Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Regressão Geograficamente Ponderada Utilizando a Distribuição Binomial Negativa

Thais Carvalho Valadares Rodrigues

Brasília

Novembro de 2011

Thais Carvalho Valadares Rodrigues

Regressão Geograficamente Ponderada Utilizando a Distribuição Binomial Negativa

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

Novembro de 2011

Ao meu esposo, idealizador deste projeto.

Agradecimentos

Agradeço a Deus por estar ao meu lado sempre.

Ao meu esposo, por me incentivar a cursar Estatística e por me dar todo apoio necessário, 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.

E aos novos amigos, que tornaram o curso mais alegre e prazeroso.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

A regressão global pressupõe que um modelo único é adequado para descrever todas as partes da região de estudo. No entanto, a força dos relacionamentos entre 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.

Com isso, surge a Regressão Geograficamente Ponderada (RGP), a fim de modelar dados espaciais não estacionários. Utilizando funções *kernel*, a RGP gera superfícies não paramétricas das estimativas dos parâmetros.

Considerando dados de contagem com superdispersão, o mais adequado é utilizar a distribuição Binomial Negativa. Por isso, o presente trabalho propõe o modelo de Regressão Binomial Negativa Geograficamente Ponderada (RBNGP). O modelo aqui proposto permite que os parâmetros β variem espacialmente, no entanto ainda mantém o parâmetro de superdispersão α constante.

Neste trabalho, a RBNGP é aplicada a um conjunto de dados reais e os resultados obtidos mostram sua superioridade com respeito aos modelos concorrentes, a saber, regressão global - Poisson e Binomial Negativa - e Regressão de Poisson Geograficamente Ponderada.

Capítulo 1

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 região na qual esta medida foi observada. 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 a superfície 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. Devido à complexidade do tema, não é objetivo deste trabalho estimar a superfície dos erros padrão, apenas das estimativas locais dos parâmetros.

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

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \ , \tag{1.1}$$

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

Uma restrição limitante do modelo básico RGP dado pela Equação (1.1) é que a distribuição dos erros ε_i 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 utilizados 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 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. A vantagem desta última é a possibilidade de modelar dados com superdispersão, visto que para a distribuição de Poisson a média e a variância são iguais.

Sendo assim, o objetivo geral do trabalho é a estimação dos parâmetros do modelo de Regressão Binomial Negativa Geograficamente Ponderada (RBNGP) utilizando os algoritmos de Newton Raphson (NR) e de Mínimos Quadrados Reponderados Iterativo (MQRI).

Os objetivos específicos são:

- Aplicar o modelo desenvolvido em um conjunto de dados reais;
- Comparar as estimativas do modelo RBNGP com as provenientes do modelo de regressão global Binomial Negativa;
- Comparar a RBNGP com a Regressão de Poisson Geograficamente Ponderada (RPGP).

Capítulo 2

Regressão Binomial Negativa

2.1 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 modelos lineares generalizados, enfatizando dois casos particulares: a regressão Normal e a regressão Binomial Negativa. 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.

2.2 Modelo Linear Generalizado

O Modelo Linear Generalizado (ou do inglês, *Generalized Linear Models - GLM*) é um conjunto de técnicas estatísticas unificadas por Nelder e Wedderburn (1972)

para distribuições pertencentes à família exponencial.

A família exponencial de Nelder e Wedderburn (1972) engloba distribuições de probabilidade que podem ser escritas de acordo com a equação

$$f(y;\theta,\phi) = \exp\left\{\phi^{-1}[\theta y - b(\theta)] + c(y,\phi)\right\}, \tag{2.1}$$

onde θ é o parâmetro canônico, ϕ é um parâmetro de perturbação e $b(.)$ e $c(.)$ são funções conhecidas.

Exemplos de distribuições que podem ser escritas conforme (2.1) são: Normal, Binomial, Binomial Negativa, Poisson e Gama. No entanto, para as distribuições biparamétricas, é necessário supor que um dos parâmetros é conhecido.

O MLG é constituído de três componentes:

i) Componente aleatório: Conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ provenientes da família exponencial (2.1).

ii) Componente sistemático: Conjunto de parâmetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ e variáveis explicativas $\mathbf{X} = (x_1, \ldots, x_n)^T$, constituindo o preditor linear $\boldsymbol{\eta}$, dado por

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r \ . \tag{2.2}$$

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

$$\eta_i = g(\mu_i) \ . \tag{2.3}$$

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).

O valor esperado e a variância da variável aleatória Y pertencente à família (2.1) podem ser calculados a partir da função geradora de cumulantes $b(\theta)$, conforme indicado na Equação (2.4).

$$E(Y) = \mu = b'(\theta) \quad e \quad V(y) = \phi b''(\theta) = \phi V(\mu) \tag{2.4}$$

A seguir serão apresentados dois casos particulares: A regressão clássica e a regressão Binomial Negativa.

2.2.1 Regressão clássica

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

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \tag{2.5}$$

A fim de classificá-la como membro da família exponencial de Nelder e Wedderburn, é necessário escrever (2.5) conforme sugerido na Equação (2.1). Assim, por

meio de operações algébricas simples, obtém-se que

$$f(y;\mu,\sigma^2) = \exp\left\{(\sigma^2)^{-1}\left[y\mu - \frac{\mu^2}{2}\right] - \frac{1}{2}\left[\log(2\pi\sigma^2) + (\sigma^2)^{-1}y^2\right]\right\}. \qquad (2.6)$$

Comparando (2.6) com (2.1), conclui-se que:

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

A função de ligação canônica é a identidade pois $\eta = g(\mu) = \theta = \mu$. Assim, temos o modelo clássico de regressão:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \ ,$$
$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \ . \qquad (2.7)$$

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

2.2.2 Regressão Binomial Negativa

A distribuição Binomial Negativa é uma alternativa robusta para modelar dados de contagem. Esta distribuição pode ser interpretada como o número de fracassos y antes da ocorrência do r-ésimo sucesso em uma sequência de ensaios de Bernoulli independentes e identicamente distribuídos com probabilidade de sucesso p. A função densidade de probabilidade (fdp) da distribuição Binomial Negativa é dada por

$$f(y;p,r) = \binom{y+r-1}{r-1} p^r (1-p)^y \ , \tag{2.8}$$

onde $y \geq 0$, $r > 0$ e $0 < p < 1$. Em termos do parâmetro de superdispersão α, tem-se que

$$f(y;p,\alpha) = \binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1} p^{\frac{1}{\alpha}} (1-p)^y \ , \tag{2.9}$$

visto que $\alpha = \frac{1}{r}$.

Considerando o parâmetro r conhecido, é possível reescrever (2.8) de acordo com a família (2.1), ou seja,

$$f(y;p,r) = \exp\left\{ [y \log(1-p) + r \log(p)] + \log\binom{y+r-1}{r-1} \right\} . \tag{2.10}$$

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

- $\phi = 1$,
- $\theta = \log(1-p)$,
- $p = 1 - \exp(\theta)$,

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

Considerando que $p = \frac{r}{\mu+r}$ e a função de ligação canônica é dada por

$$g(\mu) = \theta = \log(1-p) = \log\left(\frac{\mu}{\mu+r}\right) ,$$

chega-se à forma canônica do modelo binomial negativo:

$$\log\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+r}\right) = \mathbf{X}\boldsymbol{\beta} . \tag{2.11}$$

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 $g(\mu) = \theta = \log(\mu)$. Sendo assim, temos que

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} . \tag{2.12}$$

Uma hipótese restritiva da distribuição de Poisson é a equidispersão, ou seja, a igualdade entre a média e a variância da variável aleatória Y. Considerando que a função de variância da Binomial Negativa é dada por $V(\mu) = \mu + \alpha\mu^2$, onde $\alpha > 0$, então $V(\mu) \geq \mu$, possibilitando a modelagem de dados com superdispersão. Outra vantagem do modelo Binomial Negativo é que o mesmo engloba o modelo de Poisson,

visto que este último é o NB-2 com r tendendo a infinito (ou $\alpha \to 0$) e $\mu = \frac{rp}{1-p}$, onde p é a probabilidade de fracasso. Além disso, para $r = 1$, a distribuição Binomial Negativa é equivalente à Geométrica, modelando o número de fracassos y antes da ocorrência do primeiro sucesso, isto é,

$$f(y; p) = p(1-p)^y, \quad y \in \mathbb{Z}^+ \ . \tag{2.13}$$

2.3 Algoritmos de estimação

Dois métodos são utilizados para estimar os modelos de contagem: o método de Newton Raphson (NR) e o de Mínimos Quadrados Reponderados Iterativo (MQRI).

2.3.1 Newton Raphson

O método de Newton Raphson será utilizado para estimar o parâmetro r da Binomial Negativa. O algoritmo computacional está apresentado na Tabela 2.1.

Tabela 2.1: 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)

Note que valores iniciais para o vetor de parâmetros devem ser fornecidos. A variável *tol* é a tolerância desejada no critério de parada. Além disso, **U** é o vetor

gradiente, ou seja, é a derivada de primeira ordem da log-verossimilhança:

$$\mathbf{U} = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \, , \tag{2.14}$$

onde $L(\beta)$ é a função de log-verossimilhança. Por fim, $\mathbf{H}$ é a matriz das derivadas de segunda ordem da log-verossimilhança, denominada matriz Hessiana,

$$\mathbf{H} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \, . \tag{2.15}$$

2.3.2 Mínimos Quadrados Reponderados Iterativo

O método de Mínimos Quadrados Reponderados Iterativo será utilizado para estimar o vetor de médias $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$. O algoritmo é iterativo e os parâmetros na iteração $(m+1)$ podem ser calculados de acordo com a equação (Cordeiro e Demétrio, 2010)

$$\boldsymbol{\beta}^{(m+1)} = [\mathbf{X}'\mathbf{A}^{(m)}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{A}^{(m)}\mathbf{z}^{(m)} \, , \tag{2.16}$$

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

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{2.17}$$

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

$$a_i = \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \, , \tag{2.18}$$

onde $V(\mu_i) = b''(\theta)$.

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 $\hat{\boldsymbol{\beta}}$, sem a necessidade do processo iterativo,

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} \ . \tag{2.19}$$

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), ou seja,

$$Dev = 2\sum_{i=1}^{n} \{L(y_i; y_i) - L(\mu_i; y_i)\} \ . \tag{2.20}$$

A Tabela 2.2 apresenta o algoritmo MQRI.

Tabela 2.2: Algoritmo MQRI

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

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 2.1), e estimar μ por meio do algoritmo MQRI (Tabela 2.2).

Capítulo 3

Regressão Geograficamente Ponderada

3.1 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). 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, a Regressão Geograficamente Ponderada para a distribuição Normal é apresentada. Por fim, o modelo de RBNGP proposto neste trabalho é detalhado.

3.2 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, respecti-

vamente. A fim de descrever estes índices, é necessário compreender o conceito de matriz de proximidade espacial.

3.2.1 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 pela malha rodoviária (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 3.1.

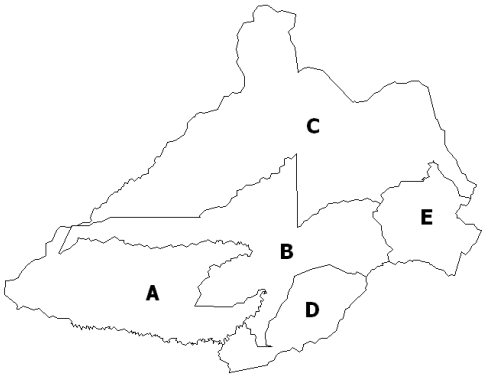

Figura 3.1: Exemplo de configuração espacial

$$\mathbf{W} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array} \qquad \mathbf{W}_p = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 0,5 & 0 & 0,5 & 0 \\ 0,25 & 0 & 0,25 & 0,25 & 0,25 \\ 0 & 0,5 & 0 & 0 & 0,5 \\ 0,33 & 0,33 & 0 & 0 & 0,33 \\ 0 & 0,5 & 0,5 & 0 & 0 \end{pmatrix} \end{array}$$

Note, por exemplo, que a área E apresenta dois vizinhos (áreas B e C). Consequentemente, a última linha das matrizes $\mathbf{W}$ e $\mathbf{W}_p$ contêm, respectivamente, pesos de 1 e $0,5$ nas colunas referentes às áreas B e C.

3.2.2 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 (3.1), 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 $\mathbf{W}$.

$$I = \frac{n}{\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.1}$$

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), $I = 1$ representa autocorrelação positiva máxima e $I = -1$ autocorrelação negativa máxima. Nota-se, por meio da Equação (3.1), 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, dado por

$$C = \frac{n-1}{2\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \, . \tag{3.2}$$

O intervalo de variação deste índice é de 0 a 2, sendo $C = 1$ ausência de autocorrelação espacial (novamente, com referência à 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 pseudo-significância (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. Contabiliza-se, então, o número s de vezes que o índice calculado foi mais extremo do que o valor de fato observado na amostra original. O p-valor do teste é

obtido pela razão $s/(m+1)$. Por não fazer pressupostos a respeito da distribuição de probabilidade dos índices, este é o teste mais utilizado.

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, é mais indicado utilizar os índices locais de autocorrelação.

3.2.3 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.

Com esta motivação, 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, respectivamente, por

$$I_i = \frac{n \times z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{i=1}^{n} z_j^2} \ , \tag{3.3}$$

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

$$C_i = \frac{\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ . \tag{3.4}$$

A significância estatística desses índices pode ser verificada por meio de 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*).

3.2.4 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 3.2 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.

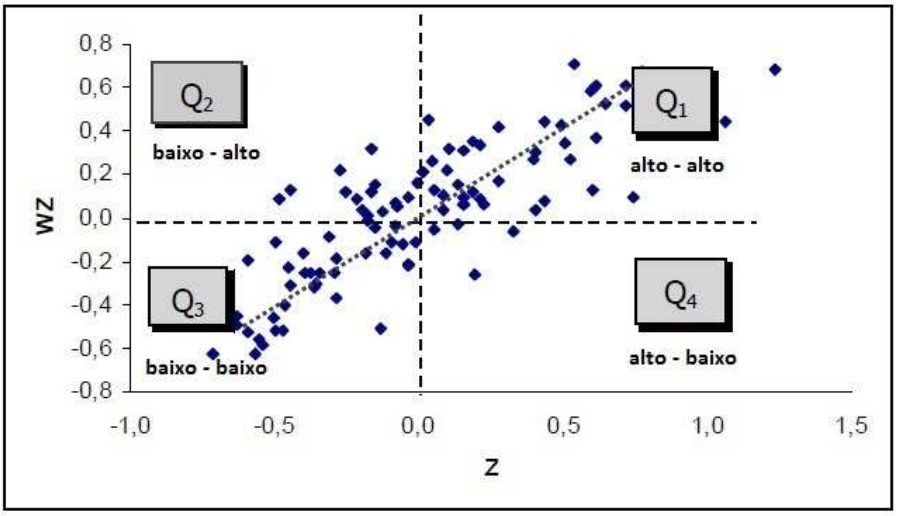

Figura 3.2: Diagrama de Espalhamento de Moran
Fonte: Druck et al. (2004) com modificações

O índice de Moran, apresentado na Equação (3.1), tem sua forma matricial dada por

$$I = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{W}\mathbf{z} \; . \tag{3.5}$$

Nota-se, a partir da Equação (3.5), que o índice de Moran é 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.

3.3 Regressão Geograficamente Ponderada

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. Sendo assim, o objetivo da RGP é fornecer estimativas não paramétricas destas superfícies contínuas utilizando a função kernel.

3.3.1 Modelo

O modelo RGP está apresentado a seguir,

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \ , \tag{3.6}$$
$$\varepsilon_i \sim N(0, \sigma^2) \ .$$

Note que os pressupostos do modelo de regressão clássica (erros normais, homocedásticos e não correlacionados) permanecem. No entanto, permitido-se 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 2.7) é um caso especial da regressão geograficamente ponderada (Equação 3.6). Esta simplificação ocorre quando não há variação espacial nos parâmetros.

A forma matricial da Equação (3.6) é dada por

$$\mathbf{y} = (\boldsymbol{\beta} \otimes \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon} \ , \tag{3.7}$$

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 i contém a estimativa dos $(k+1)$ parâmetros para a amostra i, ou seja,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \dots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \dots & \beta_k(u_2, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \dots & \beta_k(u_n, v_n) \end{bmatrix} . \tag{3.8}$$

A estimação dos parâmetros do modelo (3.6) é feita utilizando o método de mínimos quadrados ponderados (Fotheringham et al., 2002). Este método foi abordado de forma mais geral na Seção 2.3.2, 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

(2.16) simplifica-se para

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{y} \ , \tag{3.9}$$

onde $\widehat{\boldsymbol{\beta}}(u_i, v_i)$ é a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ no ponto (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 i.

Denotando $\widehat{\boldsymbol{\beta}}(u_i, v_i)$ por $\widehat{\boldsymbol{\beta}}(i)$ e $\mathbf{W}(u_i, v_i)$ por $\mathbf{W}(i)$, a Equação (3.9) pode ser reescrita como

$$\widehat{\boldsymbol{\beta}}(i) = [\mathbf{X}'\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(i)\mathbf{y} \ , \tag{3.10}$$

onde

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_{in} \end{bmatrix} . \tag{3.11}$$

Portanto, a matriz de pesos $\mathbf{W}(i)$ da Equação (3.11) deve ser calculada para cada ponto i. As possibilidades de escolha da matriz $\mathbf{W}(i)$ serão apresentadas a seguir.

3.3.2 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 3.3 apresenta um exemplo desta função.

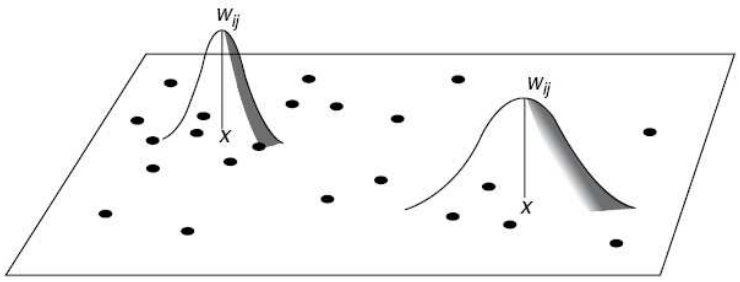

Figura 3.3: Função de ponderação espacial
Fonte: Fotheringham et al. (2002)

A seguir, estão apresentadas algumas possibilidades de escolha para a função de ponderação (Fotheringham et al., 2002).

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.

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 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. Note que se $w_{ij} = 1$ $\forall i, j$, então chega-se ao modelo de regressão clássica global apresentado na Seção 2.2.1.

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.

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.

3.3.3 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),

$$CV = \sum_{j=1}^{n} \left[y_j - \hat{y}_{\neq j}(b) \right]^2 \ , \tag{3.12}$$

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 (3.12) é o parâmetro de suavização ótimo do método de validação cruzada. Note que a Equação (3.12) 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

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

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) foi determinado para a RGP por Hurvich et al. (1998), sendo dado por

$$AIC_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + \frac{n(n + tr(\mathbf{H}))}{n - 2 - tr(\mathbf{H})}\ , \tag{3.13}$$

onde $\hat{\sigma}$ é a estimativa de máxima verossimilhança,

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

e $\mathbf{H}$ é a matriz de projeção (ou do inglês, *hat matrix*), cujas linhas $\boldsymbol{h_j}$ são dadas por

$$\boldsymbol{h_j} = \mathbf{X_j}[\mathbf{X}'\mathbf{W}(j)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(j)\ , \tag{3.14}$$

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

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 parâmetros como, aproximadamente, o traço da matriz $\mathbf{H}$, denotado por $tr(\mathbf{H})$ (Fotheringham et al., 2002), assim como ocorre nos Modelos Lineares Generalizados.

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 2.2.1). Fotheringham et al. (2002) comentam que, por ser um critério mais geral, ele é mais recomendado.

3.4 Regressão Binomial Negativa Geograficamente Ponderada

Com base na metodologia desenvolvida por Nakaya et al. (2005) para a Regressão de Poisson Geograficamente Ponderada - RPGP (ou do inglês, Geographically Weighted Poisson Regression - GWPR), desenvolvemos neste trabalho a Regressão Binomial Negativa Geograficamente Ponderada.

Como a distribuição Binomial Negativa apresenta dois parâmetros (α e $\boldsymbol{\beta}$), enquanto que a de Poisson tem apenas o $\boldsymbol{\beta}$, considerou-se que o parâmetro α da distribuição Binomial Negativa não varia espacialmente. Esta consideração foi feita com o intuito de simplificar o modelo. Sendo assim, o α será estimado de forma global, ou seja, de acordo com a regressão Binomial Negativa global. O método de estimação do $\boldsymbol{\beta}$ será explicado a seguir.

De acordo com a regressão Binomial Negativa apresentada na Seção 2.2.2, tem-se que

$$\log(\mu_i) = \sum_k \beta_k x_{ik} \ ,$$

então,

$$\mu_i = \exp\left(\sum_k \beta_k x_{ik}\right) .$$

Permitindo variação espacial aos parâmetros β_k, tem-se que

$$\mu_i = \exp\left(\sum_k \beta_k(u_i, v_i) x_{ik}\right) . \tag{3.15}$$

Assim, o modelo de RBNGP modelado em termos da média μ_i é dado por

$$y_i \sim \text{BN}\left[\exp\left(\sum_k \beta_k(u_i, v_i) x_{ik}\right), \alpha\right] . \tag{3.16}$$

Note que, enquanto que a média da distribuição varia espacialmente, o parâmetro α é mantido constante.

O método escore de Fisher (vide Seção 2.3.2) modificado fornece a solução para a estimação dos parâmetros do modelo (3.16). 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 $\mathbf{A}$ do MQRI pela matriz de pesos $\mathbf{W}(i)$ da RGP (Fotheringham et al., 2002). A estimativa de $\boldsymbol{\beta}(u_i, v_i)$ no ponto i na iteração $(m+1)$ é dada por

$$\widehat{\boldsymbol{\beta}}(u_i, v_i)^{(m+1)} = [\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)^{(m)}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{A}(u_i, v_i)^{(m)}\mathbf{z}(u_i, v_i)^{(m)} , \tag{3.17}$$

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

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \tag{3.18}$$

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

$$\mathbf{W}(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{pmatrix}, \tag{3.19}$$

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

$$\mathbf{A}(u_i, v_i) = \begin{pmatrix} a_{i1} & 0 & \dots & 0 \\ 0 & a_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{in} \end{pmatrix}. \tag{3.20}$$

Os elementos a_{ij} da diagonal ($j = 1, \dots, n$) são obtidos por meio de (2.18),

$$a_{ij} = \frac{1}{V(\mu_j)} \left(\frac{\partial \mu_j}{\partial \eta_j} \right)^2 = \frac{\mu_j(\boldsymbol{\beta}(i)^{(m)})}{1 + \alpha \times \mu_j(\boldsymbol{\beta}(i)^{(m)})}, \tag{3.21}$$

onde $\mu_j(\boldsymbol{\beta}(i)^{(m)})$ é dado por

$$\mu_j(\boldsymbol{\beta}(i)^{(m)}) = \exp\left(\sum_k \beta_k(u_i, v_i)^{(m)} x_{jk} \right). \tag{3.22}$$

Por fim, $\mathbf{z}(u_i, v_i)^{(m)}$ é o vetor da variável dependente ajustada do algoritmo MQRI para o ponto i, cujos elementos $z_{ij}^{(m)}$ ($j = 1, \dots, n$) foram apresentados na Equação

(2.17). Para a Binomial Negativa, tem-se que:

$$\mathbf{z}_{ij}^{(m)} = X\boldsymbol{\beta}(i)^{(m)} + \frac{y_j - \mu_j(\boldsymbol{\beta}(i)^{(m)})}{a_{ij}(1 + \alpha \times \mu_j(\boldsymbol{\beta}(i)^{(m)}))} \, . \tag{3.23}$$

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

Capítulo 4

Resultados

4.1 Introdução

A Regressão Binomial Negativa Geograficamente Ponderada e a Regressão de Poisson Geograficamente Ponderada foram implementadas em linguagem SAS/IML e o código encontra-se no Apêndice. Os modelos implementados foram aplicados na análise da distribuição da oferta de veículos rodoviários de carga do tipo caminhão simples no Estado do Espírito Santo, que foi explicada em função da quantidade de estabelecimentos do ramo da indústria. A unidade espacial utilizada foi a divisão municipal, que é composta por 77 municípios.

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.

Este capítulo apresenta a RBNGP aplicada a esse estudo de caso na área de Transportes. Inicialmente, uma análise exploratória é realizada a fim de avaliar a

dependência espacial e a estacionariedade dos dados. Visto que a variável dependente (frota de caminhão simples) é de contagem, os modelos de regressão Binomial Negativa e de Poisson são apresentados. Em seguida, os modelos espaciais locais de RBNGP e de RPGP são construídos e comparados, tanto entre si quanto com os de regressão global. Por fim, mostra-se que a RBNGP generaliza a regressão Binomial Negativa, a regressão de Poisson e a RPGP.

4.2 Análise exploratória

Com o intuito de verificar a dependência espacial, foi gerado um mapa coroplético da variável dependente a fim de observar tendências espaciais. Neste texto, a variável frota de caminhão simples será chamada de *Frota* e a variável número de indústrias será denominada *Indústrias*. A Figura 4.1 apresenta os mapas dessas variáveis utilizando os quintis para definir as classificações.

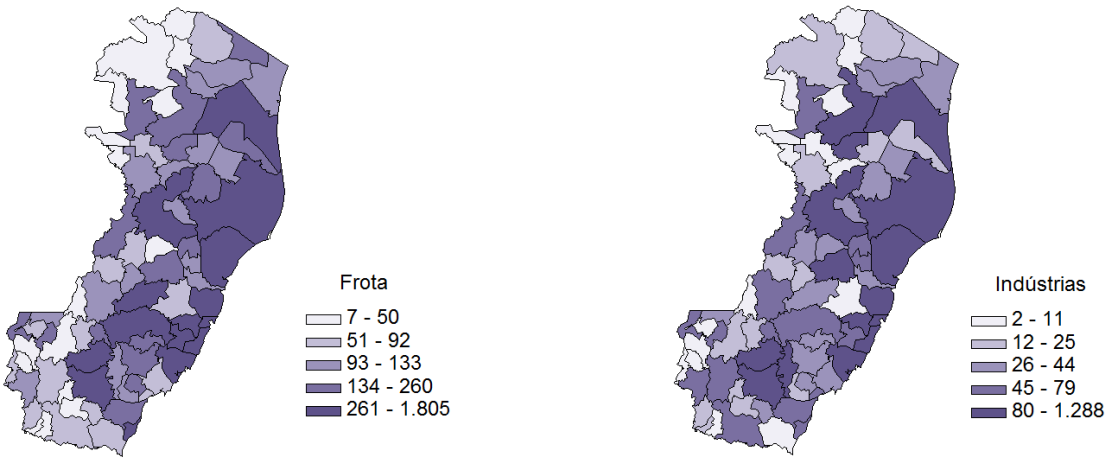

Figura 4.1: Mapa da variável dependente *Frota* e da variável independente *Indústrias*

É possível observar, a partir da Figura 4.1, 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 litorânea 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 (vide Seção 3.2.2) para a frota de caminhão simples no Estado do Espírito Santo. A matriz de proximidade espacial **W** utilizada (vide Seção 3.2.1) foi binária, indicando se a área A_i faz fronteira com a área A_j. Esta escolha foi feita devido a sua simplicidade. O valor obtido foi $I = 0,23$, conforme indicado na Figura 4.2, que ilustra o diagrama de espalhamento de Moran (vide Seção 3.2.4). Este índice caracteriza uma dependência espacial baixa com respeito a matriz **W** binária.

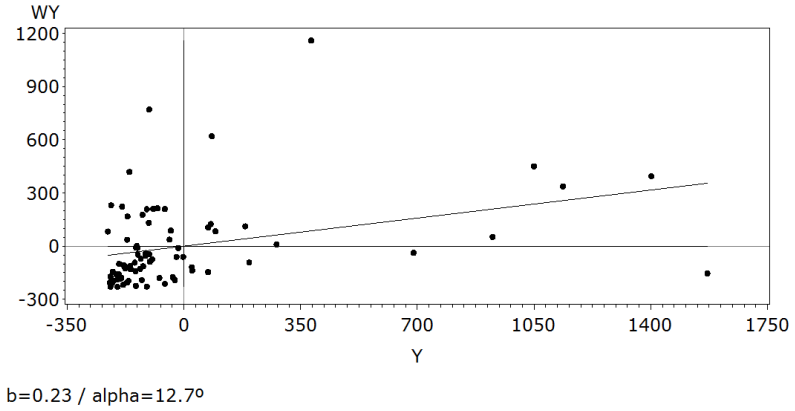

Figura 4.2: Diagrama de espalhamento de Moran

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 o mapa de espalhamento de Moran apresentado na Figura 4.3. Os municípios coloridos em tons de vermelho apresentam dependência espacial positiva (ou seja, estão no primeiro ou terceiro quadrantes da Figura 4.2), enquanto que os municípios em tons de azul tem dependência espacial negativa (quadrantes dois e quatro da Figura 4.2).

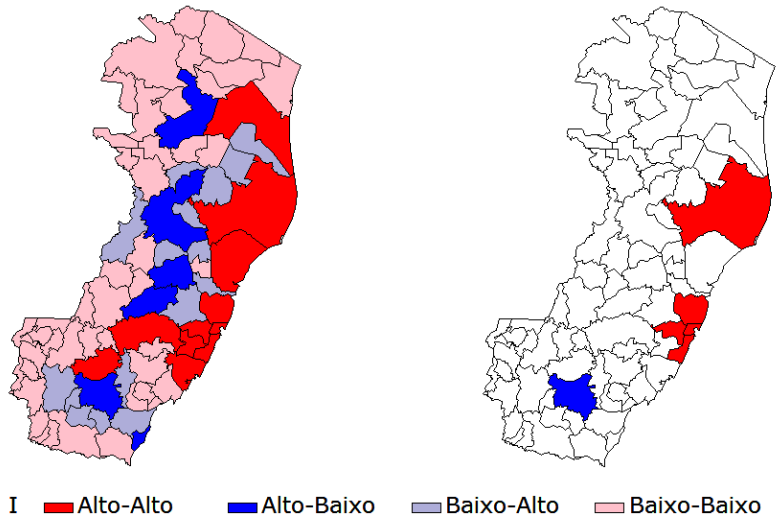

Figura 4.3: Mapa de espalhamento de Moran (esquerda) e Mapa de Moran 95% (direita)

A partir do mapa de espalhamento de Moran da Figura 4.3, notamos 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, não é adequado utilizar o índice global de Moran para caracterizar a dependência espacial. Além disso, um modelo espacial local aparenta ser mais indicado.

4.3 Regressão global

Iniciaremos a modelagem estatística pelos modelos de regressão mais simples para dados de contagem, que são a Regressão de Poisson e a Regressão Binomial Negativa. Visto que nesta seção estamos visualizando os dados de maneira global, é útil analisar o histograma e o boxplot da variável dependente $Frota$, os quais estão na Figura 4.4.

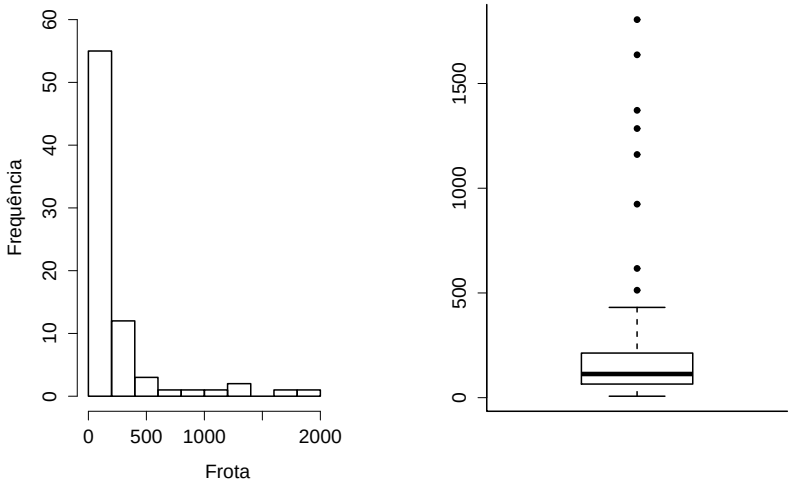

Figura 4.4: Histograma e Boxplot da variável $Frota$

Nota-se, pela Figura 4.4, que a variável $Frota$ tem uma distribuição assimétrica positiva e apresenta muitos outliers. Considerando que sua média é de $234,4$ caminhões simples, enquanto que sua variância é de $129.516,2$, temos uma variável com superdispersão.

Os resultados dos ajustes da regressão de Poisson e da regressão Binomial Negativa estão apresentados na Tabela 4.1. As colunas "Intercepto", "Indústria" e "Dispersão" indicam as estimativas pontuais dos parâmetros β_0, β_1 e α da regressão, respectivamente. Já a coluna "Par." indica o número de parâmetros estimado em

cada modelo. Note que para a Poisson temos 2 parâmetros (β_0 e β_1), enquanto para a Binomial Negativa temos 3, devido ao parâmetro α de superdispersão.

Tabela 4.1: Estimativas das regressões de Poisson e Binomial Negativa

Regressão	Intercepto	Indústria	Dispersão	Par.	Desvio	AICc
Poisson	4,9517	0,0023	0	2	9493,84	10006,9
Bin. Negativa	4,6554	0,0038	0,5156	3	83,32	923,3

As estatísticas desvio e AICc são medidas de qualidade do ajuste, com valores menores indicando um modelo mais bem ajustado. O AICc também considera a complexidade do modelo pois, além da log-verossimilhança, leva em conta a quantidade de parâmetros envolvida. Analisando a Tabela 4.1, verifica-se que tanto o AICc quanto o desvio sofreram grande redução (mais de 90%) da regressão de Poisson para a Binomial Negativa. De fato, dados de contagem que apresentam superdispersão são melhor ajustados pela distribuição Binomial Negativa.

4.4 Regressão Geograficamente Ponderada

Na análise exploratória foram constatados indícios de não estacionariedade espacial. Sendo assim, apresentamos, nesta seção, os modelos espaciais locais de RBNGP e RPGP.

4.4.1 Regressão Binomial Negativa Geograficamente Ponderada

Conforme explicado na Seção 3.3.3, os resultados da RGP dependem da estimação do parâmetro de suavização b. A escolha deste parâmetro pode ser feita de forma a minimizar uma medida da qualidade do ajuste do modelo. Neste trabalho,

decidiu-se pela minimização do AICc, por ser um critério mais geral na seleção de modelos. O algoritmo de minimização utilizado foi da divisão áurea (Zörnig, 2009). Além disso, optou-se pela escolha do parâmetro de suavização de forma fixa, ou seja, um mesmo b para todas as regiões. Esta escolha não leva em conta a concentração de pontos, sendo feita com o intuito de simplificar o algoritmo de minimização. A Figura 4.5 apresenta a busca ótima do parâmetro de suavização que minimiza o AICc.

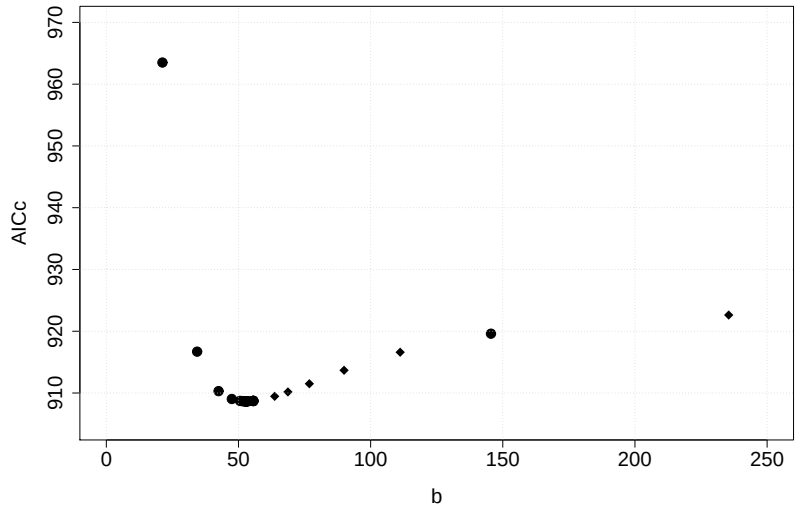

Figura 4.5: Parâmetro de suavização b da RBNGP que minimiza o AICc

Como resultado do método da divisão áurea da Figura 4.5, temos que o parâmetro de suavização ótimo é $b = 53,0684$ com um AICc de $908,66$. Considerando que a função de poderação espacial escolhida foi a gaussiana,

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

onde $b = 53,0684$, temos a superfície das estimativas dos parâmetros da Regressão Binomial Negativa Geograficamente Ponderada ilustrada na Figura 4.6.

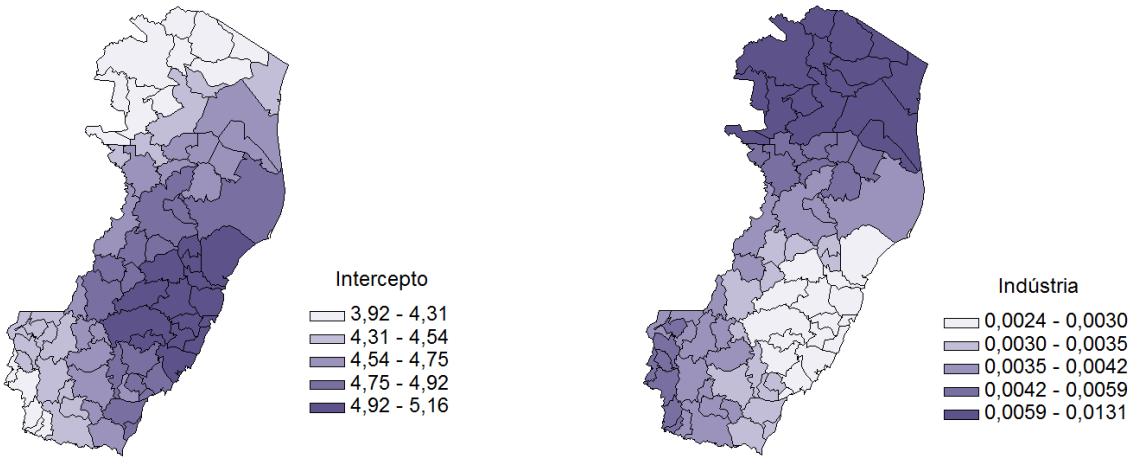

Figura 4.6: Superfície das estimativas dos parâmetros da RBNGP

O parâmetro de superdispersão, considerado constante na RBNGP proposta neste trabalho, foi $\alpha = 0,5156$, que é idêntico, por construção, ao parâmetro de dispersão da regressão Binomial Negativa global (Tabela 4.1). Já as estimativas dos parâmetros β_0 e β_1, apresentadas na Figura 4.6, variam espacialmente. Note que os valores estimados para o intercepto do modelo são mais elevados no Sudeste do Estado do Espírito Santo, onde se localiza a capital - Vitória, refletindo a maior concentração de caminhões simples nestes lugares. Já os valores de $\widehat{\beta}_1(u_i, v_i)$ são menores nessa região devido ao grande número de indústrias x_{i1} associado a forma exponencial do modelo,

$$\mu_i = \exp\left(\beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1}\right) .$$

Por exemplo, considere a equação $\mu = \exp(4 + 0,0035x)$, então um aumento de uma unidade em x, aumenta a média em $0,2$ se $x = 10$, ou aumenta μ em 6 unidades se $x = 1000$. Consequentemente, na região Sudeste, onde o número de indústrias é muito elevado, o parâmetro $\beta_1(u_i, v_i)$ é naturalmente mais baixo, não porque um

aumento marginal de 1 indústria tem efeito menor no número de caminhões na região Sudeste, mas sim pela forma exponencial presente na modelagem da média.

Além da visualização do mapa, é útil também apresentar algumas estatísticas (mínimo, média, máximo e quartis) das estimativas. Este sumário encontra-se na Tabela 4.2, cuja última coluna repete os valores da regressão Binomial Negativa global da Tabela 4.1.

Tabela 4.2: Sumário das estimativas dos parâmetros da RBNGP

Parâmetro	Mínimo	Q1	Q2	Média	Q3	Máximo	Global
Intercepto	3,92	4,39	4,63	4,62	4,87	5,16	4,6554
Indústria	0,0024	0,0032	0,0039	0,0048	0,00517	0,0131	0,0038
Dispersão	0,5156	0,5156	0,5156	0,5156	0,5156	0,5156	0,5156

Note, pela Tabela 4.2, que o modelo global captou essencialmente a mediana da variação espacial das estimativas dos parâmetros. Enquanto que a RBNGP fez uma modelagem local, levando em conta a dependência e a não estacionariedade espacial. Com a regressão espacial local, o desvio do modelo foi reduzido de $83,32$ para $55,19$, já o AICc caiu de $923,3$ para $908,66$. Lembrando que diferenças maiores do que 3 no AICc são consideradas significativas (vide Seção 3.3.3), pode-se afirmar que o modelo de RBNGP apresentou um melhor ajuste.

4.4.2 Regressão de Poisson Geograficamente Ponderada

Apesar da regressão de Poisson não ter se mostrado um bom ajuste aos dados, construiremos nesta seção a RPGP a fim de verificar os possíveis avanços que a modelagem espacial local pode trazer. A Figura 4.7 apresenta o resultado do algoritmo da divisão áurea na minimização do AICc.

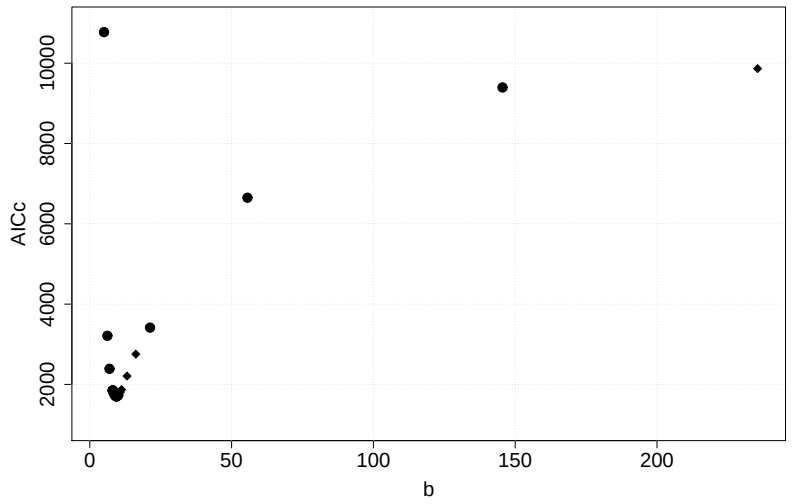

Figura 4.7: Parâmetro de suavização b da RPGP que minimiza o AICc

O parâmetro de suavização $b = 9,38$ quilômetros resultou no AICc mínimo de AICc $= 1705,1$. 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. Com isso, as regressões locais são feitas com base em um número pequeno de pontos (de 1 a 9). Este é mais um indício da inadequabilidade da regressão de Poisson para modelar a frota de caminhões simples.

Para contornar este problema, optou-se por não utilizar o b ótimo, e sim $b = 53,068$, igual ao da RBNGP. Assim, temos regressões locais com número de pontos variando entre 22 e 65 e AICc $= 6466$. Apesar do aumento do AICc em relação ao valor ótimo, a RPGP ainda é um melhor ajuste se comparada com a regressão global de Poisson, cujo AICc era de $10006,9$.

A Figura 4.8 apresenta os mapas das estimativas dos parâmetros da RPGP para $b = 53,068$. Note que as superfícies foram feitas com a mesma escala da Figura 4.6 a fim de facilitar a comparação dos modelos.

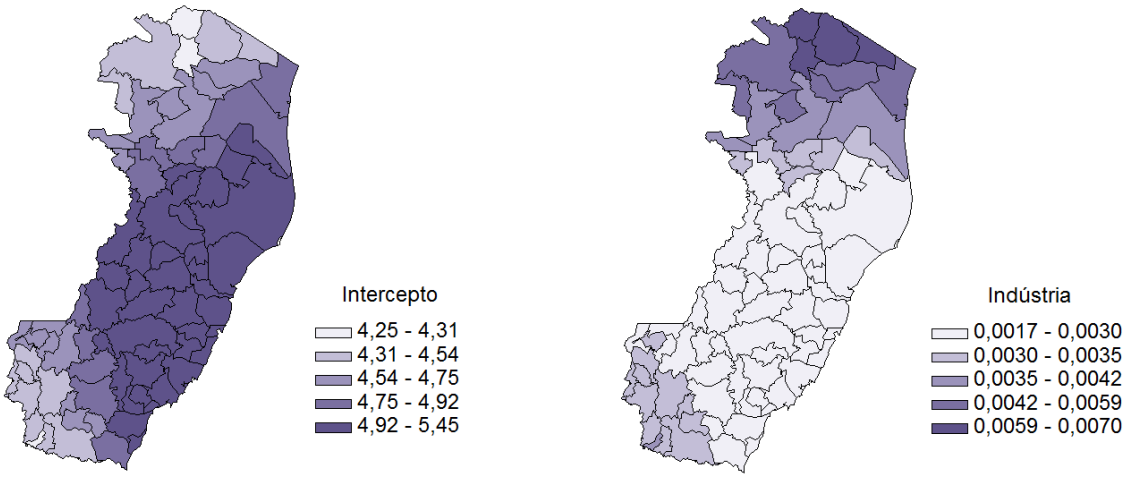

Figura 4.8: Superfície das estimativas dos parâmetros da RPGP

Comparando as superfícies das estimativas da RPGP da Figura 4.8 com as da RBNGP da Figura 4.6 notamos muitas diferenças. Em geral, a RPGP apresenta valores mais elevados para o intercepto do modelo e menores para as estimativas do parâmetro *Indústrias*. Além disso, a RPGP considera que $\alpha = 0$, enquanto que na RBNGP $\alpha = 0,5156$. A Tabela 4.3 apresenta algumas estatísticas descritivas das estimativas pontuais da RPGP. A última coluna da tabela repete os valores da regressão de Poisson (Tabela 4.1).

Tabela 4.3: Sumário das estimativas dos parâmetros da RPGP

Parâmetro	Mínimo	Q1	Q2	Média	Q3	Máximo	Global
Intercepto	4,25	4,57	4,91	4,87	5,14	5,44	4,95
Indústria	0,0017	0,0022	0,0027	0,0030	0,0035	0,007	0,0023
Dispersão	0	0	0	0	0	0	0

A fim de comparar a qualidade do ajuste e a complexidade dos modelos apresentados, a Tabela 4.4 traz algumas medidas relevantes. A coluna "Dif. Desvio" indica a diferença entre os desvios dos modelos com respeito a RBNGP. E a coluna "Dif. AICc" apresenta a diferença entre os AICc, tendo como referência também a

RBNGP.

Tabela 4.4: Comparação entre modelos

Modelo	Par.	Desvio	Dif. Desvio	AICc	Dif. AICc
Poisson	2	9493,8	9438,6	10006,9	9098,2
RPGP	6,24	5943	5887,8	6466	5557,3
Bin. Negativo	3	83,3	28,1	923,3	14,6
RBNGP	8,67	55,2	0	908,7	0

Analisando a Tabela 4.4, verificamos que o modelo de RBNGP é o que apresenta o maior número $efetivo$ de parâmetros, sendo este calculado pelo traço da matriz **H** (vide Seção 3.3.3). No entanto, além de possuir o menor desvio, a RBNGP tem o menor AICc. Vale lembrar que o AICc não só mede a qualidade do ajuste, mas também considera o seu grau de complexidade, ou seja, a sua quantidade de parâmetros. Sendo assim, conclui-se que o modelo mais indicado para descrever a frota de caminhões simples no Estado do Espírito Santo, em função da quantidade de estabelecimentos industriais, é a Regressão Binomial Negativa Geograficamente Ponderada.

4.5 Casos particulares

O modelo de Regressão Binomial Negativa Geograficamente Ponderada apresenta a vantagem de permitir uma modelagem espacial local de dados de contagem com superdipersão. Além disso, esse modelo generaliza a regressão global - Binomial Negativa e de Poisson - e a RPGP. Nesta seção apresentamos como ocorrem essas generalizações.

4.5.1 Regressão global

É simples visualizar porque a regressão global é um caso particular da RBNGP. Considere a função de ponderação espacial gaussiana,

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

e note que a medida que o parâmetro de suavização b cresce, os pesos w_{ij} da diagonal de $W(i)$ se aproximam da unidade, chegando-se assim à regressão global. Isso ocorre para todas as funções de ponderação espacial.

Um exemplo prático dessa generalização está ilustrado na Figura 4.9, que mostra o comportamento das estatísticas média (linha preta), mínimo e máximo (linhas vermelhas tracejadas) das estimativas dos parâmetros da RBNGP em função do parâmetro de suavização b. A linha azul é a estimativa do parâmetro da regressão Binomial Negativa global.

Analisando a Figura 4.9, verifica-se que, a partir de $b = 200$ quilômetros, as estimativas dos parâmetros das regressões local e global são praticamente idênticas, confirmando que a regressão Binomial Negativa é um caso particular da RBNGP. Além disso, temos que, em geral, as estimativas da regressão global (linha azul) são próximas da média das estimativas da RBNGP (linha preta). Ou seja, a regressão global modela um comportamento médio, sendo ineficiente para descrever as peculiaridades locais de cada região.

O exemplo da frota de caminhões simples é um caso no qual a RBNGP fornece um melhor ajuste. No entanto, como seria a modelagem pela RBNGP de um conjunto

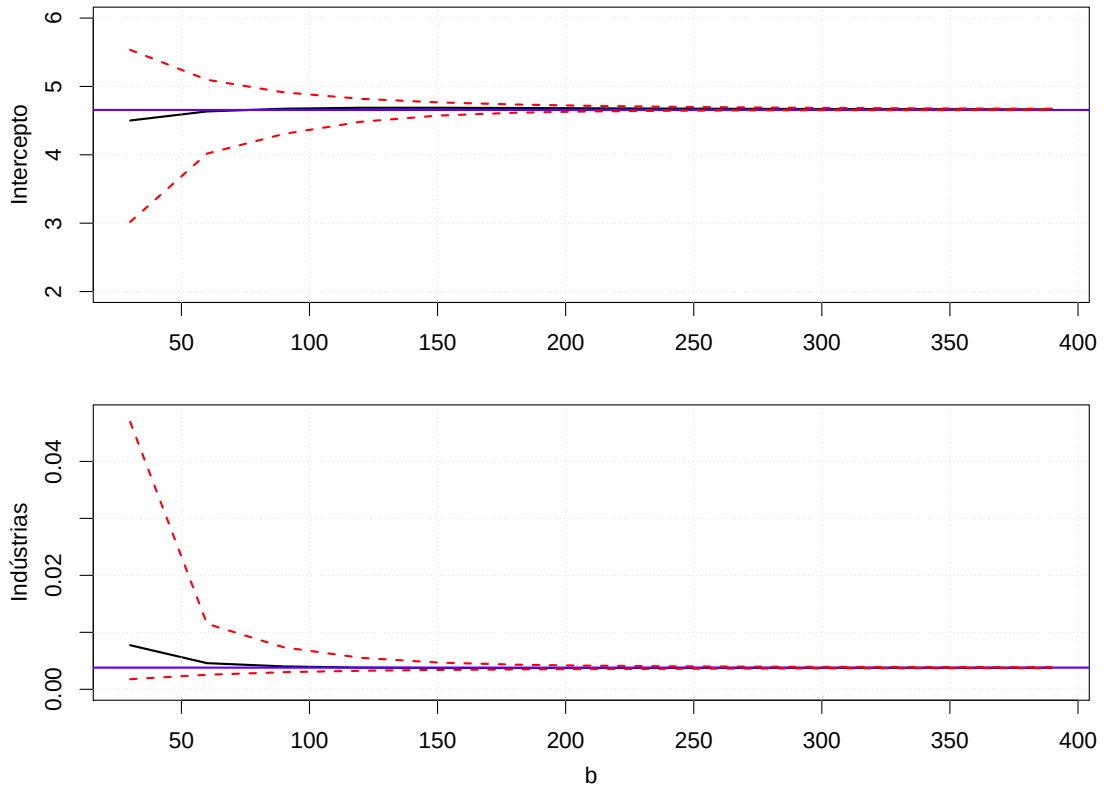

Figura 4.9: Comparação da estimativa do parâmetro da regressão global (linha azul) com as estatísticas média (linha preta), mínimo e máximo (linhas vermelhas tracejadas) das estimativas dos parâmetros da RBNGP em função do parâmetro de suavização b

de dados cujos parâmetros não variam espacialmente? Será que a RBNGP é capaz de detectar que os parâmetros são, na verdade, constantes? A fim de responder a essas perguntas, foi gerado um conjunto de dados Binomial Negativo com $\beta_0 = 1$, $\beta_1 = 0,5$ e $\alpha = \frac{1}{3}$ e modelado pela RBNGP. A Figura 4.10 apresenta o resultado do algoritmo da divisão áurea na determinação do parâmetro de suavização ótimo.

A partir da Figura 4.10, tem-se que b ótimo é $380,91$ quilômetros e AICc $= 343,94$. No entanto, note que o ponto de mínimo foi encontrado no extremo do intervalo, que é a maior distância existente entre os municípios do Estado do Espírito Santo. Com isso, a RBNGP está indicando que o mais adequado é incluir todos os pontos na regressão local ou, em outras palavras, que o modelo global deve ser mais

apropriado nesta modelagem.

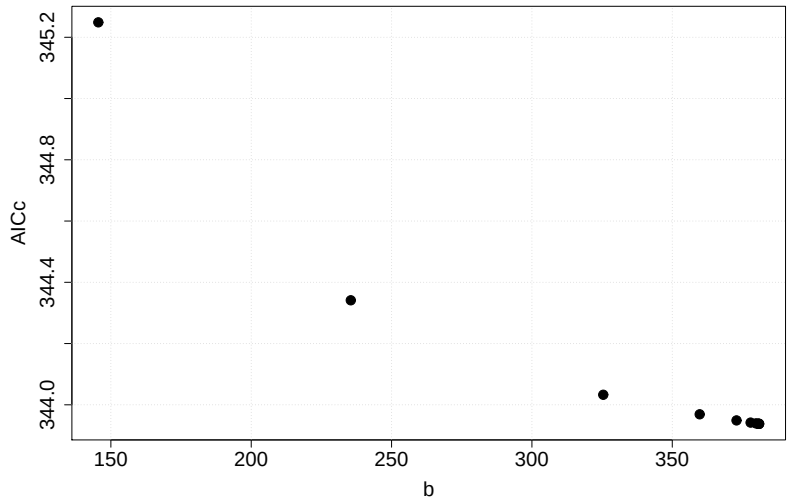

Figura 4.10: Determinação do parâmetro de suavização b que minimiza o AICc

De fato, a regressão Binomial Negativa global fornece AICc $= 343,66$ e, como a diferença entre os AICc não é significativa, optou-se pelo modelo mais parcimonioso. Sendo assim, a RBNGP indica um modelo de regressão global quando b converge para a distância máxima máxima e o AICc não é significativo.

4.5.2 Regressão de Poisson Geograficamente Ponderada

Conforme já foi dito, a distribuição Binomial Negativa com $\alpha \to 0$ tende para a distribuição de Poisson. Com base nisso, no algoritmo implementado (vide Apêndice), criamos a possibilidade do parâmetro α ser fornecido externamente pelo analista por meio da variável macro *alphag*. Sendo assim, a Figura 4.11 apresenta o resultado da RBNGP com $\alpha = 10^{-8}$.

Comparando os mapas da RPGP (Figura 4.8) com os mapas da RBNGP com $\alpha = 10^{-8}$ (Figura 4.11), feitos na mesma escala, verifica-se que ambos são equivalentes.

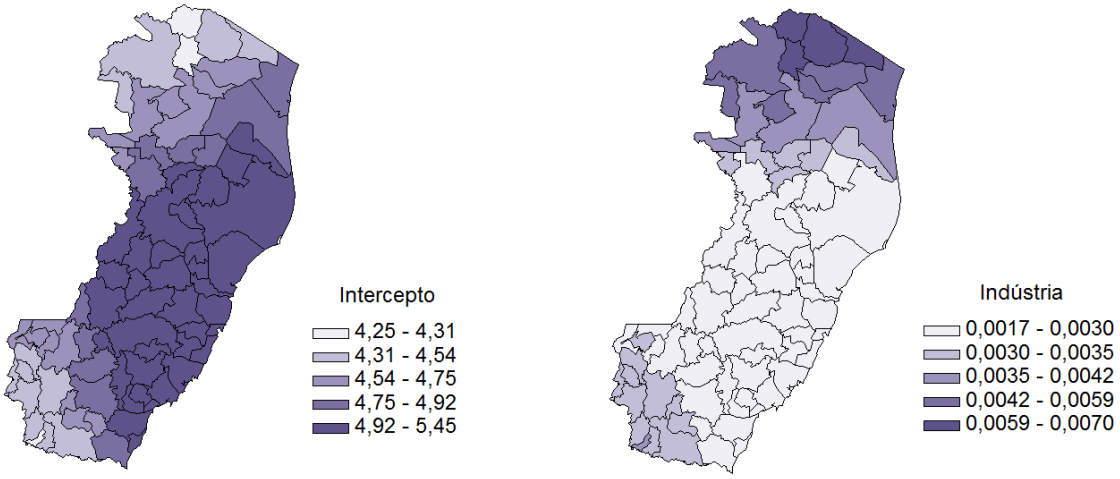

Figura 4.11: Superfície das estimativas dos parâmetros da RBNGP com $\alpha = 10^{-8}$

Este resultado também pode ser observado comparando-se a Tabela 4.5 da RBNGP com $\alpha = 10^{-8}$ com a Tabela 4.3 da RPGP.

Tabela 4.5: Sumário das estimativas dos parâmetros da RBNGP com $\alpha = 10^{-8}$

Parâmetro	Mínimo	Q1	Q2	Média	Q3	Máximo	Global
Intercepto	4,25	4,57	4,91	4,87	5,14	5,44	4,95
Indústria	0,0017	0,0022	0,0027	0,0030	0,0035	0,007	0,0023
Dispersão	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}

A última coluna da Tabela 4.5 foi feita modelando a RBNGP com $\alpha = 10^{-8}$ e $b = 1000$. Note que as estimativas são as mesmas da Regressão global de Poisson (última coluna da Tabela 4.3). Ou seja, fazendo $\alpha \to 0$ e b grande, tem-se que a RBNGP é equivalente a regressão de Poisson.

Portanto, com o algoritmo da Regressão Binomial Negativa desenvolvido é possível realizar, além dela própria, a regressão Binomial Negativa, a regressão de Poisson e a Regressão de Poisson Geograficamente Ponderada.

Capítulo 5

Conclusões

O objetivo desse trabalho foi desenvolver o modelo de Regressão Binomial Negativa Geograficamente Ponderada, a fim de modelar dados de contagem não estacionários e com superdispersão. Por simplicidade, considerou-se que o parâmetro α da RBNGP não varia espacialmente. Com isso, seu valor é igual ao da regressão Binomial Negativa global. O algoritmo da RBNGP foi implementado em linguagem SAS/IML.

A RBNGP foi utilizada para modelar a frota de caminhões simples no Estado do Espírito Santo em função da quantidade de estabelecimentos industriais. As estatísticas de qualidade do ajuste indicaram que a RBNGP foi mais adequada do que os modelos concorrentes, a saber, regressão global - Binomial Negativa e Poisson - e Regressão de Poisson Geograficamente Ponderada.

Além disso, mostrou-se que a RBNGP generaliza a regressão Binomial Negativa e a Regressão de Poisson Geograficamente Ponderada, utilizando, para isso, os dados da frota de caminhões simples do Estado do Espirito Santo.

Para trabalhos futuros, sugere-se testes com dados simulados a fim de confirmar

a validade do modelo e do algoritmo implementado. Outro aprimoramento é a elaboração do modelo de RBNGP com α estimado de forma local. Além disso, seria interessante utilizar um método de determinação do parâmetro de suavização b que leve em conta a dispersão espacial dos dados. De forma que um b pequeno (grande) fosse atribuído aonde os dados estivessem mais (menos) concentrados. Por fim, o cálculo dos erros padrão das estimativas dos parâmetros também traria enriquecimentos ao modelo.

Referências Bibliográficas

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.

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

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

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.

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

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B*, 60:271–293.

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.

Zörnig, P. (2009). *Introdução à Programação Não-Linear*. Não publicado.

Apêndice

Código SAS

```
%macro gwnbr(tab=,y=,x=,lat=,long=,h=,grid=,latg=,longg=,gwr=,method=, alphag=);
proc iml;
use &tab;
read all var {&y} into y;
read all var {&x} into x;
read all var{&long &lat} into COORD;
close &tab;
use &grid;
read all var{&longg &latg} into POINTS;
close &grid;
h=&h;
gwr="&gwr";
method="&method";
n=nrow(y);
x=j(n,1,1)||x;
m=nrow(POINTS);
bii=j(ncol(x)*m,2,0);
alphaii= j(m,2,0);
S=j(n,n,0);
yp=y-sum(y)/n;
yhat=j(m,1,0);
/*** Estimação do alpha pela regressão Binomial Negativa ***/
if gwr^="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;
        /* Newton Raphson */
        do while (abs(dpar)>0.001);
            aux1=aux1+1;
            if par<0 then do;
                par=0.00001;
            end;
            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)));
43          hess=choose(abs(hess)<1E-23,sign(hess)*1E-23,hess);
            hess=choose(hess=0,1E-23,hess);
45          par0=par;
            par=par0-inv(hess)*g;
47          if aux1>30 & par>1E5 then do;
                dpar= 0.0001;
49              aux2=aux2+1;
                if aux2=1 then par=2 ;
51              else if aux2=2 then par=1E5;
                else if aux2=3 then par=0.0001;
53          end;
            else dpar=par-par0;
55      end;
        a=1/par;
57      dev=0;
        ddev=1;
59      /* MQRI */
        do while (abs(ddev)>0.00001);
61          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));
63          b=inv((x#w)'*x)*(x#w)'*z;
            n=x*b;
65          u=exp(n);
            olddev=dev;
67          tt=y/u;
            tt=choose(tt=0,1E-10,tt);
69          dev=2*sum(y#log(tt)-(y+1/a)#log((1+a*y)/(1+a*u)));
            ddev=dev-olddev;
71      end;
        if aux2>4 then ddpar=1E-9;
73      else ddpar=par-parold;
    end;
75  %if &alphag= %then %let alphag=a;
    %else %let alphag=&alphag;
77  alphag=&alphag;
    bg=b;
79  parg=par;
end;
81  /*** Estimação do vetor de médias pelo MQRI modificado ***/
n=nrow(y);
83  do i=1 to m;
    /* Pesos da RGP */
85  d=j(1,3,0);
    do j=1 to n;
87      if abs(COORD[,1])<180 then do;
            dif=abs(POINTS[i,1]-COORD[j,1]);
89          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);
91          if round(ang,0.000000001)=1 then arco=0;
            else arco=arcos(ang);
93          d1=arco*6371 /*Earth's Radius = 6371 (aproximately)*/;
        end;
95      else d1=sqrt((POINTS[i,1]-COORD[j,1])**2+(POINTS[i,2]-COORD[j,2])**2);
```

```
        d[1]=i;
97      d[2]=j;
        d[3]=d1;
99      if j=1 then dist=d;
        else dist=dist//d;
101   end;
      w=j(n,1,0);
103   if method= "fixed" then do;
        do jj=1 to n;
105        w[jj]=exp(-0.5*(dist[jj,3]/h)**2);
        end;
107   end;
      wi=diag(w[,1]);
109   ym=sum(y)/nrow(y);
      uj=(y+ym)/2;
111   nj=log(uj);
      /* Alpha definido de forma global */
113   if gwr= "global" then alpha=alphag;
      if gwr= "poisson" then alpha=0 ;
115   dev=0;
      ddev=1;
117   cont=0;
      /* Cálculo do vetor de médias pelo MQRI modificado */
119   do while (abs(ddev)>0.000001);
        cont=cont+1;
121     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);
123     zj=nj+(y-uj)/(Ai#(1+alpha*uj));
        Ai=diag(Ai);
125     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;
127     nj=x*bi;
        nj=choose(nj>1E2,1E2,nj);
129     uj=exp(nj);
        olddev=dev;
131     uj=choose(uj<1E-150,1E-150,uj);
        tt=y/uj;
133     tt=choose(tt=0,1E-10,tt);
        if gwr= "poisson" then dev=2*sum(y#log(tt)-(y-uj));
135     else dev=2*sum(y#log(tt)-(y+1/alpha)#log((1+alpha*y)/(1+alpha*uj)));
        if cont>50 then ddev= 0.0000001;
137     else ddev=dev-olddev;
      end;
139   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);
141   Ai=diag(Ai2);
      if det(x'*wi*Ai*x)=0 then S[i,]=j(1,n,0);
143   else S[i,]= x[i,]*inv(x'*wi*Ai*x)*x'*wi*Ai;
      if gwr^="poisson" then do;
145     r=1/alpha;
        alphaii[i,1]=i;
147     alphaii[i,2]= alpha;
      end;
149   m1=(i-1)*ncol(x)+1;
      m2=m1+(ncol(x)-1);
151   bii[m1:m2,1]=i;
```

```
        bii[m1:m2,2]=bi;
153     yhat[i]=uj[i];
    end;
155 b=bii[,2];
    alphai=alphaii[,2];
157 id= bii[,1];
    ida=alphaii[,1];
159 yhat=choose(yhat<1E-150,1E-150,yhat);
    tt=y/yhat;
161 tt=choose(tt=0,1E-10,tt);
    if gwr= "poisson" then dev=2*sum(y#log(tt)-(y-yhat));
163 else dev=2*sum(y#log(tt)-(y+1/alphai)#log((1+alphai#y)/(1+alphai#yhat)));
    a2=y+1/alphai;
165 b2=1/alphai;
    c2=y+1;
167 algamma=j(n,1,0);
    blgamma=j(n,1,0);
169 clgamma=j(n,1,0);
    do i=1 to nrow(y);
171 algamma[i]=lgamma(a2[i]);
    blgamma[i]=lgamma(b2[i]);
173 clgamma[i]=lgamma(c2[i]);
    end;
175 if gwr^="poisson" then do;
        ll=sum(y#log(alphai#yhat)-(y+1/alphai)#log(1+alphai#yhat)+ algamma - blgamma -
            clgamma );
177     if gwr="global" & alphai^=1/parg then npar=trace(S);
        else npar=trace(S)+1;
179 end;
    else do;
181     ll=sum(-yhat+y#log(yhat)-clgamma);
        npar=trace(S);
183 end;
    AIC= 2*npar - 2*ll;
185 AICC= AIC +(2*npar*(npar+1))/(n-npar-1);
    print gwr method aicc dev npar;
187 create _beta_ var{id b}; append;
    create _alpha_ var{ida alphai}; append;
189 quit;
    %mend gwnbr;
```