

dos índices, este é o teste mais utilizado.

A significância estatística dos índices I de Moran e C de Geary também pode ser avaliada de forma paramétrica, visto que a distribuição assintótica desses índices é Normal. Os momentos de I e de C (Cliff e Ord, 1981) são:

$$E(I) = -\frac{1}{n-1} \,, \tag{2.3}$$

$$Var(I) = \frac{n^2 S_1 - nS_2 + 3S^2}{(n+1)(n-1)S^2} - \left(-\frac{1}{n-1}\right)^2 \,, \tag{2.4}$$

$$E(C) = 1 \,, \tag{2.5}$$

$$Var(C) = \frac{(2S_1 + S_2)(n-1) - 4S^2}{2(n+1)S^2} \,, \tag{2.6}$$

onde:

$$S = \sum_i \sum_{j \neq i} w_{ij} \,,$$

$$S_1 = 0.5 \sum_i \sum_{j \neq i} (w_{ij} + w_{ji})^2 \,,$$

$$S_2 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji} \right)^2 \,.$$

Portanto, tem-se que:

$$\frac{I - E(I)}{\sqrt{Var(I)}} \sim N(0,1) \quad \text{e} \quad \frac{C - E(C)}{\sqrt{Var(C)}} \sim N(0,1) \,, \text{ quando } n \to \infty \,.$$

É importante ressaltar a importância da escolha adequada da matriz de proximidade espacial, visto que os índices de autocorrelação espacial dependem diretamente da matriz $\mathbf{W}$. Uma escolha inapropriada de $\mathbf{W}$, por exemplo, pode levar à falsa impressão de ausência de autocorrelação espacial.

Os índices I de Moran e C de Geary consideram a hipótese de estacionariedade de segunda ordem (média e variância constantes). Quando os dados apresentarem não-estacionariedade (testes de estacionariedade serão discutidos na Subseção 2.5.4), é mais indicado utilizar os índices locais de autocorrelação.

2.2.3 Indicadores locais

O índice global enfatiza similaridades, pressupondo que todas as partes das regiões de estudo podem ser bem representadas por um valor único. No entanto, a presença de peculiaridades locais nos fazem questionar a validade dessa afirmação. Conforme apresentado no paradoxo de Simpson (Simpson, 1951), resultados opostos podem ser obtidos quando os dados são analisados conjuntamente e separadamente.

Considerando que diferentes formas de autocorrelação espacial podem existir em um conjunto de dados, Anselin (1995) elaborou os índices locais (ou do inglês, *Local Indicators of Spatial Association* - LISA), que são desagregações espaciais das estatísticas globais. Ao invés de similaridades, as estatísticas locais buscam por diferenças regionais e, por serem um conjunto de medidas, é possível mapeá-las (Fotheringham et al., 2002).

Os índices locais de Moran e de Geary são descritos por

$$I_i = \frac{n \times z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2} \ , \tag{2.7}$$

onde $z_j = y_j - \bar{y}$, e

$$C_i = \frac{\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \ . \tag{2.8}$$

A significância estatística desses índices pode ser calculada com testes de pseudo-significância da mesma forma descrita anteriormente para os índices globais (Druck et al., 2004). A presença de áreas com índices locais significativos é um indício de não estacionariedade. Assim, é útil gerar um mapa com as regiões que apresentam correlação local significativa, denominado mapa de indicadores locais (ou do inglês, *LISA map*).

2.2.4 Diagrama de espalhamento de Moran

O diagrama de espalhamento de Moran (ou do inglês, *Moran Scatterplot*) proposto por Anselin (1996) é uma forma gráfica de visualizar a dependência espacial. O objetivo é comparar o valor do atributo na área A_i com a média dos valores dos

atributos nas áreas próximas a A_i. Sendo assim, o eixo das abscissas apresenta o valor normalizado do atributo, ou seja, $\mathbf{z} = (\mathbf{y} - \bar{\mathbf{y}})/\mathbf{s}_y$, e o eixo das ordenadas contém o valor normalizado da média dos respectivos vizinhos, $\mathbf{Wz} = \mathbf{W}(\mathbf{y} - \bar{\mathbf{y}})/\mathbf{s}_y$.

A Figura 2.2 apresenta um exemplo do diagrama. Nota-se que o gráfico está dividido em quatro quadrantes, Q_1, Q_2, Q_3 e Q_4, chamados de alto-alto, baixo-alto, baixo-baixo e alto-baixo, respectivamente. O quadrante Q_1, por exemplo, contém os pontos cujo valor do atributo é alto e a média dos seus vizinhos também é alta, daí o nome alto-alto. Sendo assim, os pontos pertencentes aos quadrantes Q_1 e Q_3 indicam associação espacial positiva e os dos quadrantes Q_2 e Q_4 associação espacial negativa.

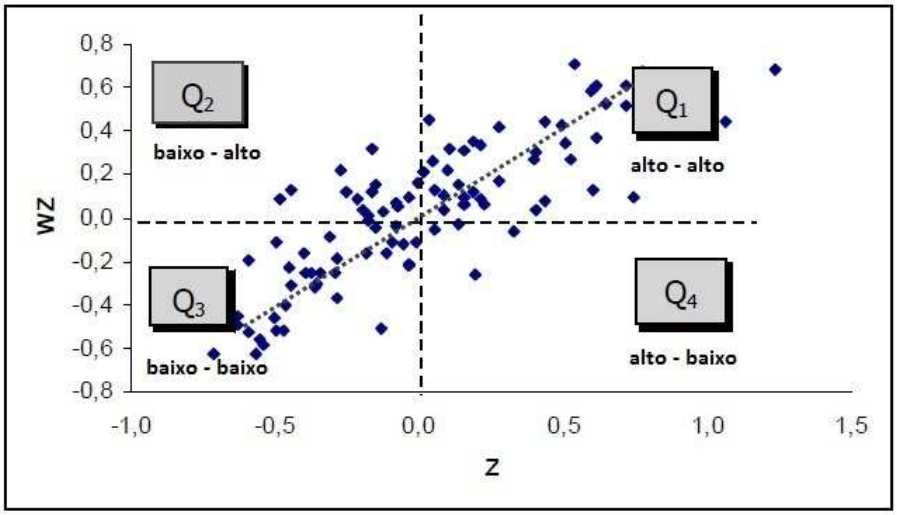

Figura 2.2: Diagrama de Espalhamento de Moran
Fonte: Druck et al. (2004) com modificações

O índice de Moran, apresentado na Equação (2.1), tem sua forma matricial descrita por

$$I = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{W}\mathbf{z} \; . \tag{2.9}$$

Nota-se, a partir da Equação (2.9), que o índice de Moran é o coeficiente angular da regressão linear de $\mathbf{Wz}$ em $\mathbf{z}$, ou seja, da reta de regressão do diagrama de dispersão de Moran (Druck et al., 2004).

O mapa de espalhamento de Moran (ou do inglês, *Box Map*) é a visualização georreferenciada do diagrama de dispersão de Moran. As áreas da região de estudo são pintadas de quatro cores, representando os quatro quadrantes.

A combinação do mapa de espalhamento de Moran com o mapa de indicadores locais dá origem ao mapa de Moran (ou do inglês, *Moran Map*). Seu intuito é indicar quais classificações do mapa de espalhamento de Moran (alto-alto, baixo-baixo, alto-baixo e baixo-alto) são significativas de acordo com a significância dos índices locais. Portanto, assim como o mapa de indicadores locais, cores no mapa de Moran também são indícios de não estacionariedade nos dados.

2.3 Regressão espacial global

A regressão espacial global é a classe de modelos de regressão espacial mais simples, pois supõe-se que é possível capturar a estrutura de correlação espacial dos dados em um ou, no máximo, dois parâmetros. Ela é indicada quando os dados apresentam estacionariedade e, consequentemente, um padrão único global é adequado para modelar o fenômeno.

A forma geral de um modelo espacial autoregressivo global (Anselin, 1988) é

$$\begin{aligned} \mathbf{y} &= \rho \boldsymbol{W_1 y} + \boldsymbol{X\beta} + \boldsymbol{u} \ , \\ \mathbf{u} &= \lambda \boldsymbol{W_2 u} + \boldsymbol{\varepsilon} \ , \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2 \mathbf{I_n}) \ . \end{aligned} \tag{2.10}$$

onde ρ e λ são os parâmetros que captam a dependência espacial na variável dependente $\mathbf{y}$ e no erro aleatório $\mathbf{u}$, respectivamente, e $\mathbf{W_1}$ e $\mathbf{W_2}$ são as matrizes de proximidade espacial de $\mathbf{y}$ e $\mathbf{u}$, respectivamente, sendo possível adotar $\mathbf{W_1} = \mathbf{W_2}$.

Apesar da forma geral (2.10) apresentar dois parâmetros espaciais, usualmente trabalha-se com $\mathbf{W_2} = 0$ (modelo espacial autoregressivo, ou do inglês, *Spatial Autoregressive Model* - SAR) ou $\mathbf{W_1} = 0$ (modelo com erro espacial autorregressivo, ou do inglês, *Spatial Error Model* - SEM). Mais detalhes sobre os modelos de regressão espacial global podem ser encontrados em Anselin (1988).

Quando o padrão de autocorrelação espacial variar na região de estudo, o modelo de regressão espacial global não será capaz de representar adequadamente a dependência espacial dos dados. Nestas situações, modelos de regressão espacial local, que permitem que os parâmetros variem no espaço, são mais recomendados.

2.4 Regressão com regimes espaciais

A regressão com regimes espaciais é uma regressão espacial local discreta. A partir da constatação da existência de regimes espaciais diferentes na região, este modelo propõe identificá-los (por exemplo, por meio do mapa de Moran) e realizar regressões em separado para cada sub-região. Utilizando variáveis indicadoras (ind) e supondo j regimes espaciais diferentes, tem-se que

$$\mathbf{y}_{ind} = \boldsymbol{X}_{ind}\boldsymbol{\beta}_{ind} + \boldsymbol{\varepsilon}_{ind},\ ind = 1, \ldots, j\ . \tag{2.11}$$

Apesar de cada região possuir seus próprios coeficientes, a estimação é feita em conjunto utilizando todas as observações disponíveis (Druck et al., 2004). Uma desvantagem dessa técnica é a dificuldade de pré-definir os regimes espaciais, isto é, discretizar o espaço em grupos supostamente homogêneos. Além disso, discontinuidades abruptas na estimação ocorrerão nos limites estabelecidos. Por fim, é possível que em algumas sub-regiões hajam poucas unidades amostrais, inviabilizando o ajuste.

Os efeitos espaciais também podem ser modelados de forma contínua, entendendo que o processo espacial modelado é, de fato, contínuo. Neste caso, temos a Regressão Geograficamente Ponderada (RGP).

2.5 Regressão Geograficamente Ponderada

A Regressão Geograficamente Ponderada auxilia a análise de dados espaciais não estacionários. O modelo permite que seus parâmetros variem espacialmente, sem limitar a forma dessa variação. A idéia da RGP é realizar um ajuste local para cada ponto da região de estudo com base nas observações mais próximas. Assim, cria-se uma função contínua $\beta_k(u_i, v_i)$ para cada parâmetro, onde (u_i, v_i) são as coordenadas espaciais do i-ésimo ponto. O objetivo da RGP é fornecer estimativas não paramétricas destas superfícies contínuas utilizando a função $kernel$.

Nesta seção será apresentado o modelo clássico de Regressão Geograficamente Ponderada. A abordagem Bayesiana da RGP não é objeto de estudo dessa dissertação. Detalhes desse assunto podem ser encontrados em LeSage (2001).

2.5.1 Modelo RGP

O modelo RGP (Fotheringham et al., 2002) está apresentado a seguir:

$$y_j = \beta_0(u_j, v_j) + \sum_k \beta_k(u_j, v_j) x_{jk} + \varepsilon_j \ , \tag{2.12}$$
$$\varepsilon_j \sim N(0, \sigma^2) \ .$$

Note que os pressupostos do modelo de regressão clássica (erros Normais, homocedásticos e não correlacionados) permanecem. No entanto, ao permitir variação espacial para os parâmetros, os problemas de autocorrelação e heterocedasticidade são reduzidos. A limitação ainda persistente é a normalidade, logo este modelo ainda não é o mais adequado para tratar dados espaciais de contagem, por exemplo.

É interessante observar que a regressão clássica (Equação 1.10) é um caso especial da Regressão Geograficamente Ponderada (Equação 2.12). Esta simplificação ocorre quando não há variação espacial nos parâmetros.

A forma matricial da Equação (2.12) é dada por

$$\mathbf{y} = (\boldsymbol{\beta} \otimes \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon} \ , \tag{2.13}$$

onde $\otimes$ é o operador que denota a multiplicação elemento a elemento. Considerando que o tamanho da amostra observada é n e o número de variáveis explicativas é k, tem-se que $\mathbf{X}$ é a matriz do modelo com dimensão $(n \times k+1)$, $\mathbf{1}$ é um vetor de 1's de dimensão $k+1$ e $\boldsymbol{\beta}$ é uma matriz $(n \times k+1)$, cuja linha j contém a estimativa dos $(k+1)$ parâmetros para a amostra j, ou seja,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \dots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \dots & \beta_k(u_2, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \dots & \beta_k(u_n, v_n) \end{bmatrix} . \tag{2.14}$$

Considerando o modelo apresentado em (2.12), a função de log-verossimilhança é

dada por

$$L(\boldsymbol{\beta}(u,v)|D) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j - \beta_0(u_j,v_j) - \sum_k \beta_k(u_j,v_j)x_{jk}\right)^2, \qquad (2.15)$$

onde D representa os dados $\{x_{jk}\}$, $\{y_j\}$ e $\{(u_j, v_j)\}$.

Como $\beta_k(u,v)$ são funções arbitrárias, a estimativa de máxima verossimilhança da Equação (2.15) não é única, sendo possível, inclusive, ajustes completamente diferentes, mas que conduzem à soma de quadrados dos erros igual a zero. Fotheringham et al. (2002) apresentam exemplos desses casos. Uma possibilidade para resolver este problema seria adotar uma forma funcional para $\beta_k(u,v)$, o que vai de encontro com a proposta da RGP. A solução encontrada foi utilizar a verossimilhança local e encontrar a estimativa para $\boldsymbol{\beta}(u,v)$ de forma local, e não global como (2.15).

Considere, então, um *ponto arbitrário* no espaço (u_i, v_i) no qual se deseja a estimativa dos parâmetros. Pelo fato de i ser um ponto arbitrário, não necessariamente é um ponto no qual há um dado observado. Com o intuito de fazer esta distinção, pontos com dados observados serão denotados por j. Agora suponha que a superfície de cada parâmetro é aproximadamente *plana* na vizinhança de i. Note que esta é uma hipótese forte para o modelo e sua aplicabilidade para determinado conjunto de dados deve ser avaliada pelo analista. A Figura 2.3 ilustra esta idéia.

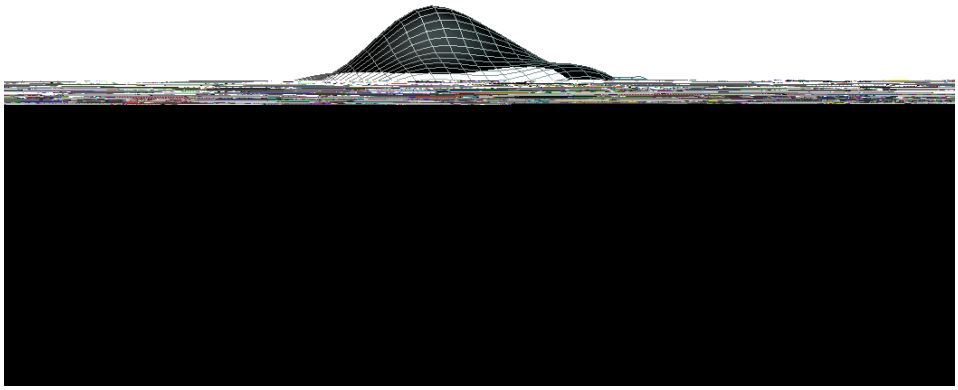

Figura 2.3: Superfície de β_k com a aproximação local no ponto i por um plano

Observe que, sob a hipótese de uma superfície de β_k com crescimento suave, a vizinhança do ponto i pode ser aproximada por um *plano*. Assim, considera-se que o parâmetro no ponto i é igual aos parâmetros dos pontos j próximos de i. Ou seja,

localmente, retorna-se ao contexto da regressão global de parâmetros constantes na região. Então, o modelo de regressão linear simples a seguir é válido para pontos j próximos de i:

$$y_j = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{jk} + \epsilon_j \ . \tag{2.16}$$

Note que a regressão (2.16) é realizada para estimar os parâmetros $\boldsymbol{\beta}$ do ponto i utilizando a informação dos pontos j $\{x_{jk}, y_j\}$, sem a necessidade de ter a informação específica de i $\{x_{ik}, y_i\}$. Considerando ainda que pontos mais próximos são mais semelhantes, pondera-se as informações existentes dos pontos observados j com pesos que decrescem a medida que eles se afastam de i. Com base nessas suposições, tem-se que a log-verossimilhança local da RGP é dada por

$$L(\boldsymbol{\beta}(u_i, v_i)|D) = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} \left[w(d_{ij}) \left(y_j - \beta_0(u_i, v_i) - \sum_k \beta_k(u_i, v_i) x_{jk} \right) \right]^2 \ , \tag{2.17}$$

onde $w(d_{ij})$ é uma função de ponderação que depende da distância entre os pontos (u_i, v_i) e (u_j, v_j).

Maximizar (2.17) é equivalente a aplicar o método de mínimos quadrados ponderados (Fotheringham et al., 2002). Este método foi abordado de forma mais geral na Seção 1.3.2.2, na qual explorou-se o método de mínimos quadrados reponderados iterativo. Como no modelo RGP estamos considerando a suposição de normalidade, não há necessidade do processo iterativo, então a Equação (1.24) simplifica-se para

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{y} \ , \tag{2.18}$$

onde $\hat{\boldsymbol{\beta}}(u_i, v_i)$ é a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ no *ponto de regressão* (u_i, v_i), e $\mathbf{W}(u_i, v_i)$ é uma matriz $n \times n$, cujos elementos fora da diagonal são zero e os elementos da diagonal, denotados aqui por w_{ij}, $j = 1, \ldots, n$, representam o peso da j-ésima observação no *ponto de regressão* i. Seja N o número de pontos de regressão, então i varia de 1 até N. Consequentemente, na RGP faz-se tantas regressões quanto o número de pontos N que se deseja estimar.

É importante ressaltar que o ponto i é um ponto *arbitrário*, pondendo, inclusive,

ser um dos pontos observados j. Neste caso, a informação dele está disponível e é utilizada na estimação. No entanto, caso não haja a informação do ponto i, não é possível estimar o valor esperado de y para ele (visto que suas covariáveis são desconhecidas), apesar de ser possível estimar seu vetor $\boldsymbol{\beta}$.

Denotando $\hat{\boldsymbol{\beta}}(u_i, v_i)$ por $\hat{\boldsymbol{\beta}}(i)$, e $\mathbf{W}(u_i, v_i)$ por $\mathbf{W}(i)$, a Equação (2.18) pode ser reescrita como

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}'\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(i)\mathbf{y} \ , \tag{2.19}$$

onde:

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{bmatrix} . \tag{2.20}$$

Observe que a matriz de pesos $\mathbf{W}(i)$ deve ser calculada para cada ponto i. As possibilidades de escolha da matriz $\mathbf{W}(i)$ serão apresentadas na Seção 2.5.2.

O erro padrão das estimativas locais do modelo RGP pode ser derivado a partir da Equação (2.19). Inicialmente, define-se a matriz $\mathbf{C}$ como

$$\boldsymbol{C} = [\mathbf{X}'\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(i). \tag{2.21}$$

Consequentemente,

$$\begin{aligned} \hat{\boldsymbol{\beta}}(i) &= \mathbf{C}\mathbf{y} \ , \\ \widehat{Var}\left[\hat{\boldsymbol{\beta}}(i)\right] &= \mathbf{C}\mathbf{C}'\hat{\sigma}^2, \end{aligned} \tag{2.22}$$

em que $\hat{\sigma}^2$ é a soma dos quadrados dos resíduos normalizados da regressão local,

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n - 2v_1 + v_2}, \tag{2.23}$$

onde $v_1 = tr(\mathbf{R})$, $v_2 = tr(\mathbf{R}'\mathbf{R})$ e $tr()$ denota o traço da matriz.

A matriz $\mathbf{R}$ é a que relaciona as *matrizes* $\hat{\boldsymbol{\mu}}$ e $\mathbf{y}$, cujas linhas $\boldsymbol{r}_j$ são dadas por

$$\boldsymbol{r}_j = \mathbf{X}_j[\mathbf{X}'\mathbf{W}(j)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(j) \ , \tag{2.24}$$

onde $\mathbf{X}_j$ é a j-ésima linha da matriz do modelo $\mathbf{X}$. O traço da matriz $\mathbf{R}$ da RGP é igual ao traço da matriz de projeção (ou do inglês, *hat matrix*) da RGP, a qual relaciona os *vetores* $\hat{\boldsymbol{\mu}}$ e $\mathbf{y}$.

Para a regressão global, tem-se que o traço da matriz de projeção é igual ao número de parâmetros do modelo. Mas, considerando que a RGP ajusta uma superfície não paramétrica para as estimativas dos parâmetros, os conceitos de número de parâmetros e graus de liberdade não fazem sentido para este modelo. No entanto, para que fosse possível implementar medidas de qualidade do ajuste e outros procedimentos inferenciais, definiu-se o número *efetivo* de graus de liberdade e o número *efetivo* de parâmetros. Sendo assim, para a RGP, o termo $2v_1 - v_2 = 2tr(\mathbf{R}) - tr(\mathbf{R}'\mathbf{R})$ é o número efetivo de parâmetros e, consequentemente, $n - 2v_1 + v_2$ é o número efetivo de graus de liberdade do resíduo. Visto que $tr(\mathbf{R}) \approx tr(\mathbf{R}'\mathbf{R})$, o número efetivo de parâmetros pode ser aproximado por v_1, não sendo necessário calcular o $tr(\mathbf{R}'\mathbf{R})$, de forma que o esforço computacional fica simplificado (Fotheringham et al., 2002).

Note que apesar do erro padrão da RGP ser local, o valor de $\hat{\sigma}^2$ dado por (2.23) é global. Consequentemente, estes erros são diferentes dos obtidos diretamente das regressões locais, os quais são baseados em estimativas locais de σ. Fotheringham et al. (2002) recomendam considerar a estimativa de σ global no cálculo dos erros padrão da RGP, no entanto, os autores também afirmam que pouca diferença foi observada entre as duas formas de estimação.

2.5.2 Função de ponderação espacial

A função de ponderação espacial é a que determina como os pesos w_{ij} da matriz $\mathbf{W}(i)$ serão calculados. A Figura 2.4 apresenta um exemplo de função de ponderação espacial.

A seguir, estão apresentadas algumas possibilidades de escolha para a função de ponderação (Fotheringham et al., 2002). A notação d_{ij} representa a distância do ponto i para a observação j, d é uma distância pré-determinada e b é o parâmetro de

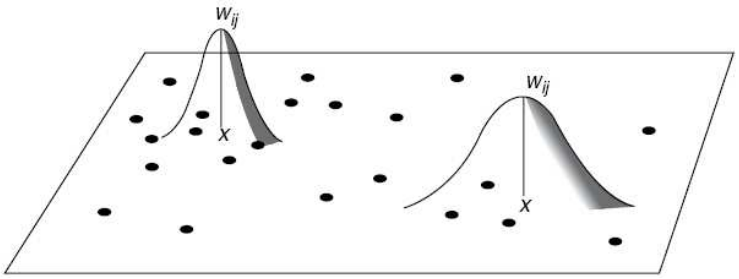

Figura 2.4: Função de ponderação espacial
Fonte: Fotheringham et al. (2002)

suavização (ou do inglês, *bandwidth*). Este parâmetro controla a variância da função de ponderação e, consequentemente, determina a velocidade de decaimento do peso com a distância.

1. $w_{ij} = 1$ se $d_{ij} < d$, e $w_{ij} = 0$ caso contrário;

2. $w_{ij} = \exp\{-\frac{1}{2}(d_{ij}/b)^2\}$;

3. $w_{ij} = [1 - (d_{ij}/b)^2]^2$ se $d_{ij} < b$, e $w_{ij} = 0$ caso contrário.

Note que se $w_{ij} = 1\ \forall i, j$, então voltamos para o modelo de regressão clássica global apresentado na Seção 1.3.1.1.

A primeira função de ponderação listada, apesar de ser mais simples, apresenta a desvantagem de ter uma descontinuidade abrupta para os pontos distantes d do ponto i, o que vai de encontro com a proposta da RGP de criar uma superfície contínua de estimação dos parâmetros. Deseja-se uma função de ponderação que decresça continuamente a medida que os pontos se distaciam. A segunda função listada, chamada de *kernel* gaussiano, é um possível candidato. Outros possíveis canditados seriam funções *quasi* gaussianas, como o *kernel* bi-quadrático apresentado no item 3.

É possível, e por vezes mais recomendado, que o parâmetro de suavização da função de ponderação varie espacialmente de acordo com a disposição dos dados observados, sendo especialmente útil quando as áreas têm tamanhos diferentes ou quando os dados não estão igualmente espaçados na região. Assim, conforme pode ser observado pela Figura 2.4, as áreas com menor densidade de pontos utilizam uma função *kernel* com maior parâmetro de suavização (ou seja, *kernel* com maior

variância), enquanto que as áreas com maior concentração de pontos empregam um parâmetro de suavização menor.

Exemplos de funções de ponderação que levam em conta a dispersão espacial dos dados estão apresentados a seguir. Para as funções 4 e 6 será necessário estimar o parâmetro N, número de pontos que devem ser incluídos na calibração do modelo.

4. $w_{ij} = 1$ se j é um dos N vizinhos mais próximos de i, e $w_{ij} = 0$ caso contrário;

5. $w_{ij} = \exp(-R_{ij}/b)$, onde R_{ij} é o posto (ou do inglês, *rank*) do ponto j com relação ao ponto i;

6. $w_{ij} = [1-(d_{ij}/b)^2]^2$ se j é um dos N vizinhos mais próximos de i, e $w_{ij} = 0$ caso contrário. Neste caso, b é a distância para o N-ésimo vizinho mais próximo.

Fotheringham et al. (2002) comentam que os resultados da RGP são relativamente insensíveis à escolha da função *kernel*, no entanto, são muito sensíveis à escolha do parâmetro de suavização. De fato, considerando, por exemplo, a função de ponderação número 2, a escolha $b \to \infty$ faz com que todas as observações tenham pesos unitários, consequentemente, temos de volta a regressão clássica (Seção 1.3.1.1). No entanto, ao reduzir progressivamente o valor de b, as estimativas dos parâmetros se tornam dependentes, cada vez mais, de observações mais próximas. Consequentemente, cada superfície estimada apresenta maior variação, ou seja, um menor grau de suavização. Sendo assim, resultados muito diferentes podem ser obtidos apenas variando o parâmetro de suavização de uma determinada função de ponderação.

2.5.3 Determinação do parâmetro de suavização

Um dos métodos de determinação do parâmetro de suavização é chamado validação cruzada (ou do inglês, *cross-validation*) e foi proposto para a regressão local por Cleveland (1979),

$$CV = \sum_{j=1}^{n} \left[y_j - \hat{y}_{\neq j}(b) \right]^2 \ , \tag{2.25}$$

onde $\hat{y}_{\neq j}(b)$ é o valor ajustado para o ponto j, omitindo-se a própria observação j desse ajuste.

O valor de b que mininiza (2.25) é o parâmetro de suavização ótimo do método de validação cruzada. Note que a Equação (2.25) é uma modificação do método de mínimos quadrados ordinários, pois considera a calibração do modelo sem a j-ésima observação. Caso a observação no ponto j fosse incluída, o valor de b que minimizaria o funcional

$$\sum_{j=1}^{n} [y_j - \hat{y}(b)]^2$$

seria $b = 0$, o que não é informativo para o modelo.

Outra forma de encontrar o parâmetro de suavização é por meio do Critério de Informação de Akaike (ou do inglês, *Akaike Information Criterion* - AIC). O AIC_c (AIC corrigido) para a RGP é dado por (Fotheringham et al., 2002)

$$AIC_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + \frac{n(n + tr(\mathbf{R}))}{n - 2 - tr(\mathbf{R})}\,, \tag{2.26}$$

onde $\hat{\sigma}$ é a estimativa de máxima verossimilhança para σ, dada por

$$\hat{\sigma} = \sqrt{\frac{\sum_j (y_j - \hat{y}_j)^2}{n}},$$

e $\mathbf{R}$ é a matriz cujas linhas foram definidas na Equação (2.24).

O parâmetro de suavização que fornece um menor AIC_c é escolhido, sendo consideradas significativas diferenças entre os AIC_c maiores do que 3 (Fotheringham et al., 2002). O critério de informação de Akaike também pode ser utilizado para comparar modelos com diferentes variáveis explicativas ou para comparar o modelo RGP com outros modelos candidatos, como por exemplo, o modelo de regressão clássica (Seção 1.3.1.1).

É importante ressaltar que o método de mínimos quadrados ponderados para a Regressão Geograficamente Ponderada produz estimativas viesadas para os parâmetros. O viés surge pois o modelo ajusta regressões locais supondo que a superfície dos parâmetros é aproximadamente *plana* nas proximidades do ponto de regressão analisado, quando na verdade os parâmetros provavelmente variam continuamente no espaço (vide Figura 2.3). Por outro lado, considerando que há não estacionariedade espacial, as estimativas do modelo de regressão global serão mais ainda viesadas, pois

o mesmo assumi que o parâmetro é constante em toda região de estudo.

O viés das estimativas da RGP, assim como a variância, dependerá do parâmetro de suavização escolhido. A escolha de um parâmetro de suavização muito grande nos traz uma estimativa precisa (com menor variância) para o parâmetro, no entanto, ao considerar pontos mais distantes na calibração do modelo, estamos introduzindo viés nesta estimativa. O outro extremo produz resultados opostos, isto é, um parâmetro de suavização pequeno produz uma estimativa pouco viesada, mas com maior variância, visto que é baseada em um tamanho de amostra menor. Portanto, a escolha deste parâmetro é um ponto crucial da Regressão Geograficamente Ponderada.

Apesar disso, Staniswalis (1989) mostra que, sob certas condições (como funções log-verossimilhança limitadas, com derivadas primeira, segunda e terceira também limitadas, e $b \to 0$ quando $n \to \infty$, vide Staniswalis (1989) para mais detalhes), os estimadores que maximizam a verossimilhança local, neste caso $\hat{\beta}_k(u_i, v_i)$, são assintoticamente Normais, não viesados e consistentes.

2.5.4 Testes de não estacionariedade

Uma maneira de testar a estacionariedade de um parâmetro da RGP é verificando a variabilidade de suas estimativas no espaço. Para o parâmetro k, tem-se que esta variabilidade pode ser calculada como

$$V_k = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\beta}_{ik} - \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{ik}\right)^2 . \tag{2.27}$$

A significância estatística de V_k pode ser avaliada por meio de um teste de aleatorização (Hope, 1968), no qual a distribuição empírica da variável V_k é construída sob a hipótese de estacionariedade espacial. Inicialmente, as coordenadas geográficas das observações são permutadas, de forma aleatória, e a RGP é realizada para esta nova configuração. Com base nas novas estimativas, calcula-se V_k por meio de (2.27). Repetindo-se este procedimento m vezes, chega-se à distribuição empírica de V_k sob a hipótese nula. Ordenando os V_k's de forma decrescente, tem-se que o p-valor do teste é dado pela razão $p/(m+1)$, onde p é o posto do V_k originalmente observado.

O teste de permutação tem a vantagem de não pressupor uma distribuição de

probabilidade para a variável V_k, mas apresenta a desvantagem de ser computacionalmente intensivo, visto que é necessário realizar a RGP um número grande m de vezes.

Leung et al. (2000a,b) formularam um teste de estacionariedade paramétrico baseado na estatística F. Este teste foi adaptado posteriormente por Fotheringham et al. (2002) a fim de facilitar o esforço computacional. Considerando este teste adaptado, tem-se que

$$\frac{V_k}{\sigma^2} \sim \boldsymbol{\chi}_g \ , \tag{2.28}$$

onde g é o número de graus de liberdade da distribuição Qui-quadrado, dado por

$$g = \text{traço}\left(\frac{1}{n}\mathbf{B}'(\mathbf{I_n} - \frac{1}{n}\mathbf{1_n})\mathbf{B}\right) \ , \tag{2.29}$$

$\mathbf{I_n}$ é a matriz identidade $n \times n$, $\mathbf{1_n}$ é uma matriz $n \times n$ de 1's e a matriz $\mathbf{B}$ é dada por

$$\mathbf{B} = \begin{bmatrix} \mathbf{e}_k'[\mathbf{X}'\mathbf{W}(1)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(1) \\ \vdots \\ \mathbf{e}_k'[\mathbf{X}'\mathbf{W}(n)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(n) \end{bmatrix} \ , \tag{2.30}$$

onde $\mathbf{e}_k$ é um vetor coluna com 1 na k-ésima posição e zero nas outras.

Além disso, tem-se que

$$\frac{f\,\hat{\sigma}^2}{\sigma^2} \sim \boldsymbol{\chi}_f \ , \tag{2.31}$$

onde $\hat{\sigma}^2$ é dado em (2.23) e f é o número efetivo de graus de liberdade do resíduo,

$$f = n - 2tr(\mathbf{R}) - tr(\mathbf{R}'\mathbf{R}) \ . \tag{2.32}$$

Então, a estatística F que testa a estacionariedade do k-ésimo parâmetro é dada por

$$\frac{V_k}{g\,\hat{\sigma}^2} \sim \boldsymbol{F}_{g,f} \ . \tag{2.33}$$

Sendo assim, caso a hipótese nula de estacionariedade espacial de $\boldsymbol{\beta}_k$ seja rejeitada, para algum k, tem-se que é mais apropriado utilizar um modelo de regressão local.

2.6 Regressão de Poisson Geograficamente Ponderada

A Regressão de Poisson Geograficamente Ponderada - RPGP (ou do inglês, Geographically Weighted Poisson Regression - GWPR) foi desenvolvida por Nakaya et al. (2005). O modelo foi utilizado pelos autores para examinar a relação entre a taxa de mortalidade e fatores sócio-econômicos na área metropolitana de Tóquio. Para isso utilizaram os dados do censo nacional e das estatísticas vitais do ano de 1990 do Japão.

Como a distribuição de Poisson e a Binomial Negativa pertencem à classe dos MLG de Nelder e Wedderburn (1972), o modelo de Regressão Geograficamente Ponderada para a distribuição Binomial Negativa será construído com base na metodologia desenvolvida por Nakaya et al. (2005) para a RPGP, que será apresentada nesta seção.

2.6.1 Modelo RPGP

Considere o modelo de Poisson parametrizado em termos da taxa μ_j/t_j, onde t_j representa o tempo de exposição ou a área na qual os eventos ocorrem, sendo considerada uma variável *offset*. Portanto, de acordo com a regressão de Poisson apresentada na Seção 1.3.1.2, temos que

$$\log\left(\mu_j\right) = \sum_k \beta_k x_{jk} + \log\left(t_j\right) \ .$$

Então,

$$\log\left(\frac{\mu_j}{t_j}\right) = \sum_k \beta_k x_{jk}$$

e

$$\mu_j = t_j \exp\left(\sum_k \beta_k x_{jk}\right) .$$

Permitindo variação espacial aos parâmetros β_k, tem-se que

$$\mu_j = t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right) . \tag{2.34}$$

Assim, o modelo de Regressão de Poisson Geograficamente Ponderada pode ser escrito como

$$y_j \sim \text{Poisson}\left[t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right)\right] . \tag{2.35}$$

A calibração do modelo (2.35) pode ser feita maximizando a função local de log-verossimilhança (as justificativas estão apresentadas na Seção 2.5.1). Para isso, considere, inicialmente, a função de log-verossimilhança global dada para a Poisson por

$$L(\boldsymbol{\beta}(u, v)|D) = \sum_{j=1}^{n} -\mu_j + y_j \log \mu_j , \tag{2.36}$$

onde D representa os dados $\{x_{jk}\}$, $\{y_j\}$ e $\{(u_j, v_j)\}$ e μ_j é definido por (2.34), logo depende de $\boldsymbol{\beta}(u, v)$. Considerando a hipótese da superfície de β_k aproximadamente *plana* nas proximidades de um ponto i qualquer, a log-verossimilhança local pode ser escrita como

$$L(\boldsymbol{\beta}(u_i, v_i)|D) = \sum_{j=1}^{n} \{ -\mu_j(\boldsymbol{\beta}(i)) + y_j \log [\, \mu_j(\boldsymbol{\beta}(i)) \,] \} \, w(d_{ij}) , \tag{2.37}$$

onde $\mu_j(\boldsymbol{\beta}(i))$ é o valor esperado de y no ponto j com base nos parâmetros do ponto i, ou seja,

$$\mu_j(\boldsymbol{\beta}(i)) = t_j \exp\left(\sum_k \beta_k(u_i, v_i) x_{jk}\right) . \tag{2.38}$$

É importante ressaltar que as médias apresentadas em (2.38) são calculadas apenas como passo intermediário na estimação dos parâmetros $\boldsymbol{\beta}(i)$. O valor de fato estimado para y no ponto j é calculado com base nos parâmetros estimados também para o ponto j, ou seja, é $\hat{\mu}_j(\boldsymbol{\beta}(j))$.

A maximização da Equação (2.37) é resolvida pelo método escore de Fisher (vide

Seção 1.3.2.2) modificado. A modificação tem o intuito de incluir, no algoritmo MQRI, a ponderação geográfica dada pela matriz de proximidade espacial $\mathbf{W}(i)$. Isto é feito multiplicando a matriz de pesos do MQRI pela matriz de pesos da RGP (Fotheringham et al., 2002). A solução desta maximização é dada por

$$\boldsymbol{\beta}(u_i, v_i)^{(m+1)} = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)^{(m)}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)^{(m)}\mathbf{z}(u_i, v_i)^{(m)} \ , \tag{2.39}$$

onde $\mathbf{X}$ é a matriz do modelo, $\mathbf{W}(u_i, v_i)$ é a matriz diagonal de pesos da RGP e $\mathbf{A}(u_i, v_i)^{(m)}$ é a matriz diagonal de pesos do MLG na iteração m para a localidade i. Os elementos $a_{ij}^{(m)}$ $(j = 1, \ldots, n)$ da diagonal de $\mathbf{A}(u_i, v_i)^{(m)}$ são dados por (vide Seção 1.3.1.2 para detalhes da Poisson)

$$a_{ij}^{(m)} = \frac{1}{V(\mu_j)}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2 = \mu_j(\boldsymbol{\beta}(u_i, v_i)^{(m)}) \ . \tag{2.40}$$

Por fim, $\mathbf{z}$ é o vetor da variável dependente ajustada do algoritmo MQRI, cujos elementos z_j foram apresentados na Equação (1.25). Para a RPGP, tem-se que:

$$z_j(\boldsymbol{\beta}(i))^{(m)} = X\boldsymbol{\beta}(i)^{(m)} + \frac{y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})}{\mu_j(\boldsymbol{\beta}(i)^{(m)})} \ . \tag{2.41}$$

Após a convergência do algoritmo (vide Seção 1.3.2.2), tem-se que

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = \mathbf{C}(u_i, v_i)\mathbf{z}(u_i, v_i) \ , \tag{2.42}$$

onde

$$\boldsymbol{C}(u_i, v_i) = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i) \ . \tag{2.43}$$

Utilizando a matriz de variância e covariância da variável dependente ajustada $\mathbf{z}$ da regressão global, dada por $\mathbf{A}^{-1}$, agora estimada no ponto i, ou seja $\mathbf{A}(u_i, v_i)^{-1}$, tem-se que a matriz de variância e covariância das estimativas dos parâmetros da

RPGP pode ser calculada como

$$\widehat{Cov}\left[\widehat{\boldsymbol{\beta}}(u_i, v_i)\right] = \mathbf{C}(u_i, v_i)\mathbf{A}(u_i, v_i)^{-1}\mathbf{C}'(u_i, v_i) \ . \tag{2.44}$$

Portanto, tem-se que o erro padrão da estimativa do k-ésimo parâmetro para a localidade i é

$$se\left[\widehat{\beta}_k(u_i, v_i)\right] = \sqrt{\widehat{Cov}\left[\widehat{\boldsymbol{\beta}}(u_i, v_i)\right]_k} \ , \tag{2.45}$$

onde $\widehat{Cov}\left[\widehat{\boldsymbol{\beta}}(u_i, v_i)\right]_k$ é o k-ésimo elemento da diagonal dessa matriz.

Note que, novamente, os erros padrão assim estimados não são os mesmos dos resultantes das regressões locais ponderadas, visto que a estimativa utilizada para a matriz de variância e covariância de $\mathbf{z}(u_i, v_i)$ não é ponderada pelos pesos $\mathbf{W}$.

Utilizando (2.45) é possível mapear a superfície não paramétrica dos erros padrão das estimativas de cada parâmetro do modelo de RPGP a fim de avaliar a confiabilidade das estimativas. No entanto, é importante ressaltar que a estimativa do erro padrão é pontual. E, apesar de ser possível construir superfícies dos limites inferior e superior de 95% de confiança para as estimativas dos parâmetros, este par de superfícies não representa um envelope de 95% de confiança para a superfície como um todo (Fotheringham et al., 2002).

A significância local da estimativa do k-ésimo parâmetro no ponto i pode ser avaliada pela estatística *pseudo t* dada por

$$t_k(u_i, v_i) = \frac{\widehat{\beta}_k(u_i, v_i)}{se\left[\widehat{\beta}_k(u_i, v_i)\right]} \ , \tag{2.46}$$

cuja distribuição é aproximadamente Normal. A fim de evitar problemas causados por múltiplos testes de hipótese, é possível utilizar a correção de Bonferroni. Esta correção ajusta o valor crítico do teste de hipótese para cima, determinando um novo nível de significância igual ao nível original dividido pelo número efetivo de parâmetros da RGP. No entanto, esta correção é conservativa, devido à dependência entre os múltiplos testes de hipótese conduzidos. E, se o número de pontos de regressão for grande, muito dificilmente os valores estimados serão significativos (Fotheringham

et al., 2002).

Note que o algoritmo escore de Fisher modificado também deve ser repetido para cada ponto de regressão i a fim de obter as estimativas locais dos parâmetros $\boldsymbol{\beta}$.

Algumas possibilidades de escolha da matriz de ponderação espacial $\mathbf{W}(u_i, v_i)$ estão apresentadas na Seção 2.5.2. Esta escolha envolve também a determinação do parâmetro de suavização, o qual pode ser encontrado de forma a minimizar a estatística AICc. A correção no AIC proposta por Hurvich e Tsai (1989) para seleção de modelos de regressão com tamanho pequeno de amostra é dada por

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}\,, \tag{2.47}$$

onde k é o número de parâmetros.

Com base na definição de AIC dada por Akaike (1974), tem-se que

$$AIC = -2L(\boldsymbol{\beta}) + 2k\,. \tag{2.48}$$

Nakaya et al. (2005) consideram $AIC = Dev + 2k$, onde Dev é o desvio dado por (1.30). Para a RPGP, as duas definições chegam a conclusões equivalentes, visto que diferem apenas de um termo constante referente a log-verossimilhança do modelo saturado. Assim, utilizando a definição (2.48) e os resultados da RPGP tem-se que

$$AICc = -2\left(\sum_{j=1}^{n} -\mu_j + y_j \log \mu_j\right) + 2tr(\mathbf{R}) + \frac{2tr(\mathbf{R})(tr(\mathbf{R})+1)}{n - tr(\mathbf{R}) - 1}\,, \tag{2.49}$$

onde μ_j é dado por (2.34) e $tr(\mathbf{R})$ é o número efetivo de parâmetros da RPGP. A matriz $\mathbf{R}$ é a que relaciona as *matrizes* $\hat{\boldsymbol{\eta}}$ e $\mathbf{z}$, ou seja, $\hat{\boldsymbol{\eta}} = \mathbf{R}\mathbf{z}$. As linhas $\boldsymbol{r}_j$ de $\mathbf{R}$ são dadas por

$$\boldsymbol{r}_j = \mathbf{X}_j[\mathbf{X}'\mathbf{W}(j)\mathbf{A}(j)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(j)\mathbf{A}(j)\,, \tag{2.50}$$

onde $\mathbf{X}_j$ é a j-ésima linha da matriz do modelo $\mathbf{X}$.

Observe que, como $\hat{\boldsymbol{\eta}}$ e $\hat{\boldsymbol{y}}$ podem ser estimados apenas para os pontos j (ou pontos observados), a matriz $\mathbf{R}$ é construída considerando que os *pontos de regressão*, que anteriormente poderiam ser qualquer ponto i no espaço, agora são exclusivamente os

próprios pontos j. Sendo assim, não é possível calcular o número efetivo de parâmetros e o AICc para a RGP realizada com o *grid* dos pontos i. Consequentemente, para este caso, não há uma técnica de determinação do parâmetro de suavização ótimo e de avaliação da qualidade do modelo. Uma possibilidade é realizar a RGP com base apenas nos pontos observados j, obtendo-se assim o parâmetro de suavização ótimo e o AICc do modelo. E utilizar estes resultados também na estimação da superfície dos parâmetros $\boldsymbol{\beta}_k$ com o *grid*.

É importante lembrar que a estatística AICc também é utilizada para comparar a RPGP com outros modelos concorrentes. Além disso, outra forma de avaliar a RPGP é por meio de testes de não estacionariedade (vide Seção 2.5.4). Ou seja, se algum parâmetro β_k for não estacionário, então provavelmente a RPGP resultará em um melhor ajuste aos dados.

Capítulo 3

Regressão Binomial Negativa Geograficamente Ponderada

3.1 Introdução

A Regressão Binomial Negativa Geograficamente Ponderada proposta nesta dissertação, denominada RBNGP, será detalhada neste capítulo. Este modelo de regressão é indicado para modelar dados espaciais de contagem não estacionários e com superdispersão. A hipótese de igualdade entre média e variância da distribuição de Poisson é flexibilizada neste modelo devido à presença do parâmetro de superdispersão α da distribuição Binomial Negativa. Com isso, a RBNGP é mais robusta do que a RPGP apresentada no capítulo anterior.

O modelo RBNGP aqui proposto permite que tanto os parâmetros β da regressão quanto o parâmetro α variem espacialmente. No entanto, devido à dificuldade de estimar a contribuição para o número de parâmetros efetivos do modelo da variação espacial do parâmetro α, também será apresentado neste trabalho o modelo de Regressão Binomial Negativa Geograficamente Ponderada com α global, o qual denominaremos de RBNGPg.

Estes modelos são úteis para resolver problemas práticos, de interesse particular ou coletivo, auxiliando a tomada de decisão, a elaboração de políticas públicas ou a maximização de lucros. Considere, por exemplo, relacionar o número de acidentes de moto nas cidades do país em função das suas características de trânsito, como a frota de motos e o percentual de *motoboys* dentre os motociclistas. Analisando o

comportamente espacial da relação entre essas variáveis é possível identificar cidades cuja intensidade desses relacionamentos é menor quando comparada com outras de mesmo porte. A compreensão dos motivos que trouxeram este melhor desempenho para algumas regiões pode ser utilizada na elaboração de políticas públicas para as regiões com pior desempenho, gerando medidas mais eficientes e direcionadas.

3.2 Modelo RBNGP

O modelo global de Regressão Binomial Negativa mais utilizado (NB-2) considera uma função de ligação logarítmica (vide Seção 1.3.1.3). Parametrizando este modelo em termos da taxa μ_j/t_j, onde t_j é uma variável *offset* tem-se que

$$y_j \sim \text{BN}\left[t_j \exp\left(\sum_k \beta_k x_{jk}\right), \alpha\right] \ . \tag{3.1}$$

A RBNGP é uma extensão do modelo global (3.1) que permite variação espacial aos parâmetros β_k e α. Sem limitar a forma funcional dessa variação, a RBNGP produz superfícies não paramétricas para as estimativas dos parâmetros. Este modelo espacial local é descrito por

$$y_j \sim \text{BN}\left[t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right), \alpha(u_j, v_j)\right] \ . \tag{3.2}$$

A estimação dos parâmetros do modelo global (3.1) é feita de forma iterativa, combinando o método NR (Tabela 1.1) com o método MQRI (Tabela 1.2). Ou seja, com base no α estimado por NR, estima-se o vetor $\boldsymbol{\beta}$ pelo MQRI. Em seguida, utilizando este novo $\boldsymbol{\beta}$, atualiza-se o valor de α, e assim sucessivamente até atingir convergência. A calibração do modelo RBNGP (3.2) proposto nesta dissestação envolve também esta alternância entre NR e MQRI. No entanto, algumas modificações nestes métodos são necessárias a fim de considerar a essência local da RBNGP.

O método MQRI modificado será primeiramente descrito. Para isso, suponha, inicialmente, que os parâmetros $\boldsymbol{\alpha}(u, v)$ são conhecidos. Então, a log-verossimilhança

da RBNGP, como função de $\boldsymbol{\beta}(u,v)$, é dada por

$$L(\boldsymbol{\beta}(u,v)|D,\boldsymbol{\alpha}(u,v)) = \sum_{j=1}^{n} \{y_j \log(\alpha_j\mu_j) - (y_j + 1/\alpha_j)\log(1+\alpha_j\mu_j) + \log[\Gamma(y_j + 1/\alpha_j)] - \log[\Gamma(1/\alpha_j)] - \log[\Gamma(y_j+1)]\} \ , \tag{3.3}$$

onde $D = \{x_{jk}\}$, $\{y_j\}$ e $\{(u_j, v_j)\}$ e, para $j = 1, \ldots, n$,

$$\mu_j = t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right) , \tag{3.4}$$

$$\alpha_j = \alpha(u_j, v_j) \ . \tag{3.5}$$

Considerando novamente a hipótese da superfície de β_k aproximadamente *plana* na vizinhança de um ponto i qualquer (vide Seção 2.5.1 para mais detalhes), a log-verossimilhança local da RBNGP pode ser escrita como

$$L(\boldsymbol{\beta}(u_i,v_i)|D,\alpha(i)) = \sum_{j=1}^{n} \{y_j \log[\alpha(i)\mu_j(\boldsymbol{\beta}(i))] - [y_j + 1/\alpha(i)]\log[1+\alpha(i)\mu_j(\boldsymbol{\beta}(i))] + \log[\Gamma(y_j + 1/\alpha(i))] - \log[\Gamma(1/\alpha(i))] - \log[\Gamma(y_j+1)]\} w(d_{ij}) \ , \tag{3.6}$$

onde, para $i = 1, \ldots, N$,

$$\mu_j(\boldsymbol{\beta}(i)) = t_j \exp\left(\sum_k \beta_k(u_i, v_i) x_{jk}\right) , \tag{3.7}$$

$$\alpha(i) = \alpha(u_i, v_i) \ . \tag{3.8}$$

A maximização da log-verossimilhança local (3.6) fornece as estimativas $\hat{\boldsymbol{\beta}}(i)$ dos parâmetros da RBNGP. Assim, com base nos resultados de Nakaya et al. (2005) para a RPGP (vide Seção 2.6.1) tem-se que a solução desta maximização é dada por

$$\boldsymbol{\beta}(u_i,v_i)^{(m+1)} = [\mathbf{X'W}(u_i,v_i)\mathbf{A}(u_i,v_i)^{(m)}\mathbf{X}]^{-1}\mathbf{X'W}(u_i,v_i)\mathbf{A}(u_i,v_i)^{(m)}\mathbf{z}(u_i,v_i)^{(m)} \ , \tag{3.9}$$

onde $\mathbf{X}$ é a matriz do modelo

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \tag{3.10}$$

$\mathbf{W}(u_i, v_i)$ é a matriz diagonal de pesos da RGP no ponto i

$$\mathbf{W}(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{pmatrix}, \tag{3.11}$$

$\mathbf{A}(u_i, v_i)^{(m)}$ é a matriz diagonal de pesos do MLG na iteração m para a localidade i

$$\mathbf{A}(u_i, v_i)^{(m)} = \begin{pmatrix} a_{i1}^{(m)} & 0 & \dots & 0 \\ 0 & a_{i2}^{(m)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{in}^{(m)} \end{pmatrix}. \tag{3.12}$$

Nesta dissertação, será utilizado o MQRI com a matriz de informação de Fisher observada (vide Seção 1.3.2.2). Assim, os elementos $a_{ij}^{(m)}$ $(j = 1, \dots, n)$ da diagonal dessa matriz para a RBNGP são dados por

$$a_{ij}^{(m)} = \frac{\mu_j(\boldsymbol{\beta}(i)^{(m)})}{1 + \alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)})} + \frac{\left[y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})\right]\left[\alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)})\right]}{1 + 2\alpha(i)\mu_j(\boldsymbol{\beta}(i)^{(m)}) + \alpha^2(i)\mu_j^2(\boldsymbol{\beta}(i)^{(m)})}. \tag{3.13}$$

Por fim, os elementos da variável dependente ajustada $\mathbf{z}(u_i, v_i)^{(m)}$ são

$$z_j(\boldsymbol{\beta}(i))^{(m)} = X\boldsymbol{\beta}(i)^{(m)} + \frac{y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})}{a_{ij}^{(m)}(1 + \alpha(i) \times \mu_j(\boldsymbol{\beta}(i)^{(m)}))}. \tag{3.14}$$

Assim como na RPGP, a matriz de variância e covariância para as estimativas dos

parâmetros pode ser estimada como

$$\widehat{Cov}\left[\widehat{\boldsymbol{\beta}}(u_i, v_i)\right] = \mathbf{C}(u_i, v_i)\mathbf{A}(u_i, v_i)^{-1}\mathbf{C}'(u_i, v_i) \ , \tag{3.15}$$

onde

$$\boldsymbol{C}(u_i, v_i) = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i) \tag{3.16}$$

e os elementos de $\mathbf{A}(u_i, v_i)$ e $\mathbf{z}(u_i, v_i)$ são dados, respectivamente, em (3.13) e (3.14).

Supondo agora que $\boldsymbol{\beta}(u_i, v_i)$ são conhecidos, os parâmetros $\alpha(i)$ serão estimados utilizando o método NR baseado na log-verossimilhança local. A fim de facilitar os cálculos, serão estimados os parâmetros $r(i)$, lembrando que $r(i) = 1/\alpha(i)$. Portanto, reescrevendo (3.6) com base nesta substituição obtém-se que

$$L(r(i)|D, \boldsymbol{\beta}(i)) = \sum_{j=1}^{n} \{y_j \log[\mu_j(\boldsymbol{\beta}(i))] - [y_j + r(i)] \log[r(i) + \mu_j(\boldsymbol{\beta}(i))] +$$
$$r(i) \log[r(i)] + \log[\Gamma(y_j + r(i))] - \log[\Gamma(r(i))] - \log[\Gamma(y_j + 1)]\} \, w(d_{ij}) \ . \tag{3.17}$$

Maximizando a log-verossimilhança local (3.17) por meio do método NR univariado, tem-se que

$$r(i)^{(m+1)} = r(i)^{(m)} - \left[H(i)^{(m)}\right]^{-1} U(i)^{(m)} \ , \tag{3.18}$$

onde $U(i)^{(m)}$ e $H(i)^{(m)}$ são as derivadas primeira e segunda da log-verossimilhança local com respeito a $r(i)^{(m)}$, ou seja,

$$U(i)^{(m)} = \frac{dL(r(i))}{dr(i)} = \sum_{j=1}^{n} \Big\{\psi\big[r(i)^{(m)} + y_j\big] - \psi\big[r(i)^{(m)}\big] + \log[r(i)^{(m)}] + 1 -$$
$$\log[r(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))] - \frac{r(i)^{(m)} + y_j}{r(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))}\Big\} \, w(d_{ij}) \ , \tag{3.19}$$

$$H(i)^{(m)} = \frac{d^2 L(r(i))}{dr^2(i)} = \sum_{j=1}^{n} \left\{ \psi'\left[r(i)^{(m)} + y_j\right] - \psi'\left[r(i)^{(m)}\right] + \frac{1}{r(i)^{(m)}} - \right.$$
$$\left. \frac{2}{r(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))} + \frac{r(i)^{(m)} + y_j}{\left[r(i)^{(m)} + \mu_j(\boldsymbol{\beta}(i))\right]^2} \right\} w(d_{ij}) \ , \quad (3.20)$$

onde $\psi(.)$ e $\psi'(.)$ são, respectivamente, as funções digama e trigama, dadas por

$$\psi(z) = \frac{d \log \Gamma(z)}{dz} \quad \text{e} \quad \psi'(z) = \frac{d\psi(z)}{dz} = \frac{d^2 \log \Gamma(z)}{dz^2} \ .$$

Pelo método delta, tem-se que para uma função $g(.)$ diferenciável em θ, se ocorre a convergência em distribuição $\hat{\theta}_n \to N(\theta, \sigma_n^2)$, então $g(\hat{\theta}_n) \to N\left(g(\theta), [g'(\theta)]^2 \times \sigma_n^2\right)$ também converge em distribuição (Casella e Berger, 2001). Considerando ainda que, sob certas condições (vide Seção 2.5.3) os estimadores que maximizam a verossimilhança local são assintoticamente Normais, não viesados e consistentes (Staniswalis, 1989). Então, tem-se que

$$\hat{\alpha}(i) = \frac{1}{\hat{r}(i)} \quad \text{e} \quad Var(\hat{\alpha}(i)) = \frac{-1}{H(i)\hat{r}^4(i)} \ , \quad (3.21)$$

pois $Var(\hat{r}(i)) = -1/H(i)$, onde $H(i)$ é dado por (3.20), e $[g'(r(i))]^2 = 1/r^4(i)$.

Assim, para cada ponto de regressão i, realiza-se os algoritmos de NR e MQRI de forma alternada até obter a convergência das estimativas dos parâmetros.

A fim de completar o ajuste do modelo, falta ainda estimar o parâmetro de suavização. Uma possibilidade seria determiná-lo de forma a minimizar o AICc. Utilizando as Equações (2.47) e (2.48), tem-se que

$$AIC_c = -2L(\boldsymbol{\beta}, \boldsymbol{\alpha}) + 2k + \frac{2k(k+1)}{n-k-1} \ , \quad (3.22)$$

onde k é o número efetivo de parâmetros e $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$ é a log-verossimilhança da RBNGP apresentada em (3.3).

O número efetivo de parâmetros da RBNGP pode ser escrito como $k = k_1 + k_2$, onde k_1 e k_2 são os números efetivos de parâmetros devido a $\boldsymbol{\beta}$ e a $\boldsymbol{\alpha}$, respectivamente. Conforme visto na Seção 2.6.1, k_1 é dado pelo traço da matriz $\mathbf{R}$, cujas linhas foram apresentadas em (2.50). No entanto, nesta dissertação, não foi possível estimar k_2, ou

seja, a contribuição da superfície de $\hat{\boldsymbol{\alpha}}$ para o número efetivo de parâmetros do modelo. Consequentemente, optou-se por estimar o parâmetro de suavização utilizando a medida de qualidade do ajuste da validação cruzada, indicada em (2.25).

Note que a indeterminação de k_2 não impede que a RBNGP seja ajustada. No entanto, dificulta a comparação de modelos visto que a complexidade da RBNGP, dada pelo número efetivo de parâmetros, é desconhecida. Ainda assim, os testes de não estacionariedade, apresentados na Seção 2.5.4, podem ser utilizados a fim de avaliar a não estacionariedade dos parâmetros e, consequentemente, justificar o uso do modelo. Visto que o parâmetro α não faz parte do componente linear da regressão, para ele é possível aplicar apenas o teste não paramétrico. Por isso, para todos os parâmetros utilizar-se-á o teste de não estacionariedade não paramétrico.

3.3 Modelo RBNGPg

A fim de contornar a dificuldade de estimação de k_2, propomos também nesta dissertação o modelo de Regressão Binomial Negativa Geograficamente Ponderada com α global, chamado de RBNGPg. Neste, a variação espacial é permitida somente aos parâmetros $\boldsymbol{\beta}$, ou seja,

$$y_j \sim \text{BN}\left[t_j \exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right), \alpha\right] . \tag{3.23}$$

No modelo RBNGPg, a estimação do parâmetro α é realizada de forma global, ou seja, supondo que *todos* os parâmetros do modelo são estacionários, estima-se uma superdispersão global α para ser utilizada nas estimativas locais de $\boldsymbol{\beta}$. Consequentemente, propõe-se que a estimativa de α da RBNGPg seja a mesma da regressão global Binomial Negativa.

Já os parâmetros $\boldsymbol{\beta}(u, v)$ são estimados pelo MQRI da mesma forma que na RBNGP, agora considerando que $\alpha(i) = \alpha$ para todo i. Note que não há a necessidade da alternância entre NR e MQRI, visto que uma vez estimado α de forma global, estima-se $\boldsymbol{\beta}(i)$, para cada ponto de regressão i, somente pelo MQRI.

Visto que não há variação espacial para α, sua contribuição para o número efetivo de parâmetros do modelo é unitária, ou seja, $k_2 = 1$. Consequentemente, a deter-

minação do parâmetro de suavização pode ser feita utilizando a estatística AICc da Equação 3.22, onde

$$L(\boldsymbol{\beta}(u,v),\alpha|D) = \sum_{j=1}^{n} \{y_j \log(\alpha\mu_j) - (y_j + 1/\alpha)\log(1+\alpha\mu_j) + \log[\Gamma(y_j + 1/\alpha)] - \log[\Gamma(1/\alpha)] - \log[\Gamma(y_j+1)]\} \ , \quad (3.24)$$

e $k = tr((R)) + 1$.

É importante ressaltar que, caso o teste de não estacionariedade para o modelo RBNGP indique que o parâmetro α é estacionário, isso não implica que o modelo RBNGPg deva ser utilizado. Lembre-se que α é estimado para o modelo RBNGPg considerando que *todos* os seus parâmetros são estacionários. Seria interessante a formulação de um modelo que estimasse α global, mas que ao mesmo tempo levasse em conta, para a estimação do próprio α, a variação espacial dos parâmetros $\boldsymbol{\beta}$.

Sendo assim, o intuito da RBNGPg é realizar a regressão geograficamente ponderada para a distribuição Binomial Negativa de uma maneira simplificada, que tem o custo de fornecer uma estimativa viesada de α e a vantagem de ser um modelo cuja complexidade é conhecida. Com isso, espera-se que a RBNGPg seja uma melhor opção para a modelagem de dados espaciais de contagem com superdispersão do que a RPGP, na qual α é fixo e zero.

Capítulo 4

Simulações e aplicações

4.1 Introdução

Este capítulo apresenta os resultados dos ajustes da RBNGP e da RBNGPg a dados reais e simulados. Os estudos de caso apresentados são exemplos simples que visam explorar as funcionalidades dos modelos propostos. Com as simulações, o intuito é avaliar se os ajustes com base na RBNGP e na RBNGPg conseguem reproduzir de forma satisfatória os exemplos simulados. Além disso, pretende-se explorar alguns casos especiais desses modelos que estão previstos na teoria. Já com as aplicações, objetiva-se mostrar a utilidade prática destes modelos teóricos na análise de dados reais. Em todos os casos, os resultados obtidos são comparados com os de alguns modelos concorrentes, a saber, RPGP e regressões globais Binomial Negativa e Poisson.

Os principais códigos desenvolvidos encontram-se no Apêndice. Todos foram implementados utilizando a linguagem SAS/IML. Uma breve descrição dos mesmos será feita a seguir, com o intuito de apresentar as ferramentas que foram utilizadas nas simulações e aplicações deste capítulo. Um maior detalhamento pode ser encontrado no Apêndice.

Para realização da RBNGP e da RBNGPg foi construída a macro $\%gwnbr$. Com ela é possível ajustar estes modelos para os pontos observados j, ou para pontos i de um *grid* especificado. Além disso, é possível selecionar a função de ponderação desejada. As opções 2, 3 e 6 apresentadas na Seção 2.5.2 estão disponíveis, as duas primeiras utilizam b fixo, enquanto que a última considera o parâmetro b variável.

A busca do parâmetro de suavização ótimo foi implementada na macro chamada *%golden*. O algoritmo de minimização utilizado foi da divisão áurea (ou do inglês, *golden search*), para detalhes desse algoritmo vide Zörnig (2011). O critério de busca pode ser escolhido dentre as estatísticas de qualidade do ajuste AICc, validação cruzada e desvio. Já as opções de função de ponderação são as mesmas da macro *%gwnbr*.

Por fim, também foi implementada a macro *%estac* que testa a estacionariedade dos parâmetros da regressão geograficamente ponderada com base no teste de aleatorização. O número de permutações é escolhido pelo analista.

Os estudos de caso apresentados neste capítulo serão ajustados utilizando estas macros desenvolvidas. Inicialmente, os modelos RBNGP e RBNGPg serão aplicados a três conjuntos de dados, os quais simulam as distribuições Binomial Negativa e Poisson, ambas com dependência espacial, e também a Binomial Negativa global. Por fim, duas aplicações a dados reais serão realizadas.

4.2 Simulação da RBNGP

O objetivo primordial dos modelos RBNGP e RBNGPg é modelar dados de contagem não estacionários e com superdispersão. Por isso, o primeiro estudo de caso é baseado na simulação de um conjunto de dados proveniente da distribuição Binomial Negativa com dependência espacial, o qual é descrito por

$$y_j \sim \text{BN}\left[\ \exp\left\{b_0(u_j,v_j) + b_1(u_j,v_j)x_{j1} + b_2(u_j,v_j)x_{j2}\right\},\ \alpha(u_j,v_j)\ \right]\ , \tag{4.1}$$

onde:

$$\begin{aligned}
&b_0(u_j,v_j) = 0.000005[abs(u_j)]^3 + 0.0005[abs(v_j)]^3\ ;\\
&b_1(u_j,v_j) = \sqrt{[u_j-\overline{u}_j]^2 \times [v_j-\overline{v}_j]^2}\ ;\\
&b_2(u_j,v_j) = 2\left\{-[(u_j-\overline{u}_j)/5]^2 - [(v_j-\overline{v}_j)/5]^2 + 0.13\right\}\ ;\\
&\alpha(u_j,v_j) = 2\left\{10^{-7}[abs(y)]^5\right\}^2\ .
\end{aligned} \tag{4.2}$$

As variáveis independentes x_{j1} e x_{j2} foram simuladas das distribuições Normal$(0,1)$ e Poisson$(m=5)$, respectivamente, sendo que esta última foi padronizada para ter também média zero e variância um. Assim, os parâmetros da regressão ficam em uma

escala comparável, facilitando a visualização da contribuição de cada covariável para o modelo.

Devido à importância deste exemplo na avaliação dos modelos propostos, duas situações, com tamanhos diferentes de amostra, serão estudadas. Inicialmente será utilizado um tamanho de amostra $n = 77$, que são os 77 municípios do estado do Espírito Santo. Aproveitou-se a configuração espacial deste estado, que será utilizada na aplicação a dados reais, também para os casos simulados. Aumentando o tamanho de amostra para $n = 504$, repetiu-se o ajuste a fim de avaliar o ganho na qualidade devido ao acréscimo de informação.

4.2.1 Tamanho de amostra $n = 77$

A fim de quantificar a dependência espacial dos dados simulados, foi calculado o índice de Moran para a variável dependente y. A matriz de proximidade espacial $\mathbf{W}$ utilizada (vide Seção 2.2.1) foi binária, indicando se a área A_i faz fronteira com a área A_j. O valor obtido foi $I = 0.5$, conforme indicado na Figura 4.1, que ilustra o diagrama de espalhamento de Moran (vide Seção 2.2.4). Este índice caracteriza uma autocorrelação espacial positiva moderada com respeito a matriz $\mathbf{W}$ binária.

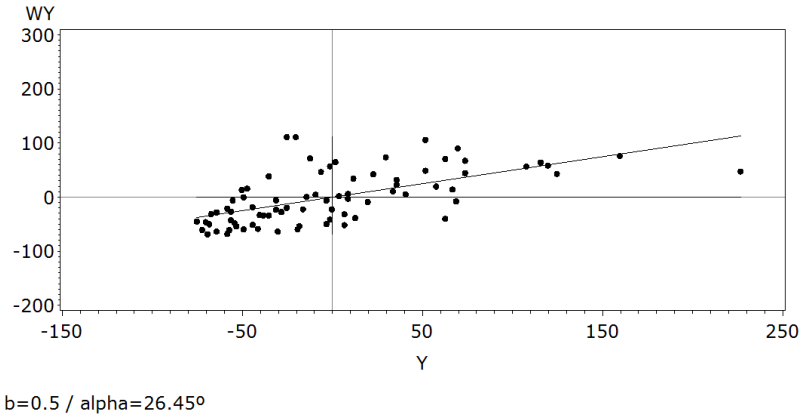

Figura 4.1: Diagrama de espalhamento de Moran da variável y

O índice de Moran global supõe que, no conjunto de dados analisado, existe uma única forma de autocorrelação espacial. No entanto, é necessário verificar se esta hipótese de estacionariedade espacial é mesmo válida. A Figura 4.2 apresenta um mapa coroplético da variável dependente y, assim como o mapa de espalhamento de Moran e o mapa de Moran de 95%.

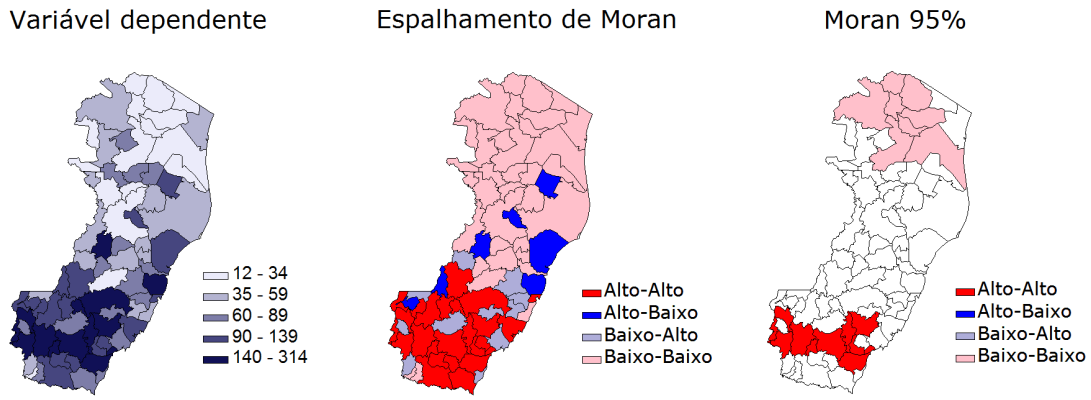

Figura 4.2: Análise exploratória da estacionariedade espacial da variável y

O mapa da variável y da Figura 4.2, cujas classificações foram definidas pelos quintis, é útil para avaliar, de forma exploratória, a existência de tendências espaciais. Observa-se, neste mapa, que valores mais elevados de y aparecem no sul do estado, enquanto que no norte as magnitudes são menores. Consequentemente, há indícios de que a variável y é não estacionária. O mapa de espalhamento de Moran também sugere esta tendência espacial. Ele mostra também que a maior parte dos municípios, os que estão coloridos em tons de vermelho, apresentam dependência espacial positiva (ou seja, estão no primeiro ou terceiro quadrantes do diagrama da Figura 4.1).

É necessário ainda avaliar se existem correlações locais em algumas regiões que são significamente diferentes das demais. O mapa de Moran de 95% indica que alguns municípios das regiões norte e sul apresentam graus de dependência diferenciados quando comparados com os demais. Portanto, a análise exploratória inicial sugere que os dados apresentam dependência espacial e são não estacionários. Consequentemente, um modelo espacial local aparenta ser mais indicado.

Além das regressões geograficamente ponderadas RBNGP, RBNGPg e RPGP, também serão avaliadas as regressões globais Binomial Negativa (RBN) e Poisson (RP), visto que estas últimas são as opções mais utilizadas para modelar dados de contagem. A Tabela 4.1 apresenta algumas medidas de qualidade e complexidade dos ajustes. A coluna b refere-se ao parâmetro de suavização, que foi encontrado minimizando o AICc, para os modelos RPGP e RBNGPg, e minimizando o CV para a RBNGP, em todos os casos utilizando o kernel biquadrático adaptativo, opção 6 da Seção 2.5.2. Neste caso, b indica o número de vizinhos mais próximos utilizado

nas regressões locais de cada modelo. Já a coluna Par. indica o número efetivo de parâmetros do modelo, enquanto a coluna $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$ apresenta sua log-verossimilhança.

Tabela 4.1: Comparação entre modelos

Modelo	b	Par.	$L(\boldsymbol{\beta}, \boldsymbol{\alpha})$	AICc
RP	-	3	-1570,4	3147
RPGP	11	44,7	-455,6	1131
RBN	-	4	-406,7	822
RBNGPg	57	8,5	-385,3	790
RBNGP	36	-	-364,8	-

Note que, na Tabela 4.1, tem-se algumas células indisponíveis. Isto ocorre pois as regressões de Poisson e Binomial Negativa não têm parâmetro de suavização, visto que são modelos globais, e para a RBNGP não é conhecido seu número efetivo de parâmetros, consequentemente seu AICc também é indeterminado. Apesar disso, algumas informações importantes podem ser extraídas da tabela. Observe que a RBN aparenta ser um melhor ajuste do que a RPGP. Ou seja, a inclusão de uma superdispersão, mesmo que global, provocou maiores ganhos de qualidade de ajuste do que a possibilidade de variação espacial para os parâmetros $\boldsymbol{\beta}$ mas com superdispersão nula. A união das vantagens de cada um desses modelos resultou em um menor AICc e uma maior log-verossimilhança para a RBNGPg. Por fim, permitindo variação espacial também para o parâmetro α, chega-se ao modelo mais verossímil RBNGP.

Nesta dissertação, optou-se por trabalhar com a log-verossimilhança do modelo ao invés do desvio. Em geral, a comparação dos desvios dos modelos é realizada quando a distribuição do componente aleatório do modelo e sua função de ligação já estão estabelecidas. Consequentemente, a log-verossimilhança do modelo saturado é idêntica para todos eles e a diferença dos desvios equivale à diferença nas log-verossimilhanças. No entanto, neste trabalho os modelos saturados são diferentes devido à estimação extra do parâmetro α. Além disso, considerando a indisponibilidade da quantidade de parâmetros e a impossibilidade de calcular o AICc, a log-verossimilhança, como sendo uma parcela do AICc, é a medida de qualidade do ajuste que surge naturalmente.

A Figura 4.3 apresenta as superfícies estimadas dos parâmetros b_0, b_1 e b_2 do componente linear das regressões geograficamente ponderadas. Os mapas reais, gerados com base nas Equações 4.2, também estão ilustrados.

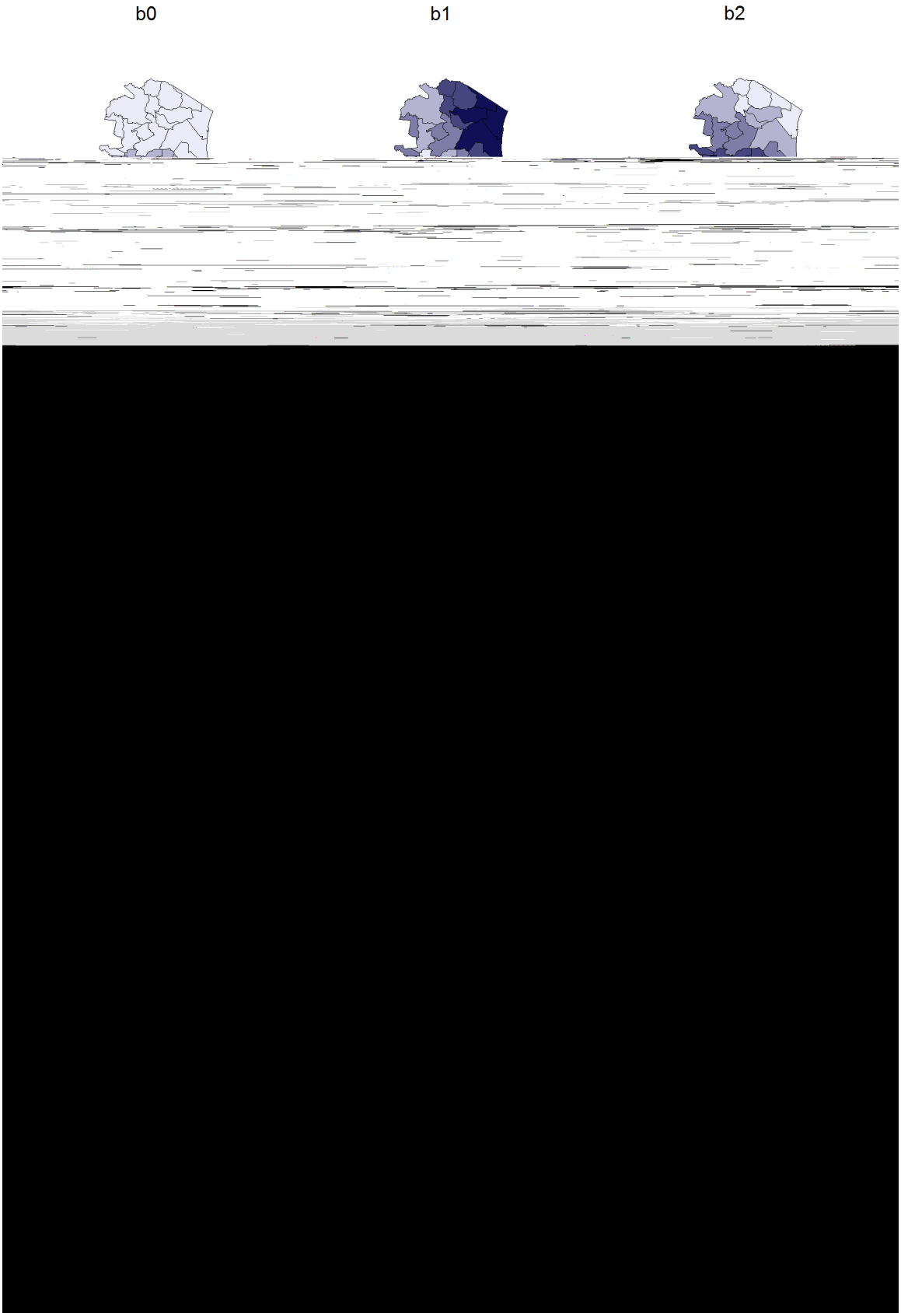

Figura 4.3: Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e b_2

A fim de facilitar a comparação, os mapas para um mesmo parâmetro foram realizados na mesma escala. Sendo assim, apenas uma escala está indicada em cada coluna da Figura 4.3. As classificações das escalas foram feitas com base nos quintis dos mapas reais de cada parâmetro. No entanto, os valores mínimo e máximo refletem as observações mais extremas considerando as estimativas deste parâmetro em todos os modelos ajustados. Portanto, caso estes valores extremos da escala sejam muito discrepantes, é importante avaliar qual modelo está gerando estes resultados atípicos. Esta forma de apresentação dos resultados será utilizada em toda a dissertação.

Analisando a primeira coluna da Figura 4.3, conclui-se que todos os modelos conseguiram captar a forma da variação espacial do intercepto do modelo, o qual assume valores maiores na região sul do estado e sofre redução gradativa na direção norte, assim como o padrão da variável y (Figura 4.2). Observe que as estimativas de b_0 da RBNGPg variam menos, sendo representadas apenas nas 3 classificações intermediárias da escala.

Os modelos também conseguiram recuperar o padrão espacial do parâmetro b_1, com valores crescendo a partir da região central. Note que, novamente, as estimativas de b_1 da RBNGPg variam pouco. Este fato ocorre pois o parâmetro de suavização ótimo para a RBNGPg foi $b = 57$ pontos (vide Tabela 4.1), ou seja, cada regressão local utilizou aproximadamente 74% dos dados totais (lembre-se que n=77 e a função de ponderação utilizada foi a adaptativa). Consequentemente, tem-se um ajuste mais suavizado para este modelo. Observe que a RPGP apresenta o ajuste menos suave, visto que seu b ótimo foi de 11 pontos, ou seja, apenas 14% dos dados totais foram utilizados em cada regressão local. Já para a RBNGP, tem-se $b = 36$ pontos (ou seja, 47% do total).

A configuração espacial do parâmetro b_2 não foi tão bem recuperada nos ajustes. As estimativas da RBNGPg e da RBNGP captaram o crescimento do parâmetro do sul para o centro do estado, mas não a sua queda do centro para o norte. Já a RPGP apresentou problemas com a magnitude dos valores estimados, sendo responsável pelos valores extremos atípicos da escala de b_2.

A Figura 4.4 apresenta os resultados do ajuste para o parâmetro α. A coluna "Status da Estimativa" indica se o intervalo de confiança de 95% contém ou não o verdadeiro valor do parâmetro.

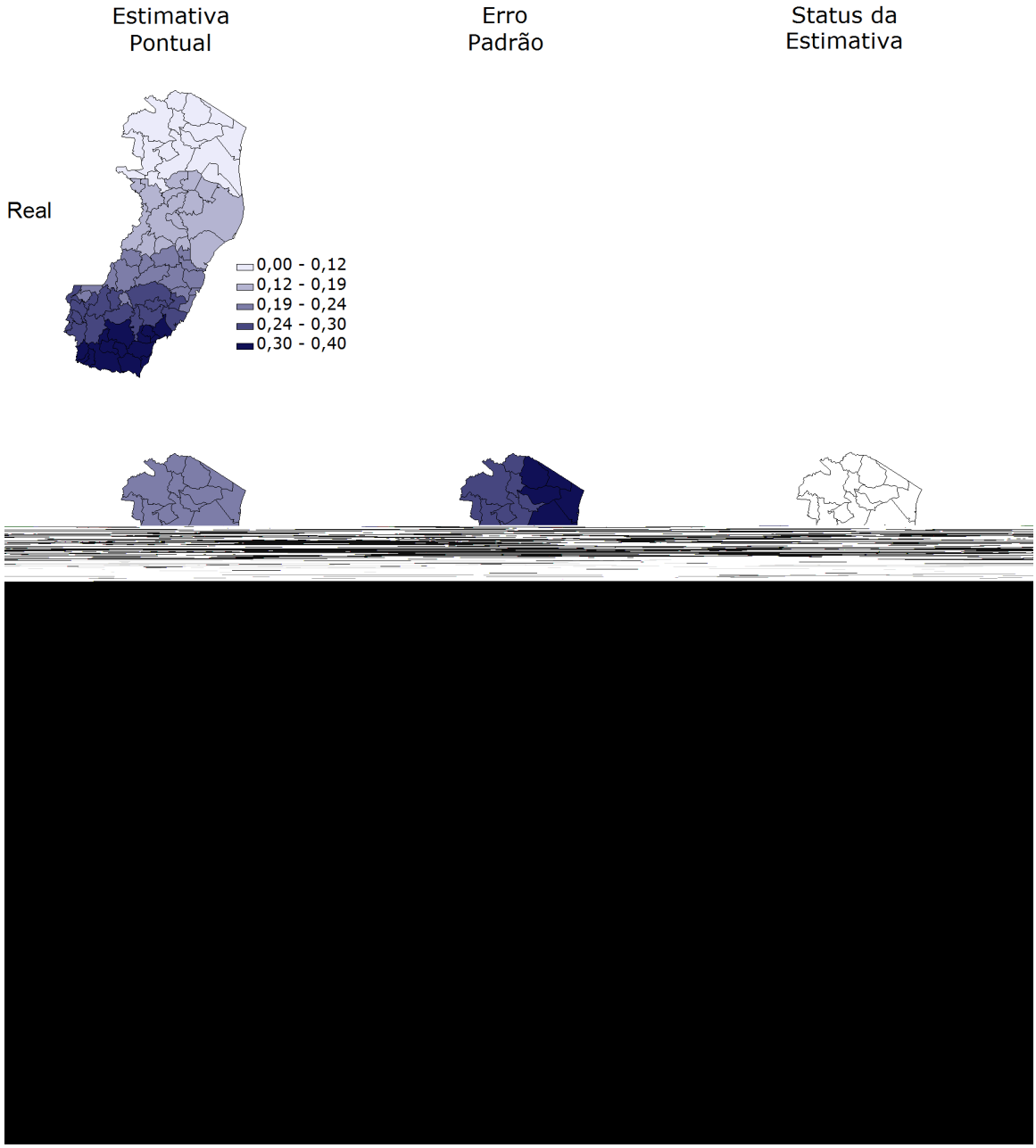

Figura 4.4: Superfícies reais e estimadas, erro padrão e status da estimativa do parâmetro α (intervalo de confiança de 95% é inferior (cor azul), ou superior (cor vermelha), ao valor real)

Observe que, como a RPGP não tem o parâmetro α, ela não está apresentada na Figura 4.4. No entanto, é possível considerar $\alpha = 0$ para a RPGP, visto que a distribuição Binomial Negativa com $\alpha \to 0$ converge para a de Poisson. Considerando que o verdadeiro valor de α varia entre 0.07 e 0.36, fixar $\alpha = 0$ não é razoável, por isso a qualidade do ajuste da RPGP foi prejudicada (vide Tabela 4.1).

A RBNGPg estimou $\alpha = 0.4$, superestimando a superdispersão em quase todo o

estado. De fato, sabe-se que o modelo RBNGPg fornece uma estimativa viesada para α, assim como a RPGP. O erro padrão de α da RBNGPg é menor pois este parâmetro é estimado globalmente, logo é baseado em um tamanho de amostra maior.

O melhor ajuste para o parâmetro α foi feito pela RBNGP. Apesar deste modelo não ter captado corretamente a variação espacial deste parâmetro, apenas no sul do estado do Espírito Santo o intervalo de confiança não conteve seu verdadeiro valor.

Os erros padrão das estimativas de b_0, b_1 e b_2 estão ilustrados na Figura 4.5.

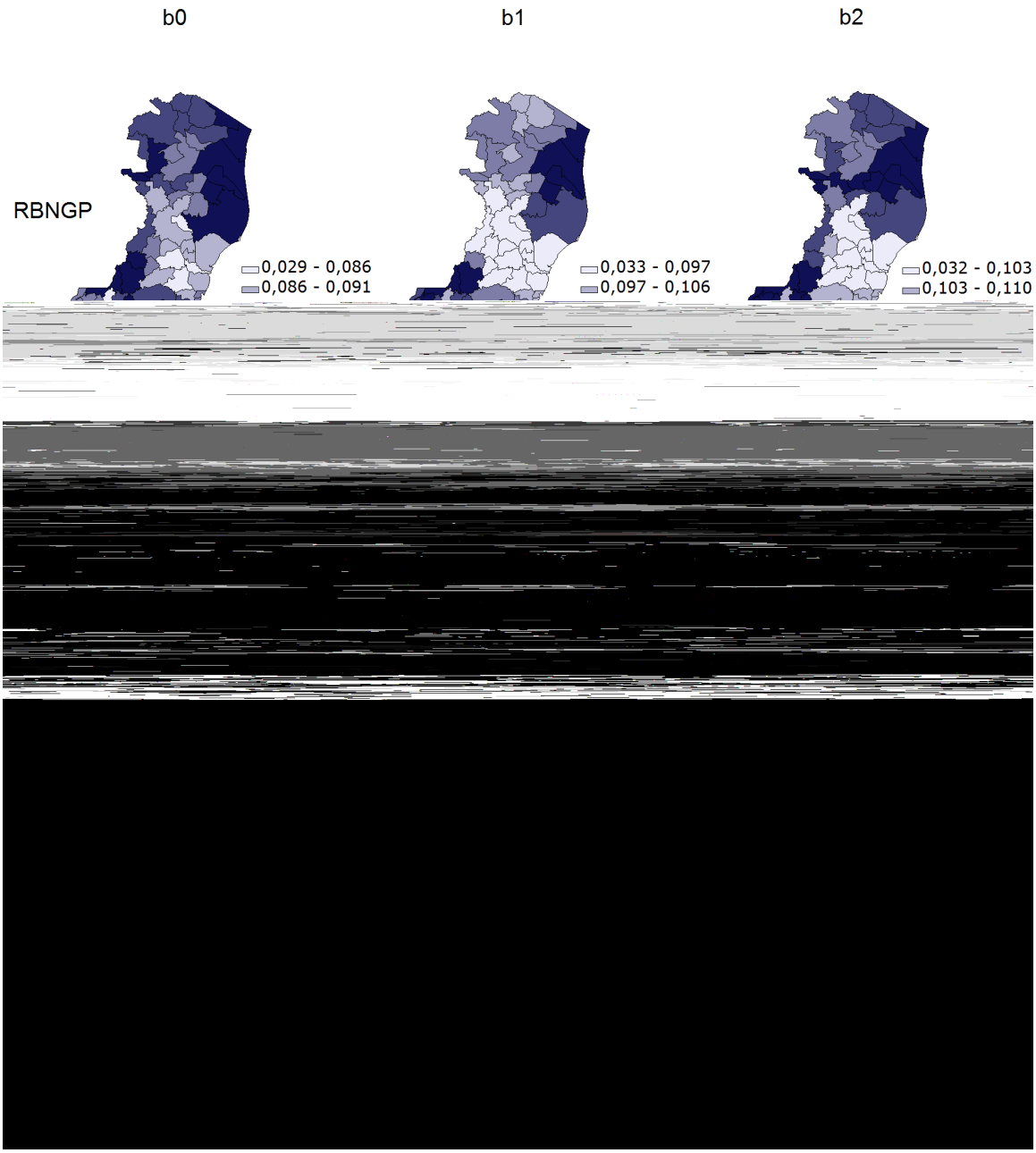

Figura 4.5: Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e b_2

Os mapas dos erros padrão da Figura 4.5 estão diretamente relacionados com os valores de α de cada modelo. Ao considerar que não há superdispersão nos dados, a RPGP estima erros padrão menores para todas as estimativas dos parâmetros, quando comparados com os erros da RBNGP. Já a RBNGPg, estimou erros maiores pois, conforme visto na Figura 4.4, superestimou a superdispersão.

A Figura 4.6 apresenta o status da estimativa de cada modelo, o qual indica se o intervalo de confiança de 95% contém ou não o verdadeiro valor de cada parâmetro.

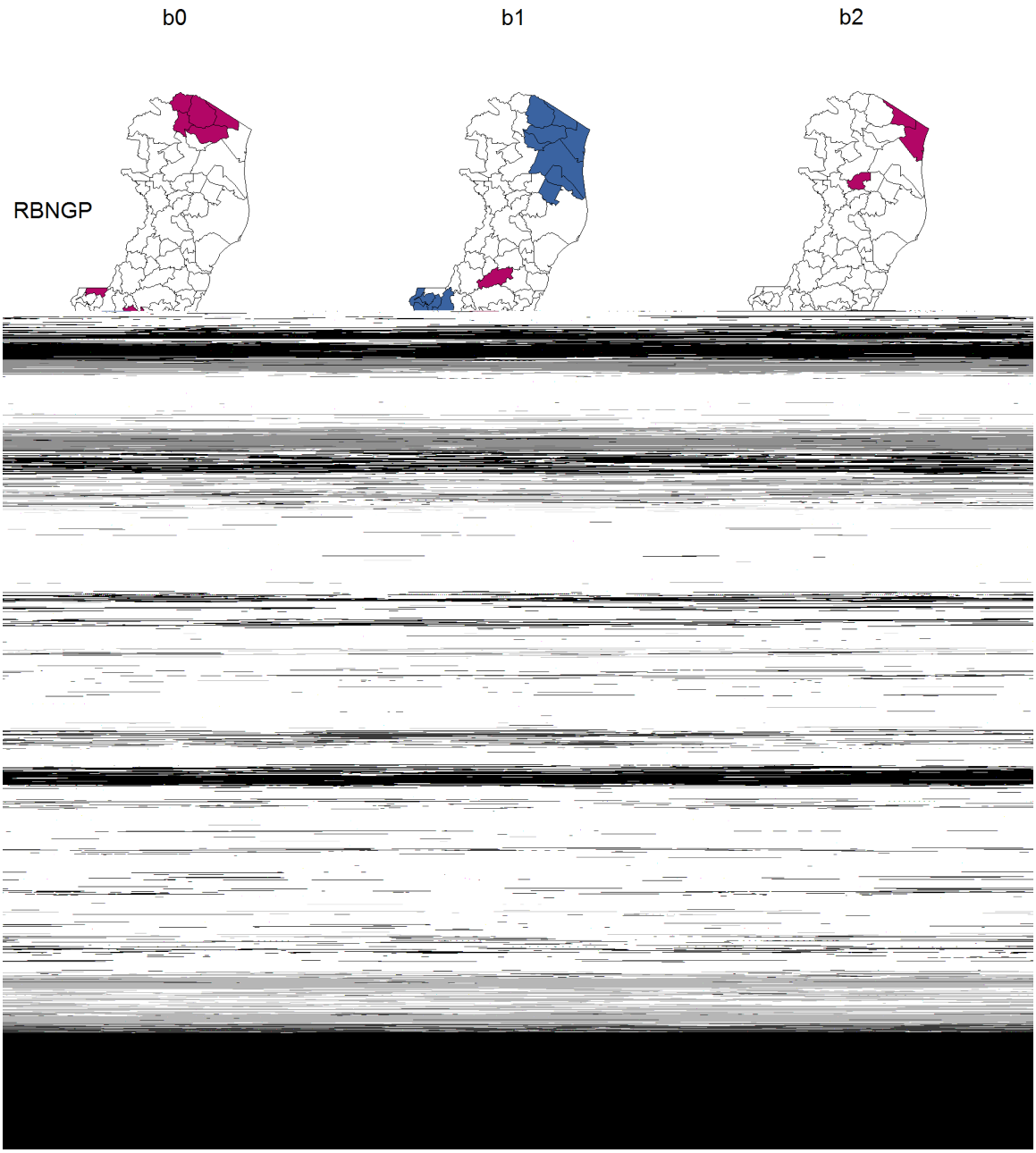

Figura 4.6: Status das estimativas dos parâmetros b_0, b_1 e b_2 (intervalo de confiança de 95% é inferior (cor azul), ou superior (cor vermelho), ao valor real)

A Figura 4.6 indica que a maior parte dos intervalos de confiança da RPGP não conteve o valor real dos parâmetros. Já para a RBNGP, aproximadamente 14% dos intervalos de b_0, 31% de b_1 e 4% de b_2 não contiveram o parâmetro real, com comportamento semelhante ao da RBNGPg. No entanto, como a estimativa do erro padrão dos modelos é pontual e vários intervalos são computados simultaneamente, espera-se que, em 5% dos casos, os intervalos não contenham o valor real do parâmetro simplesmente devido ao volume de testes conduzido. Além disso, é preciso considerar também que estes resultados são referentes a uma amostra apenas.

As estimativas dos modelos globais RBN e RP estão apresentados na Tabela 4.2, assim como as *médias* das estimativas dos parâmetros das regressões geograficamente ponderadas e os valores médios reais.

Tabela 4.2: Comparação das médias das estimativas dos parâmetros

Modelo	b_0	b_1	b_2	α
RP	4.43	0.17	0.12	-
RBN	4.43	0.19	0.13	0.40
RPGP	4.25	0.24	0.25	-
RBNGPg	4.36	0.17	0.19	0.40
RBNGP	4.33	0.19	0.22	0.18
Real	4.30	0.36	0.17	0.21

Analisando a Tabela 4.2 e a faixa de variação de cada parâmetro da Figura 4.3, conclui-se que as estimativas dos modelos globais (com exceção do α) refletem o comportamento *médio* dos parâmetros do modelo. É importante ressaltar que esta relação média pode esconder diferenças locais importantes nas relações entre as variáveis, podendo até mesmo não ser representativa da situação em nenhuma parte da região em estudo. Portanto, é importante avaliar se a hipótese de estacionariedade espacial é satisfeita para a utilização dos modelos globais de regressão. Para a RBNGP, o teste de permutação rejeita a hipótese de estacionariedade espacial para o parâmetro b_0, encontrando um p-valor de 0.1% utilizando $m = 999$ repetições.

4.2.2 Tamanho de amostra $n = 504$

Com o intuito de avaliar se os modelos propostos conseguem captar as variações espaciais dos parâmetros de forma mais precisa, aumentou-se o tamanho de amos-

tra para $n = 504$ (Figura 4.7), realizando-se novos ajustes com base no kernel bi-quadrático adaptativo. A comparação dos modelos está apresentada na Tabela 4.3.

Tabela 4.3: Comparação entre modelos

Modelo	b	Par.	$L(\boldsymbol{\beta}, \boldsymbol{\alpha})$	AICc
RP	-	3	-17698.7	35403.4
RPGP	10	316.1	-2779.7	7264.2
RBN	-	4	-2663.7	5335.4
RBNGPg	184	21.1	-2479.2	5002.5
RBNGP	109	-	-2315.3	-

Analisando a Tabela 4.3, nota-se que a RBNGPg é o modelo com menor AICc e a RBNGP apresenta a maior log-verossimilhança. Portanto, espera-se ajustes melhores para estes modelos. Além disso, percebe-se que a quantidade efetiva de parâmetros da RPGP é extremamente alta. A explicação para este fato se deve ao pequeno parâmetro de suavização estimado para a RPGP. Note que, apesar do tamanho de amostra total ser $n = 504$, as regressões locais para a RPGP são feitas apenas com base nos 10 vizinhos mais próximos, ou seja, utilizam somente 2% dos dados totais. Consequentemente, espera-se superfícies estimadas pouco suaves para este modelo. A Figura 4.8 apresenta a busca do b ótimo da RPGP pelo algoritmo da divisão áurea.

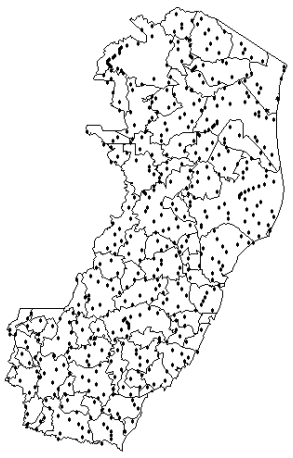

Figura 4.7: Amostra

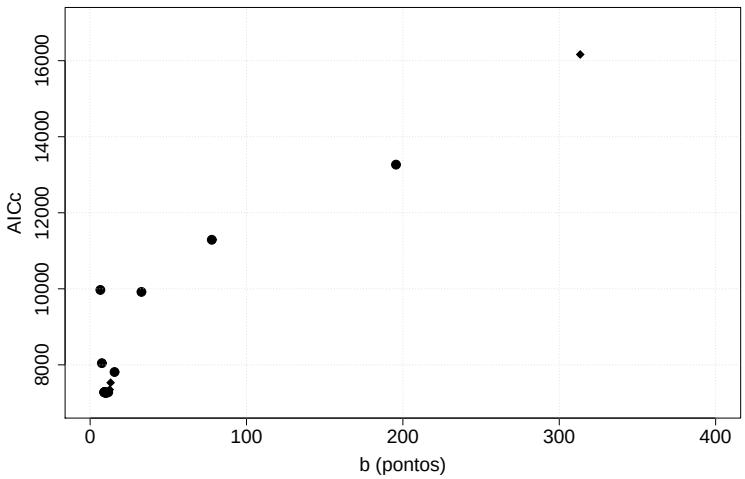

Figura 4.8: Parâmetro de suavização b da RPGP

A Figura 4.9 apresenta as superfícies reais e estimadas dos parâmetros b_0, b_1 e b_2. A fim de explorar as funcionalidades da regressão geograficamente ponderada, a estimação dos parâmetros foi feita considerando um *grid* com 17.636 pontos.

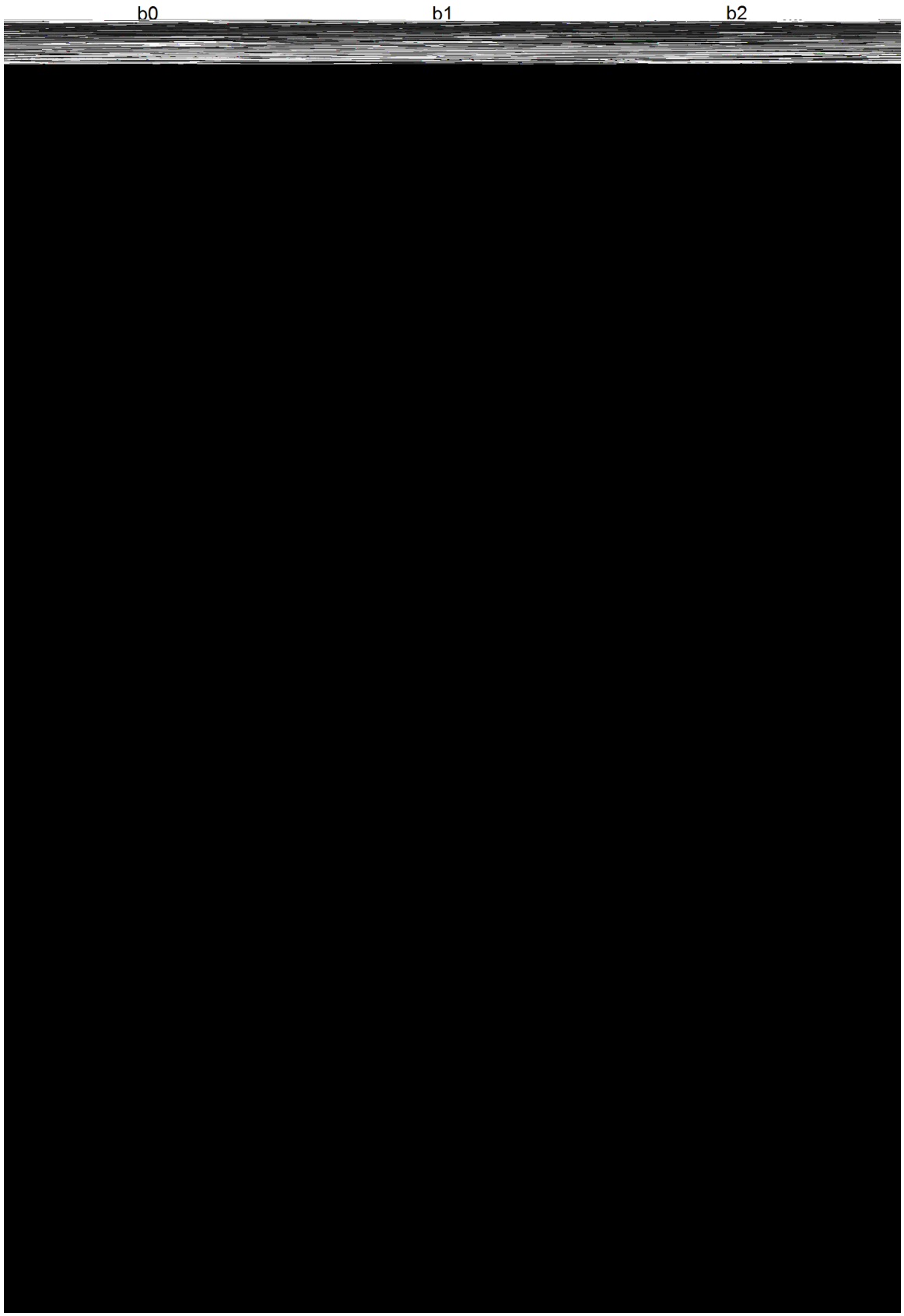

Figura 4.9: Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e b_2

Na Figura 4.9, as escalas dos mapas da RPGP foram truncadas nos valores máximo e mínimo dos parâmetros estimados pelos outros modelos. Sendo assim, para a RPGP, o azul mais escuro (mais claro) representa valores maiores (menores) ou iguais ao indicado no limite superior (inferior) da escala. Dessa forma, foi possível construir os mapas do mesmo parâmetro na mesma escala, sem que as estimativas extremas da RPGP interferissem na visualização dos resultados.

Conforme já esperado, as superfícies estimadas para os parâmetros da RPGP são muito irregulares. Além disso, para o parâmetro b_2, a RPGP não conseguiu captar nem a forma da variação, nem a magnitude das estimativas. Observe que sua superfície nos indica que quase todas suas estimativas ficaram fora dos limites da escala (mapa preenchido essencialmente pelos tons de azul mais escuro e mais claro).

Já os ajustes dos modelos RBNGPg e RBNGP ficaram muito parecidos com os mapas reais, sugerindo que estes modelos são adequados para modelar dados espaciais de contagem não estacionários e com superdispersão.

É importante ressaltar que os parâmetros foram estimados utilizando um *grid* de $N = 17.636$ pontos, por isso as superfícies da Figura 4.9 apresentam esta definição. No entanto, os ajustes foram feitos com base na amostra de $n = 504$ pontos. Os parâmetros de suavização utilizados para o ajuste com o *grid* foram os indicados na Tabela 4.3, os quais foram encontrados com base na amostra de 504 pontos, conforme explicado na Seção 2.6.1.

Os resultados do ajuste para o parâmetro α estão indicados na Figura 4.10. Para não interferir a visualização dos outros mapas, o valor de α estimado pela RBNGPg, $\hat{\alpha} = 0.57$, não foi incluído na escala. Assim, seu mapa foi colorido na cor azul escura com base na mesma idéia usada anteriormente de que esta cor, quando feita a ressalva, representa valores maiores ou iguais ao limite superior da escala.

Observe que, novamente, a RBNGPg superestimou o parâmetro de superdispersão do modelo. Além disso, o intervalo de confiança não conteve o verdadeiro valor de α para nenhum ponto. Já a estimativa pontual da RBNGP conseguiu captar tanto a forma da variação espacial, com valores crescendo na direção norte-sul, quanto a magnitude do parâmetro. Mais ainda, o status da estimativa indica que todos os intevalos de confiança de 95% contiveram o verdadeiro valor de α.

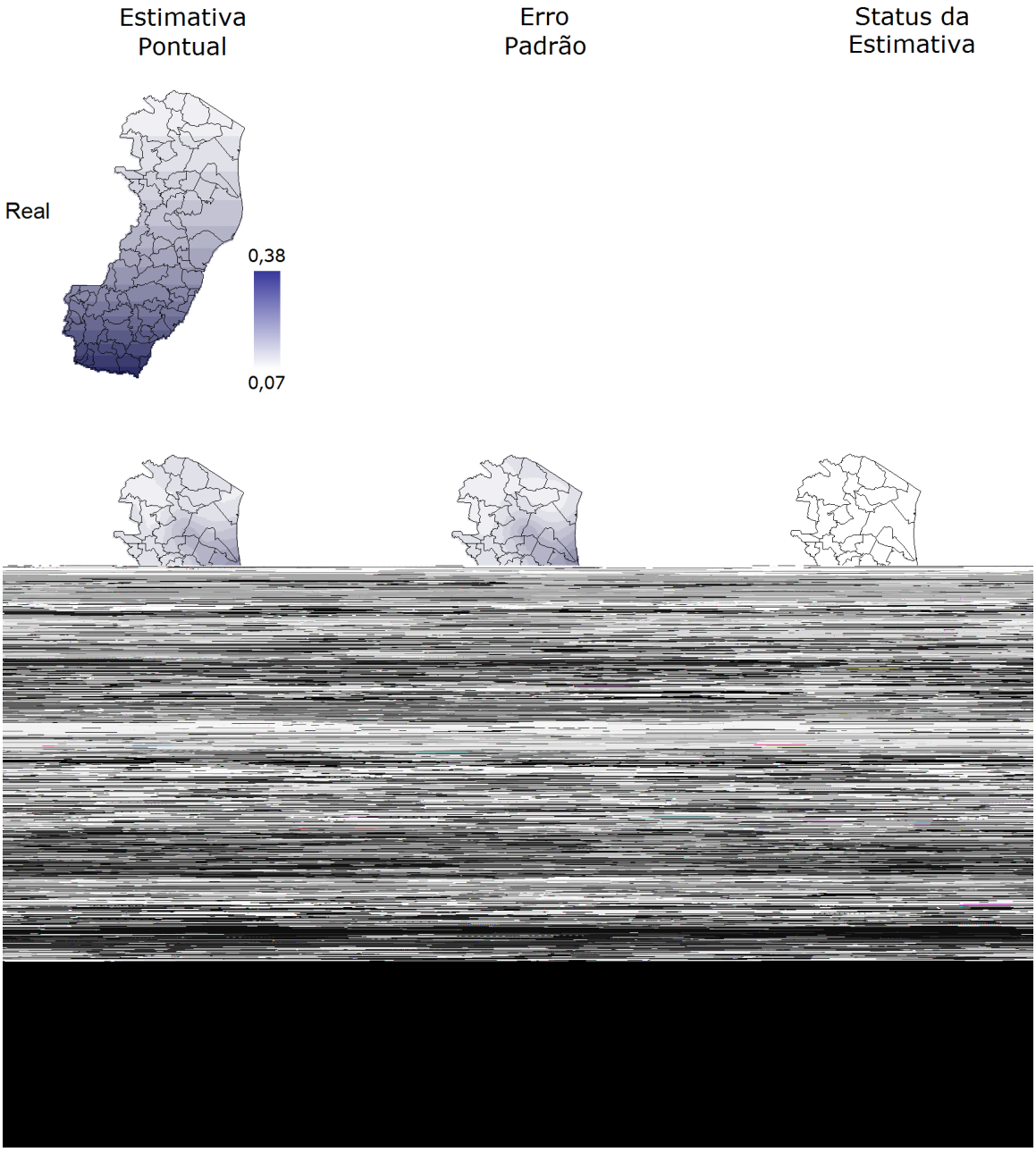

Figura 4.10: Superfícies reais e estimadas, erro padrão e status da estimativa do parâmetro α (intervalo de confiança de 95% é inferior (cor azul), ou superior (cor vermelho), ao valor real)

Os erros padrão das estimativas de b_0, b_1 e b_2 estão apresentados na Figura 4.11. Mais uma vez, os mapas da RPGP tiveram suas escalas truncadas para não comprometer a visualização dos demais.

Apesar da RPGP considerar que não há superdispersão, o que, a princípio, causaria uma redução nos erros padrão das estimativas, o fato do tamanho de amostra das regressões locais ser muito inferior ao dos outros modelos (vide Tabela 4.3) provocou

um aumento nestes erros.

Já para a RBNGPg, os erros padrão são maiores devido à superestimação do parâmetro α. Por fim, observe que os erros da RBNGP acompanham o padrão espacial da estimativa do seu parâmetro de superdispersão (vide Figura 4.10).

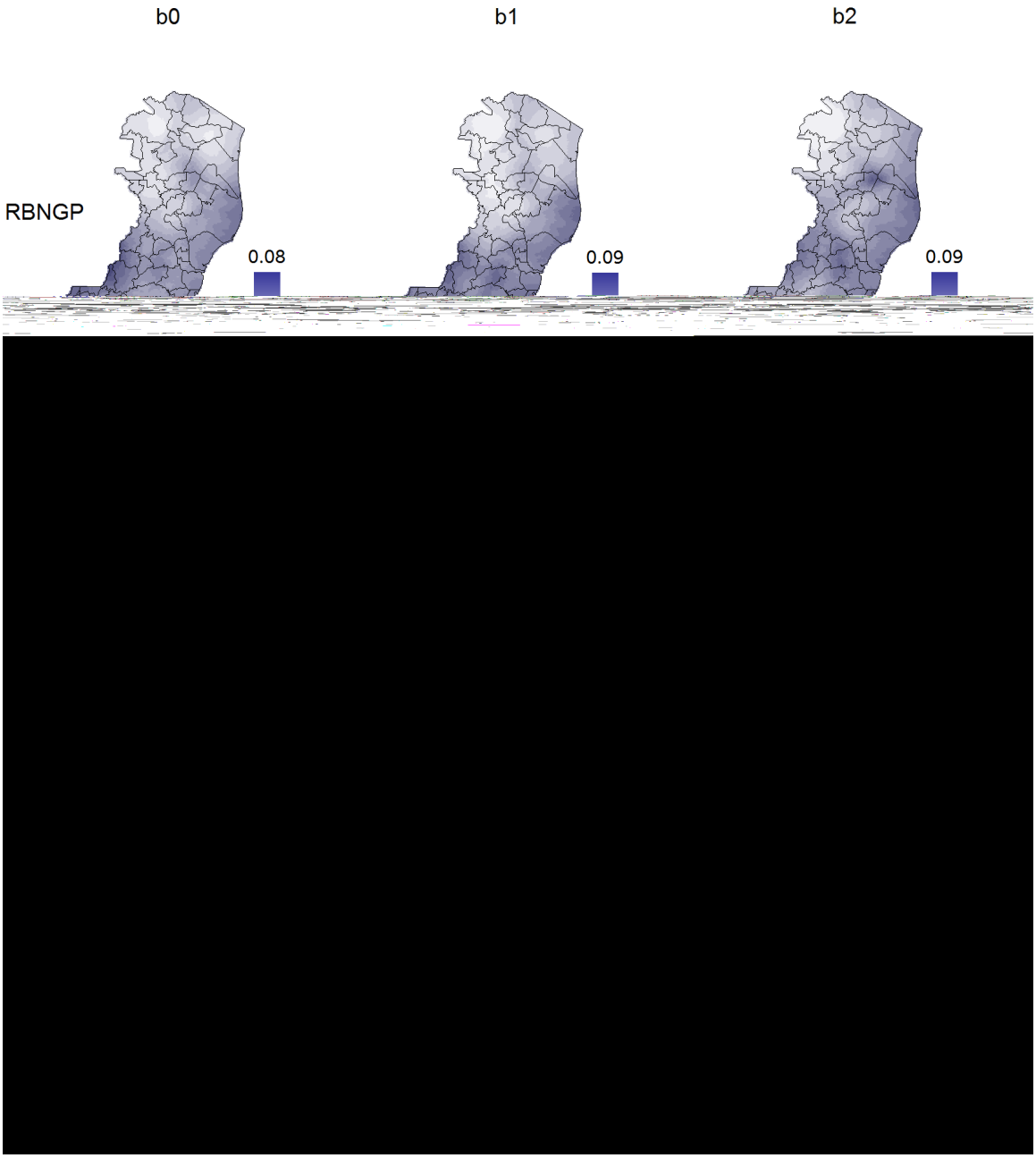

Figura 4.11: Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e b_2

A Figura 4.12 apresenta o status da estimativa de cada modelo, indicando na cor vermelha (azul) os intervalos de confiança que ficaram acima (abaixo) do verdadeiro valor do parâmetro.

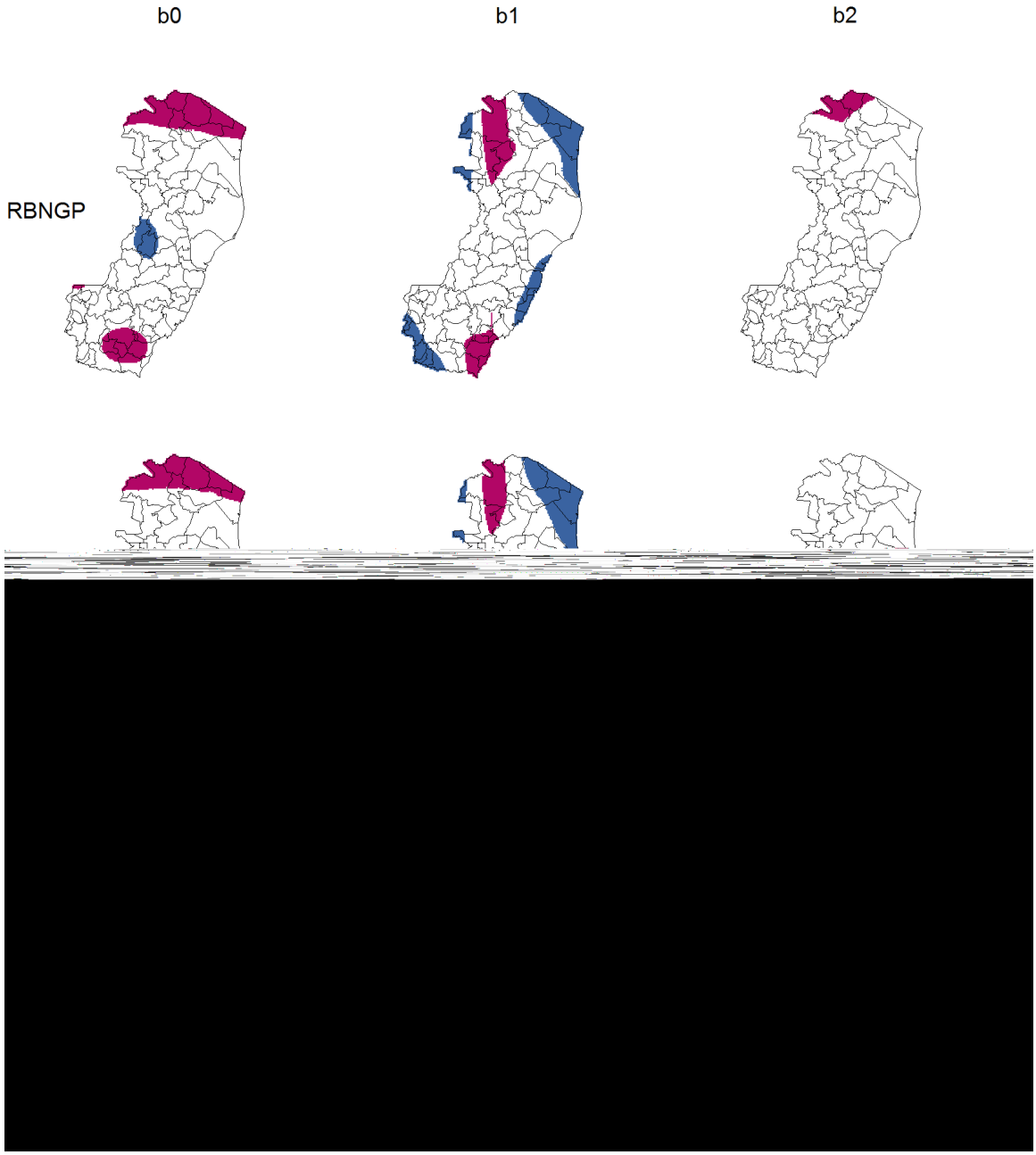

Figura 4.12: Status das estimativas dos parâmetros b_0, b_1 e b_2 (intervalo de confiança de 95% é inferior (cor azul), ou superior (cor vermelho), ao valor real)

Analisando a Figura 4.12 conclui-se que, de fato, a RPGP não se ajustou bem a este conjunto de dados. Observa-se também que os mapas da RBNGP e da RBNGPg são bem parecidos, devido às semelhanças das suas estimativas para os parâmetros β. Assim, considerando a RBNGP, 18% dos intervalos de b_0, 25% de b_1 e 3% de b_2 não contiveram o verdadeiro valor dos parâmetros.

É importante lembrar que os erros padrão de todos os modelos de regressão geograficamente ponderada são calculados com base em uma estimativa global de variância

para a variável dependente ajustada **z** (vide Seção 2.6.1). Nesta dissertação, testou-se também utilizar os erros padrão das próprias regressões locais, a fim de avaliar as diferenças dos dois métodos. Os resultados obtidos foram erros, em geral, de 20% a 30% menores utilizando a estimativa global de variância. No entanto, Fotheringham et al. (2002) comentam que pouca diferença foi observada entre as duas formas de estimação nos seus trabalhos.

Sendo assim, sugerimos, para trabalhos futuros, uma análise, por meio de testes de simulação para diferentes amostras e cenários, das diferenças entre os dois métodos de estimação do erro padrão. Além disso, seria interessante também verificar se os intervalos de confiança assim gerados são, de fato, de 95%. Com base nos resultados de apenas uma amostra não podemos confirmar esta afirmação.

A fim de testar a estacionariedade espacial, foi realizado o teste não paramétrico de aleatorização com base nos 504 pontos da amostra e $m = 999$ repetições. A hipótese nula de estacionariedade espacial foi rejeitada para b_0 e b_1, com p-valor de 0.1% para ambos. Conforme explicado na Seção 3.3, o fato do teste não rejeitar a hipótese de estacionariedade para o parâmetro α, não implica que o modelo RBNGPg é mais indicado. Lembre-se que sua estimativa de α é viesada pois considera que *todos* os parâmetros são estacionários.

Conclui-se, então, que o modelo RBNGP proposto nesta dissertação conseguiu ajustar de forma satisfatória o conjunto de dados simulados proveniente da distribuição Binomial Negativa espacial, com a ressalva de que sua complexidade ainda deve ser avaliada pois sua quantidade de parâmetros é desconhecida. Além disso, a RBNGPg também forneceu um ajuste razoável, apesar da sua estimativa viesada do parâmetro de superdispersão. Já a RPGP mostrou-se realmente inadequada para modelar dados espaciais com superdispersão. Por fim, observou-se que os modelos globais fornecem estimativas médias para os parâmetros, não sendo capazes de revelar as peculiaridades locais.

4.3 Simulação da RPGP

Nesta seção será simulado um conjunto de dados proveniente da distribuição de Poisson com dependência espacial. O intuito é verificar se o modelo RBNGP também

é capaz de ajustar a esses dados, visto que a distribuição de Poisson é um caso especial da Binomial Negativa.

As mesmas equações teóricas para b_0 e b_1 do estudo de caso anterior foram utilizadas aqui. No entanto, além do parâmetro α, que naturalmente não aparece na RPGP, foi retirado também o parâmetro b_2, a fim de tornar o exemplo menos extenso. O tamanho de amostra utilizado foi $n = 77$, referenciando-se aos 77 municípios do estado do Espírito Santo. A variável independente x_{j1} foi simulada da distribuição Normal$(0, 1)$. Sendo assim, tem-se que

$$y_j \sim \text{Poisson}\left[\ \exp\left\{b_0(u_j, v_j) + b_1(u_j, v_j)x_{j1}\right\}\ \right]\ , \tag{4.3}$$

$$b_0(u_j, v_j) = 0.000005[abs(u_j)]^3 + 0.0005[abs(v_j)]^3\ ,$$

$$b_1(u_j, v_j) = \sqrt{[u_j - \overline{u}_j]^2 \times [v_j - \overline{v}_j]^2}\ . \tag{4.4}$$

A análise exploratória, análoga ao do estudo de caso anterior, também indicou que os dados apresentam dependência e não estacionariedade espacial. O índice de Moran global foi $I = 0.33$ e os índices de Moran locais foram significativos para 10 municípios, com dependências do tipo Alto-Alto, Baixo-Baixo e Baixo-Alto.

A busca ótima do parâmetro de suavização dos modelos de regressão geograficamente ponderada foi efetuada utilizando o kernel biquadrático adaptativo, opção 6 da Seção 2.5.2. A título de ilustração, está apresentado, na Figura 4.13, o resultado do algoritmo da divisão áurea na minimização do CV para a RBNGP.

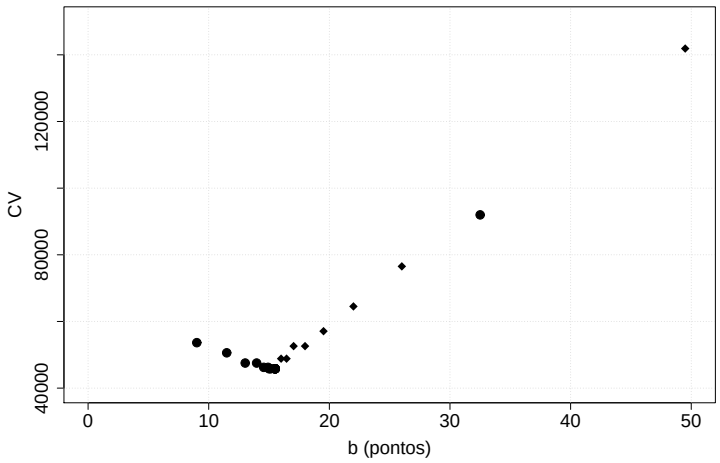

Figura 4.13: Parâmetro de suavização b da RBNGP que minimiza o CV

Sendo assim, para a RBNGP, o valor ótimo do parâmetro de suavização é $b = 15$. Utilizando os valores de b assim encontrados, as regressões geograficamente ponderadas foram realizadas e os resultados da qualidade dos ajustes estão indicados na Tabela 4.4.

Tabela 4.4: Comparação entre modelos

Modelo	b	Par.	$L(\boldsymbol{\beta}, \boldsymbol{\alpha})$	AICc
RP	-	2	-1515.6	3035.3
RPGP	15	25.5	-272.8	623.5
RBN	-	3	-401.7	809.7
RBNGPg	48	7.8	-368.3	754.1
RBNGP	15	-	-282.4	-

Conforme esperado, a RPGP apresenta as melhores medidas de qualidade do ajuste, com menor AICc e maior log-verossimilhança. No entanto, note que a log-verossimilhança da RBNGP não difere muito da RPGP, consequentemente é também um modelo candidato. Já as medidas da RBNGPg são consideravelmente piores.

Além disso, é interessante observar que, apesar dos dados serem Poisson, a regressão de Poisson não se ajustou bem, visto que os dados são espacialmente dependentes. Até mesmo a regressão Binomial Negativa apresentou um ajuste melhor, interpretando o conjunto de regressões de Poisson como uma Binomial Negativa com superdispersão.

As superfícies reais e estimadas dos parâmetros b_0, b_1 e α estão apresentadas na Figura 4.14. Os modelos real e RPGP não apresentam o parâmetro α, no entanto, para comparar com os modelos da distribuição Binomial Negativa, podemos considerar que, para eles, $\alpha = 0$.

A partir da Figura 4.14, tem-se que os três modelos ajustados recuperaram, de forma satisfatória, o intercepto do modelo. Com relação ao parâmetro b_1, nota-se que a RBNGPg não ajustou tão bem quanto a RPGP e a RBNGP, que, por sinal, tiveram ajustes muito parecidos. Este fato não foi apenas uma coincidência, observe que os valores estimados para o parâmetro α da RBNGP foram próximos de zero, e a distribuição Binomial Negativa com α próximo de zero se aproxima da de Poisson. Já a RBNGPg estimou $\alpha = 0.325$, sendo muito diferente do valor teórico zero.

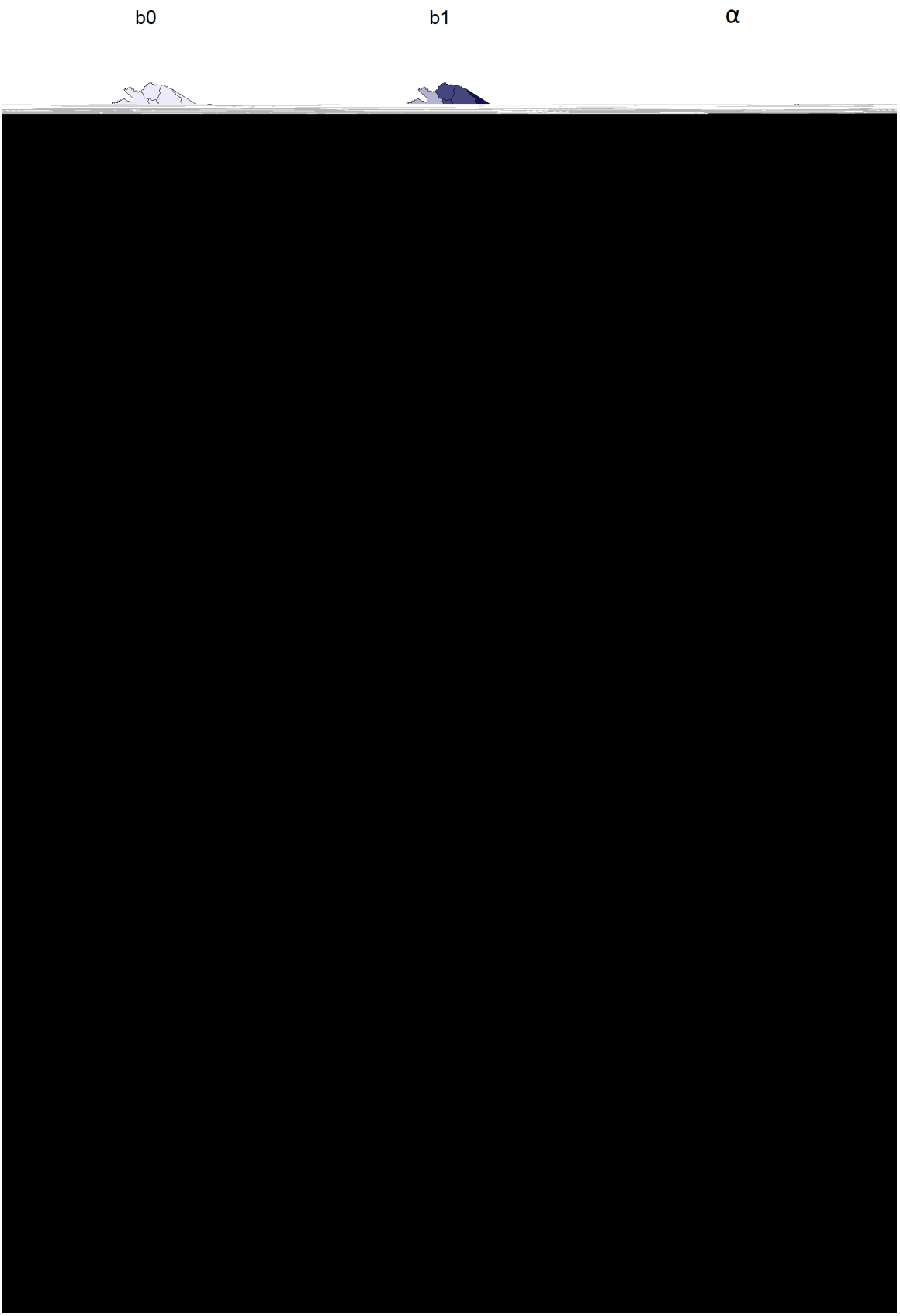

Figura 4.14: Comparação das superfícies reais e estimadas dos parâmetros b_0, b_1 e α

As superfícies dos erros padrão das estimativas dos parâmetros estão ilustradas na Figura 4.15. Observe que a estimação de um parâmetro de superdispersão muito maior do que o dos outros modelos elevou os erros padrão da RBNGPg. Os menores erros são estimados pela RPGP, na qual a superdispersão é identicamente nula.

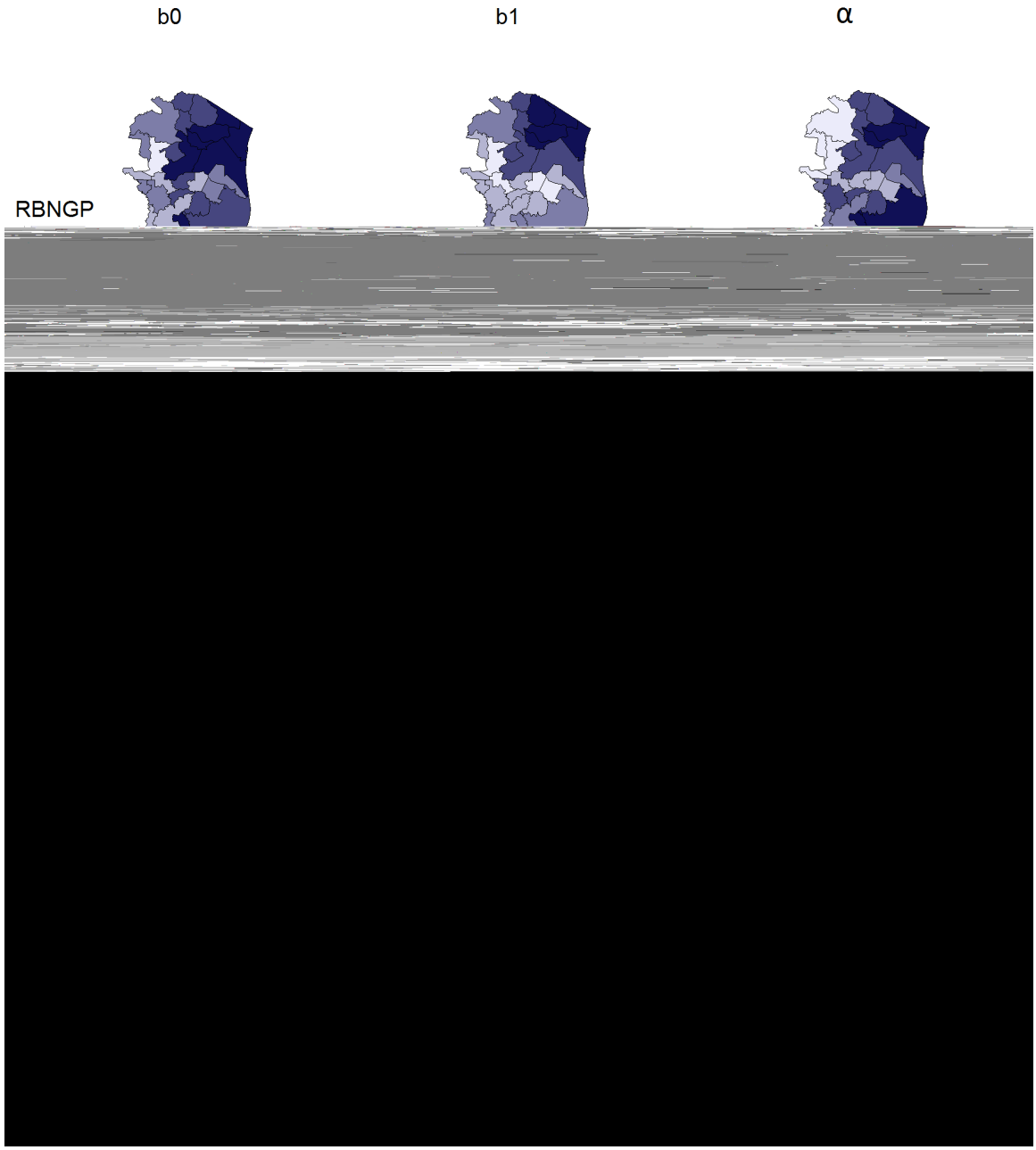

Figura 4.15: Superfícies dos erros padrão das estimativas dos parâmetros b_0, b_1 e α

Com base nos erros padrão da Figura 4.15, foram construídos intervalos de confiança de 95% para os parâmetros do modelo. A localização relativa destes intervalos com respeito aos valores teóricos dos parâmetros está apresentada na Figura 4.16.

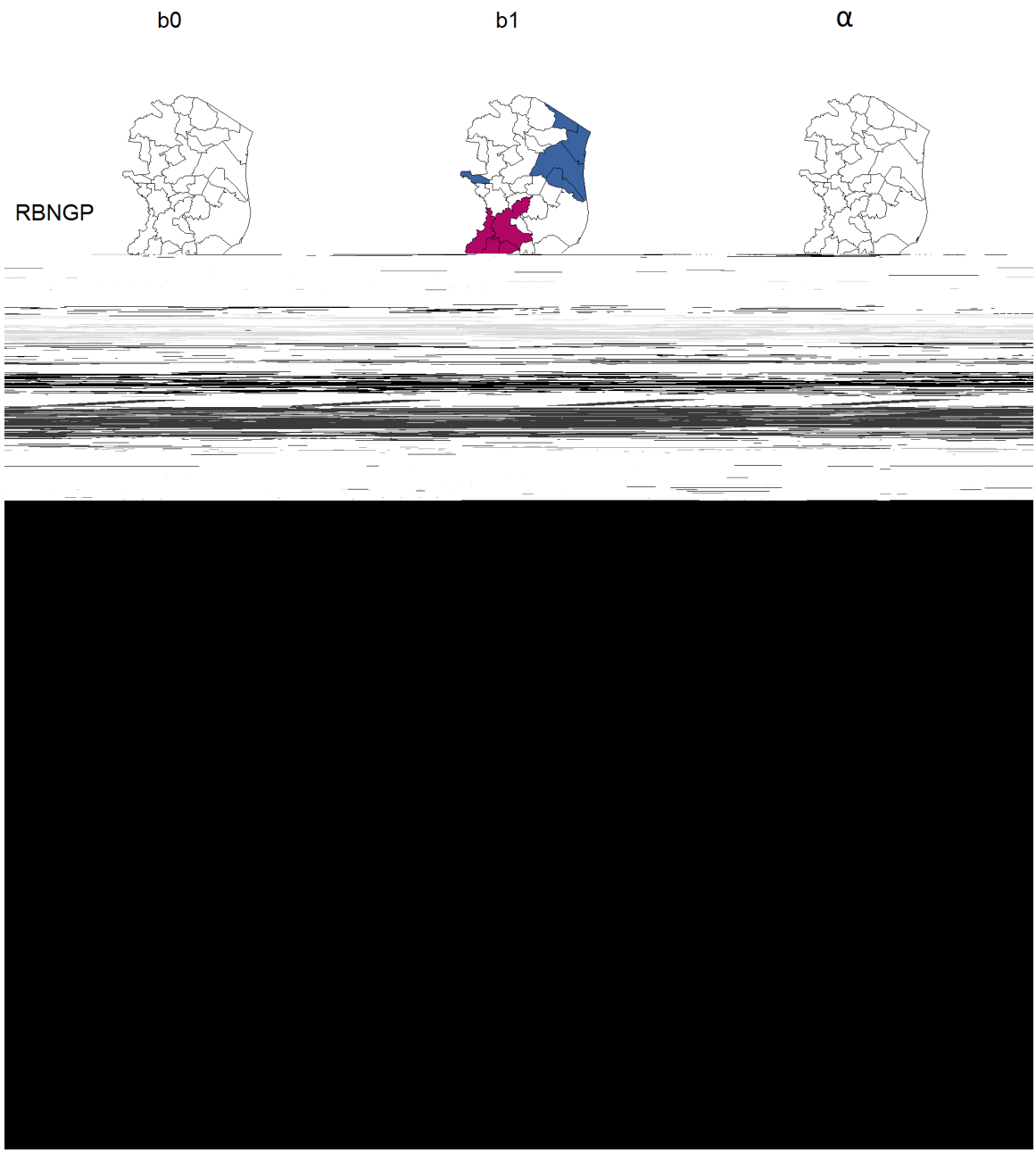

Figura 4.16: Status das estimativas dos parâmetros b_0, b_1 e α (intervalo de confiança de 95% é inferior (cor azul), ou superior (cor vermelho), ao valor real)

A Figura 4.16 apresenta um resultado muito importante. O valor $\alpha = 0$ está presente em todos os intervalos de confiança de 95% da RBNGP. Consequentemente, a RBNGP não só estimou valores pequenos para α, mas também não rejeitou a hipótese de que este parâmetro é, na verdade, zero. Portanto, tem-se um modelo robusto, capaz inclusive de identificar que o conjunto de dados é não estacionário, mas sem superdispersão.

A não estacionariedade espacial também foi confirmada pelo teste de aleatorização

para a RBNGP com $m = 999$ repetições, o qual rejeitou a hipótese nula para b_0, cujo p-valor encontrado foi de 0.1%.

A RBNGPg não tem essa habilidade de discernir entre a não estacionariedade e a superdispersão. Como o α é estimado considerando que todos os parâmetros do modelo são constantes, tem-se uma visão global que confunde os dois conceitos. A consequência é uma estimativa viesada para α. Sendo assim, a RBNGPg não é adequada para modelar dados de contagem sem superdispersão.

Na macro %gwnbr desenvolvida, cujo códico está no Apêndice, criou-se a possibilidade do analista fornecer externamente um valor para o parâmetro α ao realizar a RBNGPg. Esta flexibilidade é especialmente útil para esta situação, na qual verifica-se, pela RBNGP, que $\alpha = 0$. Assim, realiza-se a RBNGPg determinando que α não deve ser estimado, e sim fixo igual a zero. Como o parâmetro da Binomial Negativa não pode ser identicamente nulo, a macro utiliza o valor mínimo de 10^{-8}. Assim, tem-se a aproximação da RPGP por meio da RBNGPg.

Para este estudo de caso, ajustando a RBNGPg fixando $\alpha = 0$, tem-se um ajuste idêntico ao da RPGP para as estimativas dos parâmetros até, pelo menos, a sexta casa decimal. As medidas de qualidade do ajuste (log-verossimilhança, desvio, número efetivo de parâmetros, AICc e BIC) também são iguais até a terceira casa decimal.

Conclui-se, então, que o modelo RBNGP é capaz de modelar dados não estacionários também provenientes da distribuição de Poisson. Além disso, tem-se que a RBNGPg com $\alpha = 0$ substitui a RPGP, visto que fornece uma aproximação muito satisfatória.

4.4 Simulação da Regressão Global

No estudo de caso anterior, verificou-se que a regressão de Poisson geograficamente ponderada é um caso particular da regressão Binomial Negativa geograficamente ponderada e que o modelo RBNGP é capaz de identificar esta situação. O objetivo deste estudo de caso é fazer uma análise semelhante considerando agora as regressões globais Binomial Negativa e Poisson.

Assim, uma amostra de $n = 77$ observações independentes foi simulada da distribuição a seguir

$$y_j \sim \text{BN}\left[\, \exp\left\{1 + 0.5x_{j1}\right\}\, ,\ 1/3\, \right]\ , \tag{4.5}$$

onde x_{j1} é Normal$(0, 1)$.

Apesar de não haver dependência espacial nos dados simulados, o modelo RBNGP foi ajustado a fim de verificar se ele tem habilidade de identificar a estacionariedade dos parâmetros. A busca ótima do parâmetro de suavização foi efetuada utilizando o *kernel* gaussiano, opção 2 da Seção 2.5.2. A Figura 4.17 apresenta o resultado do algoritmo da divisão áurea na minimização do CV para a RBNGP.

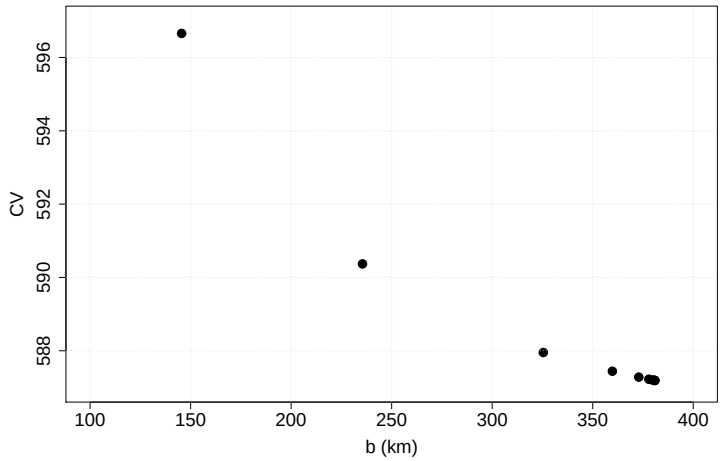

Figura 4.17: Parâmetro de suavização b da RBNGP que minimiza o CV

A partir da Figura 4.17, tem-se que o valor ótimo do parâmetro de suavização é $b = 380.9$ quilômetros, que é a maior distância entre os centros dos municípios do estado do Espírito Santo. Sendo assim, o modelo RBNGP sugere que as regressões locais sejam realizadas incluindo-se todos os pontos observados.

Consequentemente, se a função de ponderação escolhida fosse binária, como as opções 1 e 4 da Seção 2.5.2, a matriz de pesos $\mathbf{W}(i)$ seria a Identidade e, matematicamente, teria-se de volta o modelo de regressão Binomial Negativa global. Utilizando uma função de ponderação contínua, como o *kernel* gaussiano, os resultados da RBNGP não são idênticos, mas próximos, aos da RBN, como indicado na Tabela 4.5.

Tabela 4.5: Sumário das estimativas dos parâmetros da RBNGP e da RBN

Parâmetro	Min. RBNGP	Max. RBNGP	RBN
b_0	1.186	1.194	1.189
b_1	0.176	0.191	0.182
α	0.304	0.311	0.303

As colunas "Min. RBNGP" e "Max. RBNGP" da Tabela 4.5 indicam, respectivamente, os valores mínimo e máximo das estimativas de cada parâmetro da RBNGP. Observe que as estimativas da RBNGP variam pouco, se aproximando dos resultados da RBN. De fato, realizando o teste de aleatorização para a RBNGP, a hipótese nula de estacionariedade espacial não é rejeitada para nenhum parâmetro.

O *kernel* gaussiano permite a escolha de um parâmetro de suavização maior do que a distância máxima entre os pontos da amostra. No entanto, no algoritmo da divisão áurea desenvolvido, limitou-se a região de busca entre as distâncias mínima e máxima, visto que é a região de interesse prático da regressão geograficamente ponderada.

No entanto, para realizar a RBN com o código desenvolvido para a RBNGP é preciso aumentar o parâmetro de suavização, a fim de tornar mais lento o decaimento da função de ponderação. Ajustando a RBNGP com $b = 380900$ quilômetros, ou seja, mil vezes maior do que a distância máxima, tem-se que as estimativas dos parâmetros (pontuais e erros padrão) são idênticas aos da regressão global até, pelo menos, a sexta casa decimal. Mais precisão pode ser alcançada aumentando ainda mais o parâmetro b. Os mesmos resultados são obtidos utilizando o modelo RBNGPg, que já utiliza a estimativa de α da regressão Binomial Negativa global.

Simulando a regressão global de Poisson e ajustando pela RBNGP com *kernel* gaussiano, obtém-se novamente o parâmetro de suavização máximo do algoritmo da divisão áurea $b = 380.9$ quilômetros. Além das estimativas de b_0 e b_1 variarem muito pouco com este ajuste, o parâmetro α é estimado entre 6.7×10^{-6} e 9.9×10^{-6}, ou seja, próximo de zero. A regressão de Poisson pode ser realizada de forma mais precisa fixando $\alpha = 0$ e $b = 380900$ quilômetros no modelo RBNGPg.

Conclui-se que a RBNGP é um modelo robusto, capaz também de identificar a estacionariedade espacial e a ausência de superdispersão. Além disso, ele realiza as regressões globais Binomial Negativa e Poisson com o nível de precisão desejado.

4.5 Aplicação para a cidade de Tóquio

A RPGP foi desenvolvida por Nakaya et al. (2005) e utilizada pelos autores na modelagem da taxa de mortalidade na área metropolitana de Tóquio. Os mesmos dados também foram aplicados por Chen e Yang (2011) para validar a macro %gwglm, desenvolvida pelos autores em linguagem SAS para realizar a regressão geograficamente ponderada para as distribuições Normal, Poisson e Binomial. Chen e Yang (2011) comparam os resultados da macro %gwglm com os provenientes do *software* comercial GWR3.0, elaborado por Fotheringham e sua equipe para realizar a RGP para as distribuições supracitadas. Observe que essas ferramentas não realizam a RGP para a distribuição Binomial Negativa.

Devido à tradição desse conjunto de dados, o mesmo será aplicado aos modelos desenvolvidos nesta dissertação. A RBNGPg com $\alpha = 0$ será utilizada como aproximação da RPGP. Os resultados desse ajuste serão comparados com os detalhados no artigo Chen e Yang (2011). Além disso, o modelo RBNGP será ajustado a fim de verificar se os dados apresentam superdispersão.

A variável dependente do modelo é o número de óbitos na faixa etária de 25 a 64 anos no ano de 1990 na área metropolitana de Tóquio, a qual é dividida em 262 zonas. A variável *offset* utilizada é o número esperado de óbitos em cada zona, nessa faixa etária, ou seja, a população exposta ao risco. As covariáveis sócio-econômicas empregadas foram: proporção de trabalhadores com nível profissional ou técnico (x_1), proporção da população com mais de 65 anos (x_2), proporção de imóveis próprios (x_3) e taxa de desemprego (x_4). Os dados são do censo nacional e das estatísticas vitais do ano de 1990 do Japão.

A Tabela 4.6 apresenta as estimativas dos parâmetros do componente linear da RPGP com base na macro %gwglm, no *software* GWR3.0 (ambos resultados foram obtidos de Chen e Yang (2011)) e na macro %gwnbr proposta nesta dissertação, que utiliza o modelo RBNGPg com $\alpha \approx 0$ como aproximação da RPGP. O parâmetro de suavização ótimo é $b = 95$ pontos, encontrado utilizando o *kernel* biquadrático adaptativo e minimizando a estatística AICc.

Os resultados do ajuste da RBNGP também estão apresentados na Tabela 4.6. Neste caso foi utilizado o *kernel* gaussiano com $b = 18$ quilômetros. Para o parâmetro

de suavização ótimo $b = 15$ quilômetros, 20 pontos não convergiram devido à estimação de α próximo de zero e a dificuldade de convergência da distribuição Binomial Negativa neste caso. Conforme indicado na Figura 4.18, que ilustra a busca do parâmetro de suavização ótimo pelo algoritmo da divisão áurea, para $b = 18$ quilômetros, o CV está próximo do seu valor mínimo, não ocasionando grandes prejuízos na qualidade do ajuste.

Tabela 4.6: Sumário das estimativas dos parâmetros utilizando %gwglm (Chen e Yang, 2011), GWR3.0 e %gwnbr

Estimativas	%gwglm RPGP	GWR 3.0 RPGP	%gwnbr RPGP	%gwnbr RBNGP
b_0				
Mínimo	-0.942	-0.942	-0.942	-0.651
Q1	-0.003	-0.004	-0.003	-0.020
Mediana	0.099	0.099	0.099	0.095
Q3	0.260	0.259	0.260	0.161
Máximo	0.434	0.434	0.434	0.323
b_1				
Mínimo	-3.783	-3.783	-3.783	-4.268
Q1	-2.681	-2.682	-2.681	-2.487
Mediana	-2.506	-2.509	-2.506	-2.380
Q3	-1.791	-1.800	-1.791	-1.833
Máximo	1.627	1.627	1.627	1.074
b_2				
Mínimo	1.223	1.223	1.223	1.209
Q1	1.643	1.643	1.643	1.869
Mediana	2.079	2.069	2.079	2.066
Q3	2.448	2.437	2.448	2.401
Máximo	4.418	4.418	4.418	3.881
b_3				
Mínimo	-0.572	-0.572	-0.572	-0.675
Q1	-0.380	-0.381	-0.380	-0.347
Mediana	-0.316	-0.317	-0.316	-0.297
Q3	-0.205	-0.210	-0.205	-0.267
Máximo	0.152	0.152	0.152	0.085
b_4				
Mínimo	-0.056	-0.056	-0.056	-0.057
Q1	0.021	0.021	0.021	0.028
Mediana	0.039	0.038	0.039	0.051
Q3	0.079	0.078	0.079	0.079
Máximo	0.171	0.171	0.171	0.256

Fonte: Chen e Yang (2011) com modificações

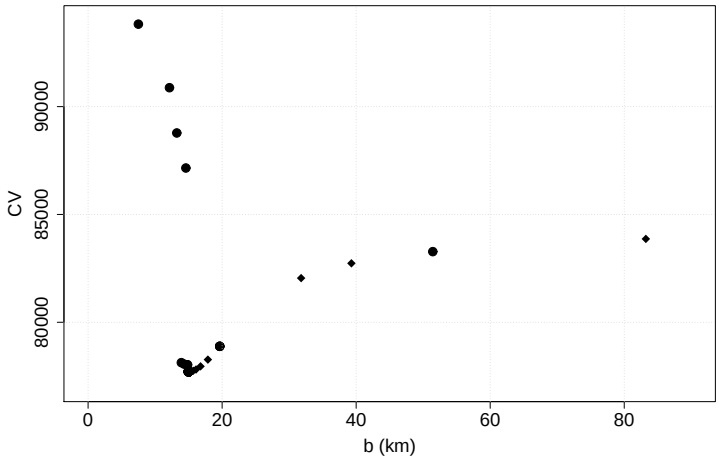

Figura 4.18: Parâmetro de suavização b da RBNGP que minimiza o CV

Comparando os resultados da Tabela 4.6 para os ajustes da Regressão de Poisson Geograficamente Ponderada, conclui-se que a aproximação realizada pela RBNGPg gerou estimativas pontuais idênticas às da macro %gwglm, considerando até a terceira casa decimal. O ajuste utilizando o *software* GWR3.0 também foi semelhante aos outros. Conclui-se, então, que a macro %gwnbr proposta nesta dissertação realiza a RPGP, com base na aproximação quando $\alpha \to 0$ da RBNGPg, de forma coerente com a proposta pelos outros autores.

No entanto, com relação às estimativas dos erros padrão, a macro %gwglm apresenta resultados diferentes, visto que esta utiliza os erros padrão provenientes das regressões locais. Lembre-se que Fotheringham et al. (2002) sugerem uma outra metodologia para este cálculo (vide Seção 2.6.1), que foi utilizada na macro %gwnbr.

A Tabela 4.7 apresenta a mediana dos erros padrão dos parâmetros estimados com as macros %gwglm, %gwnbr e pelo *software* GWR3.0.

Tabela 4.7: Medianas dos erros padrão dos parâmetros utilizando %gwglm (Chen e Yang, 2011), GWR3.0 e %gwnbr

Mediana dos erros padrão	%gwglm RPGP	GWR 3.0 RPGP	%gwnbr RPGP	%gwnbr RBNGP
b_0	0.239	0.172	0.172	0.156
b_1	0.633	0.449	0.450	0.394
b_2	0.720	0.474	0.474	0.433
b_3	0.167	0.121	0.121	0.114
b_4	0.042	0.029	0.029	0.026

Analisando a Tabela 4.7, conclui-se que as medianas dos erros padrão estimados pela macro %gwnbr estão coerentes com as provenientes do *software* GWR3.0 para a RPGP. Já as estimadas pela macro %gwnbr são, em média, 30% inferiores a essas outras. Observe que um resultado semelhante foi observado na Seção 4.2.2, na qual as duas alternativas de cálculo dos erros também foram comparadas. Já os erros padrão da RBNGP são menores pois utilizam outro parâmetro de suavização.

A Tabela 4.6 não apresenta as estimativas do parâmetro de superdipersão da RBNGP, as quais variaram entre 0.0001 e 0.01. Realizando testes de hipóteses locais para verificar se $\alpha = 0$, conclui-se que a superdispersão é significativa em apenas $5,7\%$ das zonas. No entanto, devido à realização de múltiplos testes, é natural que em 5% das zonas, aproximadamente, o resultado seja significativo mesmo na ausência de superdispersão. Além disso, as log-verossimilhanças da RPGP e da RBNGP são próximas (-985.9 e -980.7, respectivamente), indicando que as qualidades do ajuste são similares para os dois modelos. Sendo assim, conclui-se que os dados não apresentam superdispersão e, consequentemente, a RPGP é o modelo mais indicado.

Observe que a confirmação da ausência de superdispersão foi possível graças a macro %gwnbr. Outra vantagem desta macro é que foi desenvolvida em linguagem SAS/IML. Assim, levou menos de 6 segundos para fazer o ajuste da RPGP, enquanto que a %gwglm, que utiliza os procedimentos *PROC* do SAS, demorou, aproximadamente, 40 segundos. Para banco de dados maiores, esta diferença de velocidade pode ser determinante.

O objetivo deste estudo de caso foi verificar que o exemplo clássico dos dados de mortalidade de Tóquio podem ser modelados segundo a RPGP. Além disso, buscou-se mostrar que a macro %gwnbr aqui proposta realiza este ajuste e seus resultados estão de acordo com os da macro %gwglm (com a ressalva dos erros padrão) e do *software* GWR3.0. Assim, os mapas das estimativas dos parâmetros e dos erros padrão não foram realizados, pois não era objetivo interpretar o comportamento espacial dos parâmetros estimados e as relações entre as variáveis. A inclusão dos mesmos tornaria o exemplo extenso, já que se trata de 6 parâmetros, e desviaria o foco central.

4.6 Aplicação para o estado do Espírito Santo

A distribuição da frota de veículos rodoviários de carga do tipo caminhão simples no estado do Espírito Santo será analisada neste estudo de caso. Espera-se que esta variável apresente certa dependência espacial, visto que a quantidade de caminhões em cada município depende, provavelmente, das características econômicas da sua região próxima. Além disso, são dados de contagem, possivelmente não estacionários e com superdispersão.

Compreender a distribuição espacial desses veículos rodoviários de carga auxilia a administração pública na tomada de decisões, visto que esta configuração tem impacto nas necessidades de fiscalização, manutenção e ampliação das rodovias, que ainda é a principal via de transporte do país. Além disso, esta configuração espacial interessa às empresas que necessitam dos caminhões para o transporte de suas mercadorias.

As concessionárias também podem se beneficiar desta informação. A identificação de municípios com uma maior demanda represada de caminhões é determinante para a estratégia de vendas da empresa e maximização dos lucros. O uso de modelos estatísticos, em especial os modelos geograficamente ponderados, pode ser extremamente útil para lidar com o problema. O número esperado de caminhões do município pode ser explicado pelas suas características, que são traduzidas por um conjunto de covariáveis. Esta média, não observável, representa a demanda do município por caminhões com base em suas características e localização geográfica. As variações das observações em torno dessa média correspondem ao erro aleatório do modelo, composto por fatores que não foram explicitamente considerados e, principalmente, pelo descompasso entre a demanda latente e a oferta (frota observável).

É importante ressaltar que este é apenas um exemplo motivacional, a elaboração de um modelo de interesse prático extrapola o escopo dessa dissertação. Assim, apenas uma covariável será utilizada, a saber, a quantidade de estabelecimentos industriais. O objetivo não é estabelecer uma relação de causa e efeito entre o número de indústrias e o número de caminhões, e sim utilizar uma covariável associada ao porte econômico do município para melhorar a estimativa da frota esperada de caminhões.

Considerando a dependência espacial, a informação sobre os municípios vizinhos também auxilia a estimação da média. Além disso, a utilização de um modelo espacial

local permite o tratamento adequado das diferenças entre as regiões.

Municípios em que a frota observada é maior que a média estimada pelo modelo apresentam resíduos positivos, o que aponta uma saturação da frota. Por outro lado, se o número de caminhões é menor do que o valor ajustado (isto é, o valor médio para municípios com as mesmas características), temos indicação que há uma carência a ser suprida, um potencial de vendas. Assim, os municípios com resíduo mais negativo são candidatos a uma avaliação mais detalhada sobre o plano de vendas. Dessa forma, concentra-se a atenção nos lugares com maior potencial. A partir daí, outros critérios, como custo de implementação e efeito da concorrência, por exemplo, passam a ser avaliados.

A Figura 4.19 apresenta os mapas da frota de caminhões simples (variável *Frota*) e da quantidade de indústrias (variável *Indústrias*). Os quintis dessas variáveis foram utilizados para definir a escala. Os dados são do RNTRC (Registro Nacional de Transportadores Rodoviários de Carga) e do IBGE (Instituto Brasileiro de Geografia e Estatística) do ano de 2000. Eles foram utilizados por Silva (2006) na elaboração de um modelo de regressão espacial global.

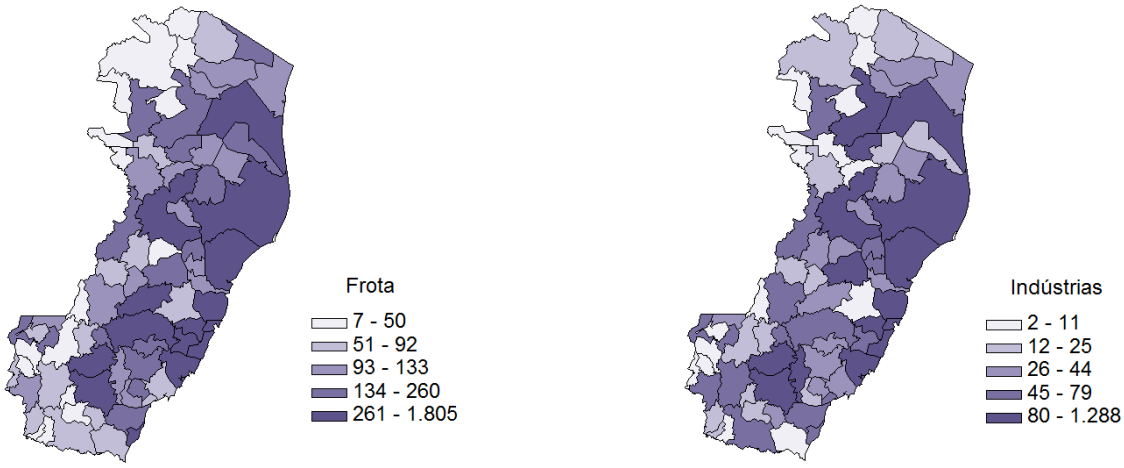

Figura 4.19: Mapa da variável dependente *Frota* e da variável independente *Indústrias*

É possível observar, a partir da Figura 4.19, que a quantidade de caminhões simples é maior na região litorânea, em especial na região sudeste do estado, onde mesmo municípios com área pequena apresentam uma frota grande de caminhões. Nota-se que, afastando-se da região costeira a concentração vai diminuindo, alcançando os menores valores na região noroeste e sul do estado. Sendo assim, conclui-se que

há indícios de que a variável *Frota* apresenta algum grau de dependência espacial. Observa-se também que a quantidade de estabelecimentos do ramo da indústria tem um comportamento semelhante ao da variável dependente.

A fim de quantificar a dependência espacial, foi calculado o índice de Moran para a variável *Frota* utilizando a matriz de proximidade espacial binária que indica se a área A_i faz fronteira com a área A_j. O valor obtido foi $I = 0.23$, conforme indicado na Figura 4.20, que ilustra o diagrama de espalhamento de Moran. Este índice caracteriza uma dependência espacial baixa com respeito a matriz **W** binária.

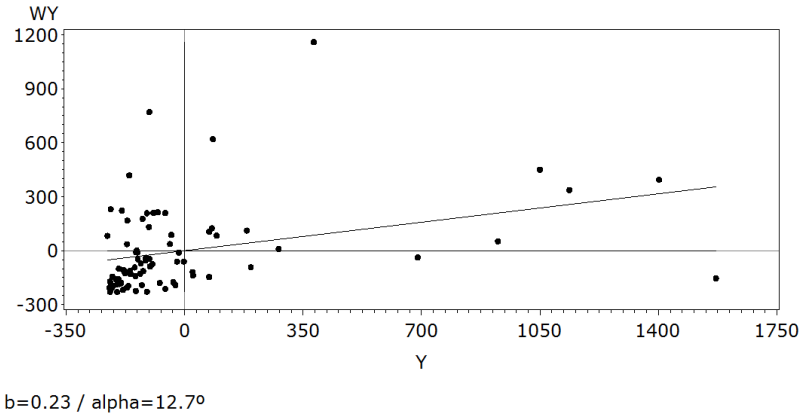

Figura 4.20: Diagrama de espalhamento de Moran

A fim de melhor explicar a dependência espacial, Silva (2006) recomenda a utilização de alguma variável não geográfica para definir a matriz de proximidade ao se trabalhar com dados de transportes como, por exemplo, a quantidade de trocas comerciais entre as unidades espaciais ou a quantidade de rodovias de ligação.

No entanto, faz-se ainda necessário verificar se a hipótese de estacionariedade espacial do índice de Moran é válida. Para isso, considere os mapas apresentados na Figura 4.21. A partir do mapa de espalhamento de Moran, nota-se novamente a polarização do litoral para o interior, com os municípios nas cores azul indicando a região de transição. Já o mapa de Moran nos indica que existem correlações locais em algumas regiões que são significativamente diferentes das demais, dando-nos indícios de não estacionariedade espacial. Consequentemente, o índice global de Moran não é adequado para caracterizar a dependência espacial. Além disso, um modelo espacial local aparenta ser mais indicado.

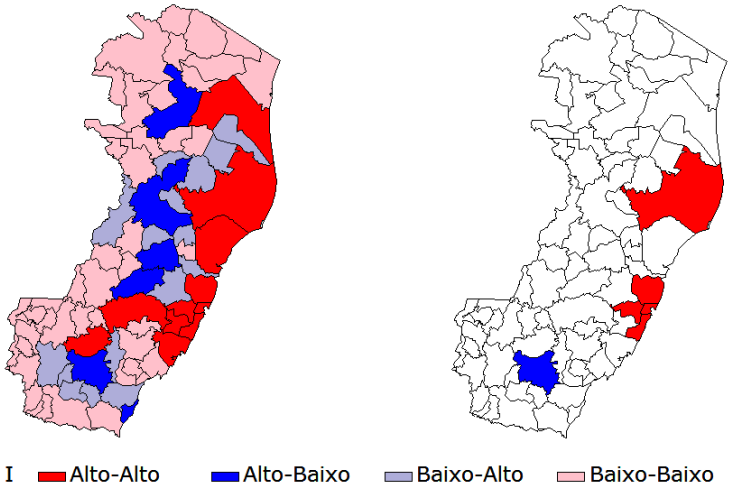

Figura 4.21: Mapa de espalhamento de Moran (esquerda) e Mapa de Moran 95% (direita)

Os modelos espaciais locais RBNGP, RBNGPg e RPGP foram ajustados, assim como as regressões RBN e RP. A Tabela 4.8 apresenta os resultados desses ajustes.

Tabela 4.8: Comparação entre modelos

Modelo	b	Par.	$L(\boldsymbol{\beta}, \boldsymbol{\alpha})$	AICc
RP	-	2	-5001.4	10006.9
RPGP	10 pontos	32.4	-1296.8	2708.2
RBN	-	3	-458.5	923.3
RBNGPg	53.07 km	8.7	-444.4	908.7
RBNGP	53.20 km	-	-440.1	-

A função de ponderação gaussiana (opção 2 da Seção 2.5.2) foi utilizada para o ajuste dos modelos RBNGP e RBNGPg. Observe que em ambos os casos o parâmetro de suavização ótimo foi aproximadamente 53 quilômetros. Já o parâmetro de suavização estimado para a RPGP com este *kernel* foi $b = 9.4$ quilômetros. No entanto, para este valor de b, os pesos da função de ponderação espacial gaussiana são praticamente nulos para áreas distantes mais de 30 quilômetros. Consequentemente, as regressões locais são feitas com base em um número pequeno de pontos (de 1 a 9).

Para contornar este problema, optou-se por utilizar o *kernel* biquadrático adaptativo (opção 5 da Seção 2.5.2). Neste caso, o limite inferior da região de busca do algoritmo da divisão áurea é escolhido como sendo o número mínimo de pontos aceitável para a realização da regressão local. Nesta dissertação, estabeleceu-se uma quantidade mínima de 10 pontos. Com isso, o AICc da RPGP foi minimizado para

$b = 10$, ou seja, no limite inferior da região de busca.

A partir da Tabela 4.8, conclui-se que a regressão de Poisson e a Regressão de Poisson Geograficamente Ponderada apresentam uma qualidade de ajuste muito inferior se comparada a dos outros modelos. A estimação de um parâmetro de suavização muito pequeno já era um indício da inadequabilidade da escolha da distribuição de Poisson. A variável frota de caminhões simples no estado do Espírito Santo apresenta superdispersão, sendo a distribuição Binomial Negativa mais indicada nessa modelagem.

A inclusão da dependência espacial nos modelos RBNGP e RBNGPg melhora a qualidade do ajuste em relação à Regressão Binomial Negativa global, diminuindo o AICc e aumentando a log-verossimilhança. Realizando o teste de não estacionariedade para a RBNGP com base em $m = 999$ repetições, tem-se p-valores iguais a 1.2%, 8.4% e 62.6% para os parâmetros β_0, β_1 e α, respectivamente. Assim, considerando um nível de significância de 10%, rejeita-se a hipótese de estacionariedade para os parâmetros regressores, reafirmando mais uma vez a preferência pelo modelo espacial local.

Conforme indicado na Tabela 4.8, os ajustes da RBNGP e da RBNGPg são parecidos, visto que estimaram praticamente o mesmo parâmetro de suavização e resultaram em log-verossimilhanças similares. No entanto, conforme explorado nos outros estudos de caso, a RBNGPg tem a desvantagem de apresentar uma estimativa viesada para o parâmetro α. Sendo assim, a frota de caminhões simples será modelada pela RBNGP. Os resultados desse ajuste estão apresentados na Figura 4.22.

A partir da Figura 4.22, nota-se que os valores estimados para o intercepto do modelo são mais elevados na região metropolitana do estado do Espírito Santo, onde se localiza a capital - Vitória, refletindo a maior concentração de caminhões simples nestes lugares. Já as estimativas b_1, referentes a variável explicativa *Indústrias*, são menores nessa região devido ao grande número de indústrias x_1 associado à forma exponencial do modelo. Os mapas da estatística *pseudo* t para estes parâmetros sugerem que eles são significativos.

É importante lembrar que a variável indústria foi utilizada como indicadora do padrão econômico do município, a fim de reduzir o erro na estimação da frota esperada de caminhões. Não há interesse no estabelecimento de uma relação de causa e efeito entre essas variáveis, nem diretamente nos valores estimados para este parâmetro.

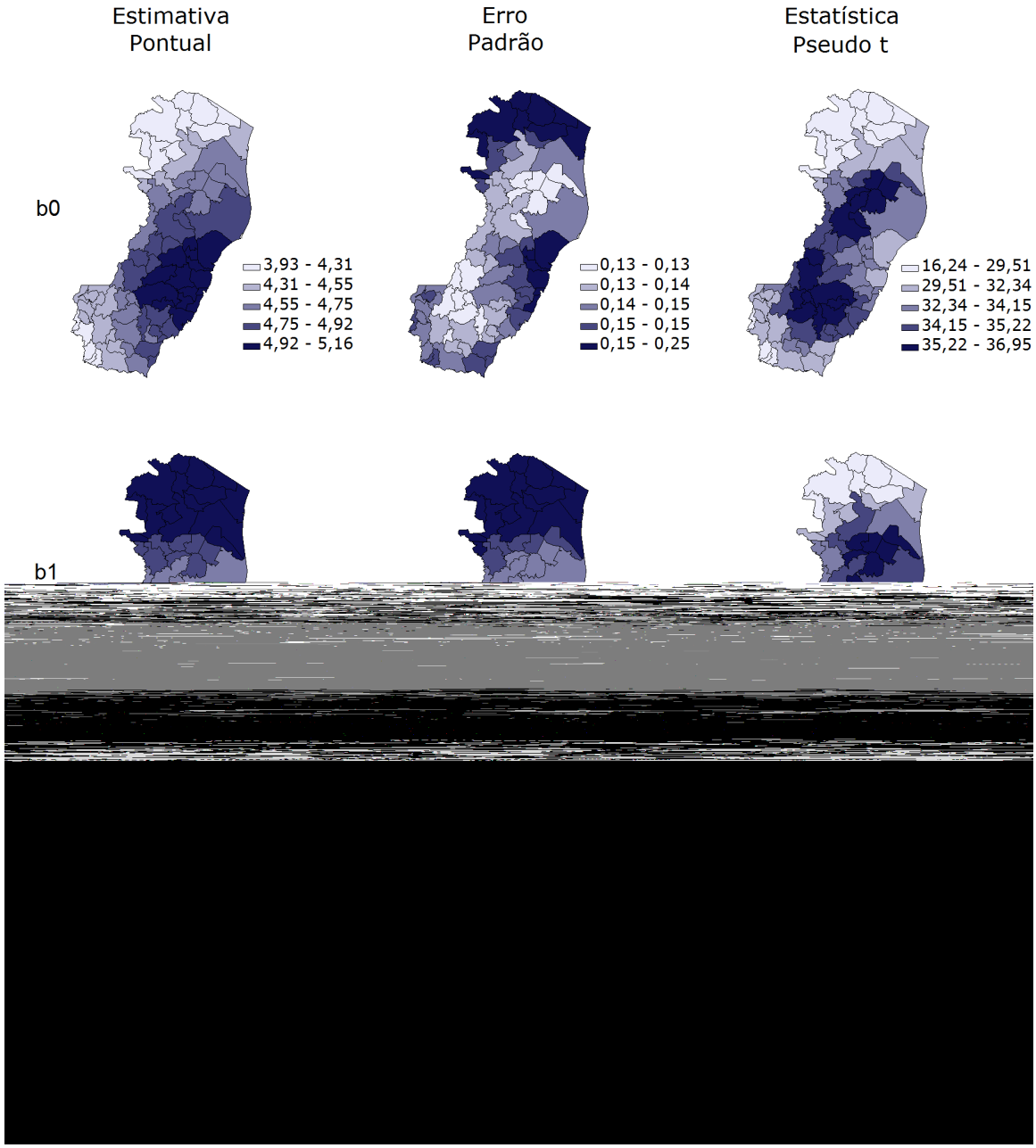

Figura 4.22: Estimativas pontuais, erros padrão e estatísticas *pseudo* t da RBNGP

No entanto, a título de ilustração, considere o município de Guarapari, localizado no litoral do estado, próximo à capital Vitória, e cuja frota média ajustada é dada por $\mu = \exp(5.04 + 0.0026x)$. Caso este município tivesse 100 indústrias a mais, estima-se que o seu número esperado de caminhões seria $\exp(100 \times 0.0026) = 1.297$ vezes maior. Como Guarapari tem 203 indústrias, sua frota esperada aumentaria de, aproximadamente, 261 para $261 \times 1.297 = 338$ caminhões.

Considerando o modelo de regressão Binomial Negativa global, a frota média

ajustada para todos os municípios do estado seria $\mu = \exp(4.66 + 0.0038x)$. Portanto, para Guarapari, o número esperado de caminhões seria de 227. Considerando que o valor observado foi 316, a estimativa do modelo RBNGP (261 caminhões) foi mais próxima. Apesar deste ser apenas um exemplo ilustrativo aplicado para um município específico, pelos resultados anteriormente apresentados, tem-se que a RBNGP fornece, de fato, um melhor ajuste.

Em relação as estimativas do parâmetro de superdispersão, nota-se, pela Figura 4.22, que elas são maiores no sul do estado e decrescem gradativamente na direção norte. No entanto, considerando o mapa das estatísticas *pseudo* t, tem-se que o valor de α na região norte do estado do Espírito Santo não é significativo.

Retomando a aplicação deste modelo para indicação de possíveis municípios candidatos para implantação de concessionárias, tem-se que o objetivo principal da modelagem está na diferença entre os valores observados e ajustados, ou seja, nos resíduos. A Figura 4.23 apresenta o mapa dos resíduos da RBNGP, enfatizando os que assumem valores negativos, que são os de maior interesse.

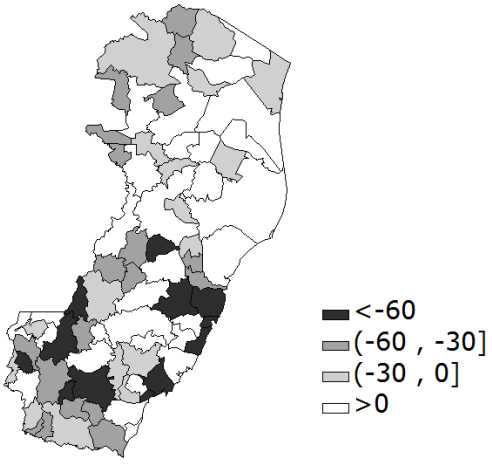

Figura 4.23: Resíduo da RBNGP

Os municípios destacados em cinza escuro no mapa da Figura 4.23 apresentam um défice de caminhões superior a 60. Observe que não é aconcelhável criar uma concessionária na região norte do estado, cuja demanda por caminhões está, pelo menos em tese, suprida ou é relativamente baixa. Já os bolsões em cinza escuro são regiões com maior potencial de vendas.

Dessa maneira, valendo-se deste gráfico, que materializa o cerne da avaliação de-

senvolvida, o investidor coloca-se em situação privilegiada em relação a seus competidores. Em resumo, a utilização de técnicas eficientes de análise de dados orientam a boa tomada de decisão. Cabe destacar que não se trata de um estudo conclusivo. A partir dele, investigações detalhadas e direcionadas são conduzidas a fim de concluir a estratégia de investimento.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Nesta dissertação foram elaborados dois modelos de regressão geograficamente ponderada utilizando a distribuição Binomial Negativa, a saber, RBNGPg e RBNGP. A diferença entre eles está na forma de estimação do parâmetro de superdispersão. Para a RBNGPg, ele é estimado globalmente, supondo que todos os parâmetros são constantes, consequentemente, fornece uma estimativa viesada. No entanto, sua simplicidade possibilita calcular o número efetivo de parâmetros do modelo. Já para a RBNGP este número é ainda desconhecido. Em contrapartida, o modelo permite que o parâmetro α também varie espacialmente.

Com base na simulação de dados espaciais não estacionários, verificou-se que a regressão global retrata apenas o comportamento médio das estimativas dos parâmetros, sendo inadequada para descrever as peculiaridades locais. Além disso, a modelagem utilizando a distribuição de Poisson, tanto pela regressão global, quanto pela regressão geograficamente ponderada, mostrou-se inadequada para dados com superdispersão.

A RBNGP apresentou o melhor ajuste para dados de contagem não estacionários e com superdispersão, como os dados reais da frota de caminhões simples do Estado do Espírito Santo. Além disso, no estudo de caso simulado, as superfícies reais dos parâmetros foram muito bem estimadas por este modelo. Apesar de superestimar o parâmetro α, a RBNGPg também ajustou de forma satisfatória os dados dessa natureza.

Neste trabalho verificou-se que uma ótima aproximação para a RPGP pode ser realizada escolhendo $\alpha = 0$ na RBNGPg. E, principalmente, foi constatado que a RBNGP tem a habilidade de identificar a ausência de superdispersão nos dados, o que já não é possível pela RBNGPg, que confunde a superdispersão com a não estacionariedade na estimação do parâmetro α.

Além disso, observou-se que tanto a RBNGP quanto a RBNGPg são capazes de detectar a estacionariedade espacial. E, aumentando o parâmetro de suavização, os modelos aqui propostos convergem para as regressões globais Binomial Negativa e Poisson.

Sendo assim, conclui-se que a RBNGP é um modelo muito robusto para o ajuste de dados de contagem, sendo especialmente útil quando os dados apresentam não estacionariedade e superdispersão. E, com exceção do caso em que $\alpha = 0$, a RBNGPg também fornece ajustes razoáveis.

5.2 Trabalhos Futuros

Naturalmente, as propriedades e características da Regressão Binomial Negativa Geograficamente Ponderada não foram todas exploradas nesta dissertação. As principais atividades que podem ser desenvolvidas em trabalhos futuros são:

- Determinar o número de parâmetros efetivos do modelo RBNGP. Não foi possível, nesta dissertação, avaliar a contribuição da variação espacial de α para este número, consequentemente, o mesmo está indefinido. O número efetivo de parâmetros do modelo permite uma análise da complexidade do ajuste, o cálculo de estatísticas importantes como o AICc e o BIC, além de testes formais da qualidade do ajuste. Assim, essa determinação é de grande valia para a RBNGP.

- Comparar, por meio de testes de simulação, os dois métodos de cálculo do erro padrão da regressão geograficamente ponderada. Conforme sugerido por Fotheringham et al. (2002), foi utilizada nesta dissertação a estimativa global de variância da variável dependente ajustada no cálculo do erro padrão do modelo. No entanto, de acordo com os resultados obtidos neste trabalho, esta estimativa é menos conservadora do que as provenientes das próprias regressões locais, que

seria a outra possibilidade. Além disso, sugere-se também que seja feita uma avaliação dos níveis dos intervalos de confiança, verificando, por exemplo, se a confiança de 95% é realmente observada na prática. É importante ressaltar que os intervalos são pontuais, de fato, não se trata de um envelope de 95% para a superfície dos parâmetros como um todo.

- Realizar um detalhamento formal das técnicas de diagnóstico do modelo, como avaliação da multicolinearidade (vide Wheeler e Tiefelsdorf (2005)), análise de resíduos, detecção de *outliers* e observações influentes. Estes tópicos não foram tratados nesta dissertação, mas são importantes na avaliação da adequabilidade do ajuste.

- Estender o modelo RBNGP para o caso semi-paramétrico, no qual a variação espacial é permitida apenas para os parâmetros não estacionários. Ganhos na qualidade do ajuste da RPGP semi-paramétrica são apresentados em Nakaya et al. (2005).

- Tornar o código desenvolvido em linguagem SAS/IML mais robusto, menos sucetível a problemas de convergência e mais veloz para trabalhar com grandes bancos de dados. Para o maior tamanho de amostra utilizado nesta dissertação ($n = 504$), a macro *%golden* levou 30 minutos para encontrar o parâmetro de suavização ótimo do modelo RBNGP. Já a macro *%estac* demorou, aproximadamente, 24 horas para realizar o teste de estacionariedade com $m = 999$ repetições. E, utilizando o grid de 17.636 pontos e o tamanho de amostra $n = 504$, a macro *%gwnbr* precisou de 45 minutos para realizar o ajuste da RBNGP.

Referências Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Santa Barbara, EUA.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (1996). The Moran Scatterplot as ESDA Tool to Assess Local Instability in Spatial Association. *Spatial Analytical Perspectives on GIS, Londres, UK*.

Assunção, R. M. (2003). Índices de auto-correlação espacial. Departamento de estatística - UFMG. Notas de aula.

Casella, G. & Berger, R. L. (2001). *Statistical Inference*, (2nd ed.). Duxbury.

Chen, V. Y.-J. & Yang, T.-C. (2011). Sas macro programs for geographically weighted generalized linear modeling with spatial point data: Applications to health research. *Computer Methods and Programs in Biomedicine*.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cliff, A. D. & Ord, J. K. (1981). *Spatial Processes: Models and Applications*. London: Pion Ltd.

Cordeiro, G. M. & Demétrio, C. G. B. (2010). *Modelos Lineares Generalizados e Extensões*. Não publicado.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, (2nd ed.). Chapman and Hall,CRC.

Druck, S., Carvalho, M. S., Câmara, G., & Monteiro, A. M. V. (2004). *Análise Espacial de Dados Geográficos*. EMBRAPA.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression*. Wiley.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5:115–145.

Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted poisson regression. *Accident Analysis and Prevention*, 42:678–688.

Hardin, J. W. & Hilbe, J. M. (2001). *Generalized Linear Models and Extensions*. Stata Press.

Hilbe, J. M. (2011). *Negative Binomial Regression*, (2nd ed.). Cambridge University Press.

Hope, A. C. A. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society*, 30:582–598.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Kobayashi, T. & Lane, B. (2007). Spatial heterogeneity and transit use. *11th World Conference on Transportation Research*.

LeSage, J. P. (2001). A Family of Geographically Weighted Regression Models. *Department of Economics University of Toledo*, pages 1–36.

Leung, Y., Mei, C.-L., & Zhang, W.-X. (2000a). Statistical tests for spacial nonstationarity based on the geographically weighted regression model. *Environment and Planning A*, 32:9/32.

Leung, Y., Mei, C.-L., & Zhang, W.-X. (2000b). Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning A*, 32:871–890.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in Medice*, 24:2695 – 2717.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135:370–384.

Silva, A. R. (2006). Avaliação de modelos de regressão espacial para análise de cenários do transporte rodoviário de carga. Master's thesis, ENC-FT-UnB.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B)*, 13:238–241.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84:276–283.

Tobler, W. R. (1979). Cellular Geography. *Philosophy in Geography*. Edited by S. Gale and G. Olssen, (Dordrecht: Reidel).

Wheeler, D. & Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7:161–187.

Zörnig, P. (2011). *Introdução à Programação Não-Linear*. Editora UnB.

Apêndice

A.1 Macro %golden

Código SAS

```
/*************************************************************************/
/* Macro do algoritmo da divisão áurea para a busca do parâmetro de suavização
/*
/* Parâmetros
/*   data  = nome do banco de dados SAS que contém as variáveis de interesse
/*   y     = nome da variável dependente
/*   x     = nome das variáveis independentes separadas por espaço
/*   lat   = nome da variável referente a latitude ou coordenada y do sistema UTM
/*   long  = nome da variável referente a longitude ou coordenada x do sistema UTM
/*   kernel= escolha da função de ponderação, existem 3 opções:
/*            GAUSS kernel gaussiano fixo;
/*            BIQUAD kernel biquadrático fixo;
/*            ADAPTIVE kernel biquadrático adaptativo;
/*    stat = Estatística que define o critério de busca, existem 3 opções:
/*            AICC: Critério de Informação de Akaike Corrigido;
/*            CV:  Validação cruzada (Cross Validation);
/*            DEV: Desvio (Deviance);
/*    gwnbr= escolha da RGP, existem 3 opções:
/*            GLOBAL: Realiza a RBNGPg;
/*            LOCAL: Realiza a RBNGP;
/*            POISSON: Realiza a RPGP;
/*    offset = nome da variável offset (opcional)
/*************************************************************************/
%macro golden(data=,y=,x=,lat=,long=,kernel=,stat=, gwnbr=, offset=);
proc iml;
use &data;
  read all var {&y} into y;
  read all var {&x} into x;
  read all var{&long &lat} into COORD;
  n=nrow(y);
  %if &offset= %then %do; offset=j(n,1,0); %end;
  %else %do; read all var {&offset} into offset; %end;
close &data;
x=j(n,1,1)||x;
kernel= "%lowcase(&kernel)";
stat="%lowcase(&stat)";
gwnbr="%lowcase(&gwnbr)";
print kernel stat gwnbr;
start dist(coord,n);
  d=j(1,3,0);
```

```
41    nome={"idi" "idj" "d"};
      create _dist_ from d[colname=nome];
43    do i=1 to n;
            do j=i+1 to n;
45            if abs(coord[,1])<180 then do;
                  dif=abs(COORD[i,1]-COORD[j,1]);
47                raio=arcos(-1)/180;
            ang=sin(COORD[i,2]*raio)*sin(COORD[j,2]*raio)+cos(COORD[i,2]*raio)*cos(COORD[
                j,2]*raio)*cos(dif*raio);
49          arco=arcos(ang);
                d[1]=i;
51              d[2]=j;
                d[3]=arco*6371;
53              append from d;
              end;
55            else do;
                  d[1]=i;
57                d[2]=j;
                  d[3]=sqrt((COORD[i,1]-COORD[j,1])**2+(COORD[i,2]-COORD[j,2])**2);
59                append from d;
              end;
61          end;
      end;
63    close _dist_;
    finish dist;
65  run dist(coord,n);
    use _dist_;
67    read all into d;
      maxd=int(max(d[,3])+1);
69    free d;
    close _dist_;
71  if  kernel= "adaptive" then do;
      h0= 10 ; h3= n;
73  end;
    else if kernel= "biquad" | kernel= "gauss" then do;
75    h0= 0 ; h3= maxd;
    end;
77  r=0.61803399; c=1-r;
    if kernel= "adaptive" then tol=0.9; else tol=0.1;
79  h1=h0+(1-r)*(h3-h0);
    h2=h0+r*(h3-h0);
81  print h0 h1 h2 h3;
    start cv(h) global(kernel, n, coord, x, y, stat, maxd, gwnbr, offset);
83    alphaii= j(n,2,0);
      yhat=j(n,1,0);
85    S=j(n,n,0);
      if gwnbr="global" then do;
87      ym=sum(y)/nrow(y);
        u=(y+ym)/2;
89      n=log(u);
        par=1; ddpar=1; j=0; aux2=0;
91      do while (abs(ddpar)>0.00001);
          aux1=0; dpar=1; parold=par;
93        do while (abs(dpar)>0.001);
            aux1=aux1+1;
95          if par<0 then do;
              par=0.00001;
97          end;
```

```
        par=choose(par<1E-10,1E-10,par);
99          g=sum(digamma(par+y)-digamma(par)+log(par)+1-log(par+u)-(par+y)/(par+u));
            hess=sum(trigamma(par+y)-trigamma(par)+1/par-2/(par+u)+(y+par)/((par+u)#(
                par+u)));
101     hess=choose(abs(hess)<1E-23,sign(hess)*1E-23,hess);
        hess=choose(hess=0,1E-23,hess);
103         par0=par;
            par=par0-inv(hess)*g;
105     if aux1>30 & par>1E5 then do;
          dpar= 0.0001;
107       aux2=aux2+1;
          if aux2=1 then par=2 ;
109       else if aux2=2 then par=1E5;
          else if aux2=3 then par=0.0001;
111     end;
        else dpar=par-par0;
113   end;
      a=1/par; dev=0; ddev=1; i=0;
115   do while (abs(ddev)>0.00001);
        i=i+1;
117         w=(u/(1+a*u))+(y-u)#(a*u/(1+2*a*u+a*a*u#u));
            z=n+(y-u)/(w#(1+a*u)) - offset;
119         b=inv((x#w)'*x)*(x#w)'*z;
            n=x*b + offset;
121         u=exp(n);
            olddev=dev;
123     tt=y/u;
        tt=choose(tt=0,1E-10,tt);
125         dev=2*sum(y#log(tt)-(y+1/a)#log((1+a*y)/(1+a*u)));
            ddev=dev-olddev;
127   end;
      if aux2>4 then ddpar=1E-9;
129   else ddpar=par-parold;
    end;
131 alpha=a;
  end;
133 n=nrow(y);
  aux2=0;
135 do i=1 to n;
    d=j(1,3,0);
137 dist=d;
        do j=1 to n;
139       if abs(coord[,1])<180 then do;
              dif=abs(COORD[i,1]-COORD[j,1]);
141           raio=arcos(-1)/180;
        ang=sin(COORD[i,2]*raio)*sin(COORD[j,2]*raio)+cos(COORD[i,2]*raio)*cos(COORD[
            j,2]*raio)*cos(dif*raio);
143     if i=j then arco=0;
              else arco=arcos(ang);
145           d1=arco*6371;
          end;
147       else d1=sqrt((COORD[i,1]-COORD[j,1])**2+(COORD[i,2]-COORD[j,2])**2);
      d[1]=i; d[2]=j; d[3]=d1;
149   if j=1 then dist=d;
      else dist=dist//d;
151 end;
    u=nrow(dist);
153 w=j(u,1,0);
```

```
      if kernel= "gauss" then do;
155     if stat="cv" then do;
              do jj=1 to u;
157           if dist[jj,3]<=maxd*0.8 & dist[jj,3]^=0 then w[jj]=exp(-0.5*(dist[jj,3]/h)*
                *2);
              else w[jj]= 0;
159             end;
        end;
161     else do;
              do jj=1 to u;
163           if dist[jj,3]<=maxd*0.8 then w[jj]=exp(-0.5*(dist[jj,3]/h)**2);
              else w[jj]= 0;
165             end;
        end;
167   end;
      else if kernel= "biquad" then do;
169     if stat="cv" then do;
              do jj=1 to u;
171           if dist[jj,3]<=h & dist[jj,3]^=0 then w[jj]=(1-(dist[jj,3]/h)**2)**2;
              else w[jj]= 0;
173             end;
        end;
175     else do;
              do jj=1 to u;
177           if dist[jj,3]<=h then w[jj]=(1-(dist[jj,3]/h)**2)**2;
              else w[jj]= 0;
179             end;
        end;
181   end;
      else if kernel= "adaptive" then do;
183     call sort(dist,{3});
        dist=dist||(1:n)';
185     w=j(n,2,0);
        hn=dist[h,3];
187     if stat="cv" then do;
          do jj=1 to n;
189         if dist[jj,4]<= h & dist[jj,3]^=0 then w[jj,1]=(1-(dist[jj,3]/hn)**2)**2;
            else w[jj,1]=0;
191         w[jj,2]=dist[jj,2];
          end;
193     end;
        else do;
195       do jj=1 to n;
            if dist[jj,4]<=h then w[jj,1]=(1-(dist[jj,3]/hn)**2)**2;
197         else w[jj,1]=0;
            w[jj,2]=dist[jj,2];
199       end;
        end;
201     call sort(w,{2});
      end;
203   wi=diag(w[,1]);
      ym=sum(y)/nrow(y);
205   uj=(y+ym)/2;
      nj=log(uj);
207   if i=1 | aux2=5 then par=1; else par=alphaii[i-1,2];
      ddpar=1; j=0; count=0; aux2=0;
209   do while (abs(ddpar)>0.000001);
        aux1=0;
```

```
211         dpar=1;
            parold=par;
213         if gwnbr="global" | gwnbr="poisson" then do;
              dpar=0.00001;
215           if gwnbr="global" then par=1/a;
            end;
217         /* calculating alpha=1/par, where par=theta */
            do while (abs(dpar)>0.001);
219           aux1=aux1+1;
              if gwnbr="local" then do;
221             par=choose(par<1E-10,1E-10,par);
                    g=sum((digamma(par+y)-digamma(par)+log(par)+1-log(par+uj)-(par+y)/(par+
                        uj))#w[,1]);
223                 hess=sum((trigamma(par+y)-trigamma(par)+1/par-2/(par+uj)+(y+par)/((par+
                        uj)#(par+uj)))#w[,1]);
              end;
225           hess=choose(abs(hess)<1E-23,sign(hess)*1E-23,hess);
              hess=choose(hess=0,1E-23,hess);
227               par0=par;
                  par=par0-inv(hess)*g;
229           if par<=0 then do;
                count=count+1;
231             if count<10 then par=0.000001;
                else par=abs(par);
233           end;
              if aux1>30 & par>1E5 | aux1>200 then do;
235             dpar= 0.0001;
                aux2=aux2+1;
237             if aux2=1 then par=2 ;
                else if aux2=2 then par=1E5;
239             else if aux2=3 then par=0.0001;
              end;
241           else do;
                dpar=par-par0;
243             if par<1E-3 then dpar=dpar*100;
              end;
245         end;
            if gwnbr="poisson" then alpha=0;
247         else alpha=1/par;
            dev=0; ddev=1; cont=0;
249         /* calculating the parameters estimates */
            do while (abs(ddev)>0.000001);
251           cont=cont+1;
              uj=choose(uj>1E100,1E100,uj);
253           aux= (alpha*uj/(1+2*alpha*uj+alpha*alpha*uj#uj));
                  Ai=(uj/(1+alpha*uj))+(y-uj)#aux;
255           Ai=choose(Ai<1E-5,1E-5,Ai);
                  zj=nj+(y-uj)/(Ai#(1+alpha*uj)) - offset;
257           Ai=diag(Ai);
              if det(x'*wi*Ai*x)=0 then bi=j(ncol(x),1,0);
259           else bi=inv(x'*wi*Ai*x)*x'*wi*Ai*zj;
                  nj=x*bi + offset;
261           nj=choose(nj>1E2,1E2,nj);
                  uj=exp(nj);
263               olddev=dev;
              uj=choose(uj<1E-150,1E-150,uj);
265           tt=y/uj;
              tt=choose(tt=0,1E-10,tt);
```

```
267           if gwnbr="poisson" then dev=2*sum(y#log(tt)-(y-uj));
              else dev=2*sum(y#log(tt)-(y+1/alpha)#log((1+alpha*y)/(1+alpha*uj)));
269           if cont>50 then ddev= 0.0000001;
                  else ddev=dev-olddev;
271         end;
            j=j+1;
273         if gwnbr="global" | gwnbr="poisson" | aux2>4 | ddpar=0.0000001 then ddpar=1E-9;
            else do;
275           ddpar=par-parold;
              if par<1E-3 then ddpar=ddpar*100;
277         end;
        end;
279     Ai2=(uj/(1+alpha*uj))+(y-uj)#(alpha*uj/(1+2*alpha*uj+alpha*alpha*uj#uj));
        if Ai2[><,]<1E-5 then Ai2=choose(Ai2<1E-5,1E-5,Ai2);
281     Ai=diag(Ai2);
        if det(x'*wi*Ai*x)=0 then S[i,]=j(1,n,0);
283     else S[i,]= x[i,]*inv(x'*wi*Ai*x)*x'*wi*Ai;
        yhat[i]=uj[i];
285     alphaii[i,1]=i;
        alphaii[i,2]= alpha;
287   end;
      alpha= alphaii[,2];
289   yhat=choose(yhat<1E-150,1E-150,yhat);
      tt=y/yhat;
291   tt=choose(tt=0,1E-10,tt);
      if gwnbr="poisson" then dev=2*sum(y#log(tt)-(y-yhat));
293   else dev=2*sum(y#log(tt)-(y+1/alpha)#log((1+alpha#y)/(1+alpha#yhat)));
      a2=y+1/alpha; b2=1/alpha; c2=y+1;
295   algamma=j(n,1,0); blgamma=j(n,1,0); clgamma=j(n,1,0);
      do i=1 to nrow(y);
297     algamma[i]=lgamma(a2[i]); blgamma[i]=lgamma(b2[i]); clgamma[i]=lgamma(c2[i]);
      end;
299   if gwnbr^="poisson" then do;
        ll=sum(y#log(alpha#yhat)-(y+1/alpha)#log(1+alpha#yhat)+ algamma - blgamma -
            clgamma );
301     npar=trace(S)+1;
      end;
303   else do;
        ll=sum(-yhat+y#log(yhat)-clgamma);
305     npar=trace(S);
      end;
307   /*AIC= 2*npar + dev;*/
      AIC= 2*npar -2*ll;
309   AICC= AIC +(2*npar*(npar+1))/(n-npar-1);
      CV=(y-yhat)'*(y-yhat);
311   res=cv||aicc||npar||dev;
      return (res);
313 finish;
    if stat="cv" then do;
315   pos=1;
      create golden var{h1 res1 h2 res2};
317 end;
    else do;
319   if stat="aicc" then pos=2;
      else pos=4;
321   create golden var{h1 res1 npar1 h2 res2 npar2};
    end;
323 res1=cv(h1); npar1=res1[3]; res1=res1[pos];
```

```
    res2=cv(h2); npar2=res2[3]; res2=res2[pos];
325 append;
    do while(abs(h3-h0) > tol*2);
327     if res2<res1 then do;
            h0=h1;
329     h1=h2;
            h2=c*h1+r*h3;
331         res1=res2;
        npar1=npar2;
333     res2=cv(h2);
        npar2=res2[3];
335     res2=res2[pos];
        end;
337     else do;
            h3=h2;
339         h2=h1;
            h1=c*h2+r*h0;
341         res2=res1;
        npar2=npar1;
343     res1=cv(h1);
        npar1=res1[3];
345     res1=res1[pos];
        end;
347   append;
    end;
349 if kernel= "adaptive" then do;
      xmin = (h3+h0)/2;
351   h2=ceil(xmin); h1=floor(xmin);
      golden1 = cv(h1); g1= golden1[pos];
353   golden2= cv(h2); g2= golden2[pos];
      npar1=golden1[3]; res1=golden1[pos];
355   npar2=golden2[3]; res2=golden2[pos];
      append;
357   if g1<g2 then do;
        xmin=h1;
359     npar=golden1[3];
        golden=g1;
361   end;
      else do;
363     xmin=h2;
        npar=golden2[3];
365     golden=g2;
      end;
367 end;
    else do;
369   xmin = (h3+h0)/2;
      golden = cv(xmin);
371   npar=golden[3];
      golden=golden[pos];
373 end;
    h1 = xmin; res1 = golden; npar1=npar;
375 h2 = .; res2 = .; npar2=.;
    append;
377 if stat="cv" then print golden xmin;
    else print golden xmin npar;
379 quit;
    %mend golden;
```

A.2 Macro %gwnbr

Código SAS

```
/*************************************************************************/
/* Macro que realiza a Regressão Geograficamente Ponderada
/*
/* Parâmetros
/*   data  = nome do banco de dados SAS que contém as variáveis de interesse
/*   y     = nome da variável dependente
/*   x     = nome das variáveis independentes separadas por espaço
/*   lat   = nome da variável referente a latitude ou coordenada y do sistema UTM
/*   long  = nome da variável referente a longitude ou coordenada x do sistema UTM
/*   h     = valor do parâmetro de suavização (em distância ou número de pontos)
/*   grid  = nome do banco de dados SAS que contém as coordenadas do grid (opcional)
/*   latg  = nome da variável latitude ou coordenada y UTM do grid (opcional)
/*   longg = nome da variável longitude ou coordenada x UTM do grid (opcional)
/*   gwnbr = escolha da RGP, existem 3 opções:
/*           GLOBAL: Realiza a RBNGPg;
/*           LOCAL: Realiza a RBNGP;
/*           POISSON: Realiza a RPGP;
/*   kernel= escolha da função de ponderação, existem 3 opções:
/*           GAUSS kernel gaussiano fixo;
/*           BIQUAD kernel biquadrático fixo;
/*           ADAPTIVE kernel biquadrático adaptativo;
/*   alphag = valor fixo para o parâmetro alpha da RBNGPg (opcional)
/*   offset = nome da variável offset (opcional)
/*   id     = nome da variável de identificação das áreas
/*************************************************************************/
%macro gwnbr(data=,y=,x=,lat=,long=,h=,grid=,latg=,longg=,gwnbr=,kernel=, alphag=,
    offset=, id=);
proc iml;
use &data;
  read all var {&y} into y;
  read all var {&x} into x;
  read all var{&long &lat} into COORD;
  n=nrow(y);
  %if &offset= %then %do; offset=j(n,1,0); %end;
  %else %do; read all var {&offset} into offset; %end;
  %if &grid= %then %do;
    read all var{&long &lat} into POINTS;
    read all var{&id} into id_;
  %end;
close &data;
%if &grid^= %then %do;
  use &grid;
  read all var{&longg &latg} into POINTS;
  read all var{&id} into id_;
  close &grid;
%end;
x=j(n,1,1)||x;
yhat=j(n,1,0);
h=&h;
kernel= "%lowcase(&kernel)";
gwnbr="%lowcase(&gwnbr)";
m=nrow(POINTS);
bii=j(ncol(x)*m,2,0); alphaii= j(m,2,0);
```

```
xcoord=j(ncol(x)*m,1,0); ycoord=j(ncol(x)*m,1,0);
&id= j(ncol(x)*m,1,0);
sebi=j(ncol(x)*m,1,0); varmu=j(n,1,0); sealphai= j(m,1,0);
S=j(n,n,0);
yp=y-sum(y)/n;
probai=j(m,1,0); probbi=j(m,1,0);
yhat=j(m,1,0);
res= j(m,1,0);
if gwnbr^="poisson" then do;
  ym=sum(y)/nrow(y);
  u=(y+ym)/2;
  n=log(u);
  par=1; ddpar=1; j=0; aux2=0;
  do while (abs(ddpar)>0.00001);
    aux1=0;
    dpar=1;
    parold=par;
    do while (abs(dpar)>0.001);
      aux1=aux1+1;
      if par<0 then par=0.00001;
      par=choose(par<1E-10,1E-10,par);
          g=sum(digamma(par+y)-digamma(par)+log(par)+1-log(par+u)-(par+y)/(par+u));
          hess=sum(trigamma(par+y)-trigamma(par)+1/par-2/(par+u)+(y+par)/((par+u)#(
              par+u)));
      hess=choose(abs(hess)<1E-23,sign(hess)*1E-23,hess);
      hess=choose(hess=0,1E-23,hess);
          par0=par;
          par=par0-inv(hess)*g;
      if aux1>30 & par>1E5 then do;
        dpar= 0.0001;
        aux2=aux2+1;
        if aux2=1 then par=2 ;
        else if aux2=2 then par=1E5;
        else if aux2=3 then par=0.0001;
      end;
      else dpar=par-par0;
    end;
    a=1/par; dev=0; ddev=1; i=0;
    do while (abs(ddev)>0.00001);
      i=i+1;
          w=(u/(1+a*u))+(y-u)#(a*u/(1+2*a*u+a*a*u#u));
          z=n+(y-u)/(w#(1+a*u)) - offset;
          b=inv((x#w)'*x)*(x#w)'*z;
          n=x*b + offset;
          u=exp(n);
          olddev=dev;
      tt=y/u;
      tt=choose(tt=0,1E-10,tt);
          dev=2*sum(y#log(tt)-(y+1/a)#log((1+a*y)/(1+a*u)));
          ddev=dev-olddev;
    end;
    if aux2>4 then ddpar=1E-9;
    else ddpar=par-parold;
  end;
  %if &alphag= %then %let alphag=a;
  %else %if &alphag=0 %then %let alphag=1e-8;
  %else %let alphag=&alphag;
  alphag=&alphag;
```

```
  bg=b;
  parg=par;
end;
if gwnbr="global" then print alphag aux2;
n=nrow(y);
aux2=0;
do i=1 to m;
  d=j(1,3,0);
  do j=1 to n;
    if abs(COORD[,1])<180 then do;
          dif=abs(POINTS[i,1]-COORD[j,1]);
          raio=arcos(-1)/180;
      ang=sin(POINTS[i,2]*raio)*sin(COORD[j,2]*raio)+cos(POINTS[i,2]*raio)*cos(COORD[
          j,2]*raio)*cos(dif*raio);
      if round(ang,0.000000001)=1 then arco=0;
          else arco=arcos(ang);
          d1=arco*6371;
        end;
        else d1=sqrt((POINTS[i,1]-COORD[j,1])**2+(POINTS[i,2]-COORD[j,2])**2);
    d[1]=i; d[2]=j; d[3]=d1;
    if j=1 then dist=d;
    else dist=dist//d;
  end;
  w=j(n,1,0);
  if kernel= "gauss" then do;
        do jj=1 to n;
      w[jj]=exp(-0.5*(dist[jj,3]/h)**2);
        end;
  end;
  else if kernel= "biquad" then do;
        do jj=1 to n;
      if dist[jj,3]<=h then w[jj]=(1-(dist[jj,3]/h)**2)**2;
      else w[jj]= 0;
        end;
  end;
  else if kernel= "adaptive" then do;
    w=j(n,2,0);
    call sort(dist,{3});
    dist=dist||(1:n)';
    hn=dist[h,3];
    do jj=1 to n;
      if dist[jj,4]<=h then w[jj,1]=(1-(dist[jj,3]/hn)**2)**2;
      else w[jj,1]=0;
      w[jj,2]=dist[jj,2];
    end;
    call sort(w,{2});
  end;
  wi=diag(w[,1]);
  ym=sum(y)/nrow(y);
  uj=(y+ym)/2;
  nj=log(uj);
  ddpar=1; j=0; count=0; aux2=0;
  if i=1 | aux2=5 | count=4 then par=1; else par=alphaii[i-1,2];
  do while (abs(ddpar)>0.000001);
    dpar=1;
    if ddpar=1 then parold=1.8139;
    else parold=par;
    aux1=0;
```

```
if gwnbr="global" | gwnbr="poisson" then do;
  dpar=0.00001;
  if gwnbr="global" then par=1/alphag;
end;
/* calculating alpha=1/par, where par=theta=r */
do while (abs(dpar)>0.001);
  aux1=aux1+1;
  if gwnbr="local" then do;
    par=choose(par<1E-10,1E-10,par);
        g=sum((digamma(par+y)-digamma(par)+log(par)+1-log(par+uj)-(par+y)/(par+uj
            ))#w[,1]);
        hess=sum((trigamma(par+y)-trigamma(par)+1/par-2/(par+uj)+(y+par)/((par+uj
            )#(par+uj)))#w[,1]);
  end;
      par0=par;
  hess=choose(abs(hess)<1E-23,sign(hess)*1E-23,hess);
  hess=choose(hess=0,1E-23,hess);
      par=par0-inv(hess)*g;
  if par<=0 then do;
    count=count+1;
    if count=1 then par=0.000001;
    else if count=2 then par=0.0001;
    else par=1/alphag;
  end;
  if aux1>30 & par>1E5 | aux1>200 then do;
    dpar= 0.0001;
    if aux2=0 then par=1/alphag + 0.0011;
    if aux2=1 then par=2 ;
    else if aux2=2 then par=1E5;
    else if aux2=3 then par=0.0001;
    aux2=aux2+1;
  end;
  else do;
    dpar=par-par0;
    if par<1E-3 then dpar=dpar*100;
  end;
end;
if gwnbr= "poisson" then alpha=0;
else alpha=1/par;
dev=0; ddev=1; cont=0;
/* calculating the parameters estimates */
do while (abs(ddev)>0.000001);
  cont=cont+1;
      Ai=(uj/(1+alpha*uj))+(y-uj)#(alpha*uj/(1+2*alpha*uj+alpha*alpha*uj#uj));
  Ai=choose(Ai<1E-5,1E-5,Ai);
      zj=nj+(y-uj)/(Ai#(1+alpha*uj))-offset;
  Ai=diag(Ai);
  if det(x'*wi*Ai*x)=0 then bi=j(ncol(x),1,0);
  else bi=inv(x'*wi*Ai*x)*x'*wi*Ai*zj;
      nj=x*bi + offset;
  nj=choose(nj>1E2,1E2,nj);
      uj=exp(nj);
      olddev=dev;
  uj=choose(uj<1E-150,1E-150,uj);
  tt=y/uj;
  tt=choose(tt=0,1E-10,tt);
  if gwnbr= "poisson" then dev=2*sum(y#log(tt)-(y-uj));
  else dev=2*sum(y#log(tt)-(y+1/alpha)#log((1+alpha*y)/(1+alpha*uj)));
```

```
      if cont>50 then ddev= 0.0000001;
224           else ddev=dev-olddev;
    end;
226   j=j+1;
    if gwnbr="global" | gwnbr="poisson" | aux2>4 | count>3 then ddpar=1E-9;
228   else do;
      ddpar=par-parold;
230   if par<1E-3 then ddpar=ddpar*100;
    end;
232 end;
    if aux2>4 then probai[i]=1;
234 if count>3 then probai[i]=2;
    Ai2=(uj/(1+alpha*uj))+(y-uj)#(alpha*uj/(1+2*alpha*uj+alpha*alpha*uj#uj));
236 if Ai2[><]<1E-4 | Ai2[<>]>1E15 then do;
    Ai2=choose(Ai2<1E-4,1E-4,Ai2);
238   Ai2=choose(Ai2>1E15,1E15,Ai2);
    end;
240 Ai=diag(Ai2);
/*  Erro padrão proveniente das regressões locais */
242 /*  C=inv(x'*wi*Ai*x);*/
/*  varb= C;*/
244 /*  seb=sqrt(vecdiag(varb));*/
    C=inv(x'*wi*Ai*x)*x'*wi*Ai;
246 varb= C*inv(Ai)*C';
    seb=sqrt(vecdiag(varb));
248 %if &grid= | &grid=&data %then %do;
      if det(x'*wi*Ai*x)=0 then S[i,]=j(1,n,0);
250   else S[i,]= x[i,]*inv(x'*wi*Ai*x)*x'*wi*Ai;
      coveta=x*varb*x';
252   matuj= diag(uj);
      covmu= matuj*coveta*matuj;
254   covmu= vecdiag(covmu);
      varmu[i,1]= covmu[i];
256 %end;
    if gwnbr^="poisson" then do;
258   ser=sqrt(1/abs(hess));
      r=1/alpha;
260   sealpha=ser/(r**2);
      sealphai[i,1]=sealpha;
262   alphaii[i,1]=i;
      alphaii[i,2]= alpha;
264 end;
    m1=(i-1)*ncol(x)+1;
266 m2=m1+(ncol(x)-1);
    sebi[m1:m2,1]=seb;
268 bii[m1:m2,1]=i;
    bii[m1:m2,2]=bi;
270 xcoord[m1:m2,1]= POINTS[i,1];
    ycoord[m1:m2,1]= POINTS[i,2];
272 &id[m1:m2,1]= id_[i,1];
    %if &grid= | &grid=&data %then %do;
274   yhat[i]=uj[i];
    %end;
276 end;
  tstat= bii[,2]/sebi;
278 probtstat=2*(1-probnorm(abs(tstat)));
  if gwnbr^="poisson" then do;
280 atstat= alphaii[,2]/sealphai;
```

```
  aprobtstat=2*(1-probnorm(abs(atstat)));
end;
b=bii[,2];
alphai=alphaii[,2];
_id_= bii[,1];
_ida_=alphaii[,1];

%if &grid= | &grid=&data %then %do;
  yhat=choose(yhat<1E-150,1E-150,yhat);
  tt=y/yhat;
  tt=choose(tt=0,1E-10,tt);
  if gwnbr="poisson" then dev=2*sum(y#log(tt)-(y-yhat));
  else dev=2*sum(y#log(tt)-(y+1/alphai)#log((1+alphai#y)/(1+alphai#yhat)));
  if gwnbr^="poisson" then do;
    a2=y+1/alphai; b2=1/alphai;
    algamma=j(n,1,0); blgamma=j(n,1,0);
    do i=1 to nrow(y);
      algamma[i]=lgamma(a2[i]);
      blgamma[i]=lgamma(b2[i]);
    end;
  end;
  c2=y+1;
  clgamma=j(n,1,0);
  do i=1 to nrow(y);
    clgamma[i]=lgamma(c2[i]);
  end;
  if gwnbr^="poisson" then do;
    ll=sum(y#log(alphai#yhat)-(y+1/alphai)#log(1+alphai#yhat)+ algamma - blgamma -
        clgamma );
    if gwnbr="global" & alphai^=1/parg then npar=trace(S);
    else npar=trace(S)+1;
  end;
  else do;
    ll=sum(-yhat+y#log(yhat)-clgamma);
    npar=trace(S);
  end;
  resord=y-yhat;
  sii=vecdiag(S);
  res=resord/sqrt(varmu#(1-sii));
  res=_ida_||COORD[,1]||COORD[,2]||y||yhat||res||varmu;
  /* AIC com base em Nakaya et al. (2005). */
  /*AIC= 2*npar + dev; */
  AIC= 2*npar - 2*ll;
  AICC= AIC +(2*npar*(npar+1))/(n-npar-1);
  BIC= npar*log(n) - 2*ll ;
  print gwnbr kernel ll dev npar aic aicc bic;
  create _res_ from res[colname={"_id_" "xcoord" "ycoord" "yobs" "yhat" "res" "varmu
      "}];
  append from res;
%end;
%else %do; print gwnbr kernel; %end;

create _beta_ var{_id_ &id xcoord ycoord b sebi tstat probtstat};
append;
xcoord=COORD[,1];ycoord=COORD[,2];
&id=id_;
create _alpha_ var{_ida_ &id xcoord ycoord alphai sealphai atstat aprobtstat probai};
append;
```

```
%let nvar=0;
%do %while(%scan(%str(&x),&nvar+1)~=);
    %let nvar=%eval(&nvar+1);
%end;
use _beta_;
  read all into b;
close _beta_;
n=nrow(b);
npar=&nvar+1;
%do i=0 %to &nvar;
  b&i=j(1,8,0);
  nome={"_id_" "&id" "xcoord" "ycoord" "b" "sebi" "tstat" "probtstat"};
  create b&i from b&i[colname=nome];
  do i=1 to (n/npar);
    b&i[1,]=b[(i-1)*npar+&i+1,];
    append from b&i;
  end;
%end;
quit;
%mend gwnbr;
```

A.3 Macro %estac

Código SAS

```
/*****************************************************************************/
/* Macro para o teste de não estacionariedade de aleatorização
/*
/* Parâmetros
/*   data  = nome do banco de dados SAS que contém as variáveis de interesse
/*   y     = nome da variável dependente
/*   x     = nome das variáveis independentes separadas por espaço
/*   lat   = nome da variável referente a latitude ou coordenada y do sistema UTM
/*   long  = nome da variável referente a longitude ou coordenada x do sistema UTM
/*   h     = valor do parâmetro de suavização (em distância ou número de pontos)
/*   grid  = nome do banco de dados SAS que contém as coordenadas do grid (opcional)
/*   latg  = nome da variável latitude ou coordenada y UTM do grid (opcional)
/*   longg = nome da variável longitude ou coordenada x UTM do grid (opcional)
/*   gwnbr = escolha da RGP, existem 3 opções:
/*           GLOBAL: Realiza a RBNGPg;
/*           LOCAL: Realiza a RBNGP;
/*           POISSON: Realiza a RPGP;
/*   kernel= escolha da função de ponderação, existem 3 opções:
/*           GAUSS kernel gaussiano fixo;
/*           BIQUAD kernel biquadrático fixo;
/*           ADAPTIVE kernel biquadrático adaptativo;
/*   alphag = valor fixo para o parâmetro alpha da RBNGPg (opcional)
/*   offset = nome da variável offset (opcional)
/*   id     = nome da variável de identificação das áreas
/*   rep    = número de réplicas
/*****************************************************************************/
%macro estac(data=,y=,x=,lat=,long=,h=,grid=,latg=,longg=,gwnbr=,kernel=, alphag=,
    offset= , id=, rep=);
%let nvar=0;
%do %while(%scan(%str(&x),&nvar+1)~=);
    %let nvar=%eval(&nvar+1);
%end;
%gwnbr(data=&data,y=&y,x=&x,lat=&lat,long=&long,h= &h,
grid=&grid,latg=&latg,longg=&longg, gwnbr=&gwnbr, kernel=&kernel, alphag=&alphag,
    offset=&offset,id=&id);

%macro vk(par);
proc iml;
use _beta_;
  read all into b;
close _beta_;
use _alpha_;
  read all var {alphai} into alpha;
close _alpha_;
n=nrow(b);
npar=&par+1; n=n/npar;
vk=0;
%do i=0 %to &par;
  b&i=j(n,1,0);
  do i=1 to n;
    b&i[i,1]=b[(i-1)*npar+&i+1,5];
  end;
  vk&i=sum((b&i - b&i[:] )##2)/n ;
```

```
  vk=vk||vk&i;
%end;
vka= sum((alpha - alpha[:] )##2)/n ;
vk=vk||vka;
idx = setdif(1:(npar+2),1);
vk = vk[,idx];
create vk from vk;
append from vk;
quit;
%mend vk;

%macro perm(data=,id=,xcoord=,ycoord=);
proc iml;
  use &data;
  read all var{&id &xcoord &ycoord} into tab;
  close &data;
  n=nrow(tab);
    u = 1:n;
    call randgen(u, "Uniform");
    _u_=rank(u);
  create perm var{_u_};
  append;
quit;
data perm; merge perm &data(drop=&id &xcoord &ycoord);run;
proc sort data=perm; by _u_;run;
data perm; merge perm &data(keep=&id &xcoord &ycoord); run;
%mend perm;

%vk(&nvar);
data vk2; set vk; i=1; run;
%do it=2 %to (&rep+1);
  %perm(data=&data, id=&id, xcoord=&long, ycoord=&lat);
  %gwnbr(data=perm,y=&y,x=&x,lat=&lat,long=&long,h= &h,
  grid=&grid,latg=&latg,longg=&longg, gwnbr=&gwnbr, kernel=&kernel, alphag=&alphag,
      offset=&offset,id=&id);
  %vk(&nvar);
  data vk; set vk; i=&it; run;
  proc append base=vk2 data=vk force; run;
%end;
proc iml;
use vk2;
  read all into x;
close vk2;
nvar=ncol(x)-1; n=nrow(x);
count=j(1,nvar,0);
do v=1 to nvar;
  do i=1 to n;
    if x[i,v]>=x[1,v] then count[v]=count[v]+1;
  end;
end;
count=count/n*100;
print count;
varnames="b0":"b&nvar"||"alpha";
create testing_stationarity from count [colname=varnames];
append from count;
quit;
%mend estac;
```