Universidade de Brasília
Departamento de Estatística

gwzinbr: um pacote `R` para a Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

Jéssica Vasconcelos de Abreu
Juliana Magalhães Rosa

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Jéssica Vasconcelos de Abreu
Juliana Magalhães Rosa

gwzinbr: um pacote `R` para a Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

Orientador: Prof. Dr. Alan Ricardo da Silva

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Agradecimentos

Primeiramente, gostaríamos de agradecer ao nosso orientador Prof. Dr. Alan Ricardo da Silva pela confiança que depositou em nós, pela paciência ao nos ensinar e pela cobrança nos momentos necessários. Agradecemos também aos demais professores e funcionários do Departamento de Estatística (EST) na Universidade de Brasília (UnB) pelos aprendizados e apoio durante a graduação. Por fim, um agradecimento especial para as nossas famílias, para os nossos amigos e para os nossos colegas de curso pelo companheirismo e carinho, que foram essenciais para enfrentarmos os desafios do ensino superior e chegarmos aonde estamos hoje.

Resumo

Este trabalho teve como objetivo principal a implementação computacional do modelo de Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada (RBNIZGP) em linguagem `R`. Para isso, foi feita, primeiramente, uma análise teórica dos tópicos que servem como base para esse modelo: Modelos Lineares Generalizados (MLG) e Regressão Geograficamente Ponderada (RGP). Em seguida, foi analisada a macro GWZINBR em SAS, já utilizada anteriormente para ajustar o referido modelo a dados de COVID-19 na fase inicial da pandemia na Coreia do Sul. Com o entendimento da macro e de comportamentos específicos dos dois *softwares*, `R` e SAS, foi possível a criação do pacote `gwzinbr` no `R`, já disponível na plataforma CRAN. As funções do pacote são explicadas de forma detalhada no presente relatório. Com o intuito de testar o funcionamento desse novo pacote e comparar suas saídas com aquelas obtidas através do SAS, um estudo de caso foi feito com os mesmos dados da COVID-19 já citados e seus resultados foram visualizados e analisados com o apoio de tabelas, mapas e outras imagens.

Palavras-chaves: Regressão Geograficamente Ponderada; Binomial Negativa Inflcaionada de Zeros; estatística espacial; `R`.

Lista de Tabelas

2.1 Funções de ligação para o modelo binomial 13

2.2 Parâmetros das distribuições binomial, Poisson e binomial negativa 15

2.3 Funções de ligação para o modelo binomial negativo inflacionado de zeros . 15

3.1 Modificações na RBNIZGP que resultam em outros modelos de regressão . 26

4.1 Descrição das variáveis preditoras de dados da COVID-19 do *KCDC* . . . 30

5.1 Parâmetros da função `Golden()` 36

5.2 Valores de saída da função `Golden()` 37

5.3 Valores de saída da função `gwzinbr()` 38

5.4 Tempo de processamento da função `Golden` por modelo no SAS e no `R`. . . 38

5.5 Tempo de processamento da função `gwzinbr` por modelo no SAS e no `R`. . 39

6.1 Valores encontados pela função `Golden` para o parâmetro de suavização, por modelo, método e critério. 42

Lista de Figuras

3.1 Funções de ponderação espacial (a) Fixas (b) Adaptáveis 20

3.2 Efeito do ajuste na matriz de ponderação espacial 24

3.3 Relação entre os modelos binomial negativo inflacionado de zeros, Poisson inflacionado de zeros, binomial negativo e Poisson 26

4.1 Distribuição espacial das ocorrências de COVID-19 em diferentes fases de pandemia na Coreia do Sul, 2020 31

6.1 Saída da função `Golden` no SAS. 41

6.2 Saída da função `Golden` no `R`. 41

6.3 Medidas de ajustamento e medidas resumo das estimativas dos parâmetros do modelo RBNIZGP no SAS. 43

6.4 Medidas de ajustamento e medidas resumo das estimativas dos parâmetros do modelo RBNIZGP no `R`. 43

6.5 Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros do modelo RBNIZGP no SAS. 44

6.6 Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros do modelo RBNIZGP no `R`. 45

6.7 Medidas para as estimativas dos parâmetros inflacionados de zero para o modelo RBNIZGP no SAS. 45

6.8 Medidas para as estimativas dos parâmetros inflacionados de zero para o modelo RBNIZGP no `R`. 46

6.9 Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros inflacionados de zero do modelo RBNIZGP no SAS. 47

6.10 Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros inflacionados de zero do modelo RBNIZGP no R. 47

6.11 Estimativas dos parâmetros e medidas de ajustamento para a versão global do modelo RBNIZGP no SAS. 48

6.12 Estimativas dos parâmetros e medidas de ajustamento para a versão global do modelo RBNIZGP no SAS. 48

6.13 Distribuição espacial do número de casos de COVID-19 na Coreia do Sul durante fase inicial da pandemia 49

6.14 Mapa da Coreia do Sul 50

6.15 Distribuição espacial do nível de aglomeração (*Crowding*) na Coreia do Sul 50

6.16 Distribuição espacial das estimativas do parâmetro *Crowding* no modelo RBNIZGP na Coreia do Sul 51

6.17 Distribuição espacial das estimativas do parâmetro *Crowding* inflacionado de zeros no modelo RBNIZGP na Coreia do Sul 52

6.18 Distribuição espacial das estimativas do parâmetro de superdispersão no modelo RBNIZGP na Coreia do Sul 52

Sumário

1 Introdução **8**

1.1 Objetivos 9

1.1.1 Objetivo Geral 9

1.1.2 Objetivos Específicos. 9

2 Modelos Lineares Generalizados **10**

2.1 Introdução. 10

2.2 Família Exponencial 10

2.3 Estrutura Geral 11

2.4 Modelos Binomial e Binomial Negativo 12

2.5 Modelo Binomial Negativo Inflacionado de Zeros. 15

3 Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada **17**

3.1 Introdução. 17

3.2 Especificações da Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada. 17

4 Materiais e Métodos **29**

4.1 Introdução. 29

4.2 Materiais. 29

4.3 Métodos 31

5 Pacote `gwzinbr` no `R` **33**

5.1 Introdução 33

5.2 Aspectos Computacionais 33

5.3 Funções 36

5.4 Tempo de Execução 37

6 Resultados **40**

6.1 Introdução 40

6.2 Golden 40

6.3 gwzinbr 42

6.4 Visualização dos Resultados 49

7 Conclusões **54**

Referências 54

Capítulo 1

Introdução

Na área de saúde, é comum observar fenômenos de contagem com características espaciais os quais apresentam a particularidade de conter excesso de zeros. Isso significa que, apesar de algumas regiões apresentarem contagens altas, a maioria dos locais investigados não registrou nenhuma ocorrência do evento de interesse.

A análise de estruturas que incorporam esse tipo de informação pode contemplar, ao menos, dois importantes métodos. O primeiro é a Regressão Geograficamente Ponderada (RGP), proposta por Brunsdon et al. (1996), que é uma técnica de modelagem de dados que lida com a condição de não estacionariedade espacial, ou seja, com processos que são variantes no espaço. A RGP possibilita o mapeamento e a estimação de parâmetros para cada localização no espaço, ao invés de ter uma superfície de tendência a eles ajustada. Já algumas extensões de modelos lineares generalizados (Nelder e Wedderburn, 1972) permitem o ajuste de dados de contagem, em particular, aqueles que apresentam inflação de zeros.

É possível citar como exemplos as distribuições Poisson Inflacionada de Zeros (PIZ) (Lambert, 1992) e Binomial Negativa Inflacionada de Zeros (BNIZ) (Hall, 2000; Garay et al., 2011; Yau et al., 2003).Tendo em vista essas características e também a espacialidade e a superdispersão dos dados, foi proposta a Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada (RBNIZGP) (Da Silva e De Sousa, 2023).

A partir de algoritmo desenvolvido no *software* SAS para essa regressão, foi elaborado um estudo de caso para analisar contágios de COVID-19 na Coreia do Sul (Weinstein et al., 2021). Todavia, considerando que o `R` é um dos *softwares* estatísticos mais utilizados atualmente e sendo este uma ferramenta de acesso livre, é importante que exista uma função da RBNIZGP nessa linguagem para que os pesquisadores interessados possam

fazer uso desta metodologia com maior facilidade.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral implementar o algoritmo da RBNIZGP na linguagem de programação `R`, com base na macro já desenvolvida para o modelo no *software* estatístico SAS.

1.1.2 Objetivos Específicos

- Compreender a implementação da técnica RBNIZGP em SAS, a partir da macro já existente;
- Replicar o algoritmo em linguagem de programação `R`;
- Ajustar o modelo RBNIZGP nos dois *softwares*;
- Conferir compatibilidade dos resultados entre os *softwares* a partir de um mesmo conjunto de dados.

Capítulo 2

Modelos Lineares Generalizados

2.1 Introdução

Modelos Lineares Generalizados (MLG) constituem uma classe de modelos estatísticos que são extensões do modelo linear clássico, acomodando diferentes tipos de dados por meio de diferentes distribuições de probabilidades para a variável resposta. Introduzidos por Nelder e Wedderburn (1972), os MLG também permitem a análise da relação entre uma variável resposta Y e variáveis independentes $X_1, ...X_n$, mas considerando situações diversas como, por exemplo, distribuição não normal da resposta. São especialmente úteis na modelagem de dados de contagem, a partir de uma estrutura geral a qual pode lidar com características como a natureza discreta desse tipo de informação e eventuais casos de superdispersão.

Além disso, na formulação original dos autores dessa teoria, a distribuição da variável resposta Y pertence, regularmente, à família exponencial uniparamétrica e que, por sua vez, contempla modelos como o Gaussiano, Poisson, binomial, binomial negativo (este último, apenas no caso em que a precisão ou a dispersão é fixada) entre outros. Sendo assim, o objetivo deste Capítulo é explorar objetivamente os componentes de um modelo linear generalizado, introduzindo a família exponencial de distribuições de probabilidade e, especificamente, os modelos binomial e binomial negativo.

2.2 Família Exponencial

Dizemos que a distribuição de uma variável aleatória Y pertence à família exponencial multiparamétrica (e que, por sua vez, é uma generalização da família exponencial uniparamétrica) se sua função (densidade) de probabilidade puder ser expressa da seguinte

forma:

$$f(y;\boldsymbol{\theta}) = h(y)\exp\left\{\sum_{i=1}^{k}\eta_i(\boldsymbol{\theta})t_i(y) - b(\boldsymbol{\theta})\right\} \tag{2.2.1}$$

em que $\boldsymbol{\theta}$ é um vetor de k parâmetros; $h(y)$ e $t(y) = (t_1(y), ..., t_k(y))^\top$ são funções com valores reais y observados para Y e que não dependem de $\boldsymbol{\theta}$; e $b(\boldsymbol{\theta})$ e $\eta(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), ..., \eta_k(\boldsymbol{\theta}))^\top$ são funções com valores reais dos parâmetros, possivelmente definidos pelo vetor $\boldsymbol{\theta}$.

A partir dessa expressão, pode-se definir a forma canônica da família exponencial para um caso particular, quando $\eta(\theta)$ e $t(y)$ são funções do tipo identidade, retornando o mesmo valor usado como argumento. Logo:

$$f(y;\theta) = h(y)\exp\{\theta y - b(\theta)\} \tag{2.2.2}$$

Tal forma foi ampliada por Nelder e Wedderburn (1972) pela introdução do parâmetro $\phi > 0$, associado à dispersão da distribuição, de modo que:

$$f(y;\theta;\phi) = h(y)\exp\left\{\frac{\theta y - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{2.2.3}$$

onde θ é um parâmetro canônico, ϕ é um parâmetro de dispersão e $b(.)$ e $c(.)$ são funções conhecidas. Desse modo, algumas distribuições podem ser descritas conforme a Equação 2.2.3, a exemplo da Poisson, binomial e binomial negativa.

2.3 Estrutura Geral

Um modelo linear generalizado consiste em três componentes:

- Componente aleatório: consiste em um conjunto de variáveis aleatórias independentes $Y_1, ..., Y_n$ pertencentes à família exponencial uniparamétrica.
- Componente sistemático: uma função linear que pode ser definida como

$$\eta_i = \boldsymbol{x_i}^\top\boldsymbol{\beta} \tag{2.3.1}$$

  onde o vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, ...\beta_p)^\top$, cujos valores usualmente são desconhecidos e devem ser estimados a partir dos dados, e a i-ésima observação das p variáveis explicativas $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})^\top$, com $i = 1, ..., n$, compõem o preditor linear η.

- Função de ligação: função que vincula a média $E(Y_i) = \mu_i$ ao preditor linear, de forma que:

$$\eta_i = g(\mu_i) = \boldsymbol{x_i}^\top\boldsymbol{\beta} \tag{2.3.2}$$

em que g(.) é uma função monótona e diferenciável. Se a função de ligação for selecionada de modo que $g(\mu_i) = \theta_i = \eta_i$, o preditor linear modelará o parâmetro canônico θ_i de maneira direta, e isso é conhecido como função de ligação canônica.

2.4 Modelos Binomial e Binomial Negativo

A distribuição binomial é representativa de dados de proporções para eventos binários. Se $Y \sim B(m, \pi)$, então Y é o número de sucessos em m ensaios de Bernoulli independentes e π é a probabilidade de sucesso em cada ensaio (Dobson e Barnett, 2008).

A função de probabilidade na sua forma da família exponencial é a seguinte (Dobson e Barnett, 2008):

$$f(y; m, \pi) = \exp\left(y \log\left(\frac{\pi}{1-\pi}\right) + m \log(1-\pi) + \log\binom{m}{y}\right) \tag{2.4.1}$$

com $0 < \pi < 1$ e $m \in \mathbb{N}^*$. Sob este modelo, Y possui média $E(Y) = m\pi$ e variância $Var(Y) = m\pi(1-\pi)$. Assim, tem-se que $Var(Y) < E(Y)$. Algumas possíveis funções de ligação para o modelo binomial estão apresentadas na Tabela 2.1. A função de ligação que será utilizada neste trabalho é a função de ligação canônica Logit.

Nesse caso, o logaritmo da função de verossimilhança sob este modelo é dado por:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ y_i(\boldsymbol{x_i}^\top \boldsymbol{\beta}) - \log\left(1 + e^{\boldsymbol{x_i}^\top \boldsymbol{\beta}}\right) + \log\binom{m}{y_i} \right\} \tag{2.4.2}$$

em que n é o tamanho da amostra.

A qualidade do ajuste pode ser conferida pelo desvio (ou *deviance*), que mede a discrepância entre o logaritmo da função de verossimilhança do modelo completo (para o qual tem-se tantos parâmetros quanto observações) e do modelo ajustado. Quanto menor o resultado da função desvio, melhor o ajuste do modelo aos dados. No caso do modelo binomial, o desvio é dado por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i) \log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right) \right] \tag{2.4.3}$$

onde $\hat{\mu}_i$ é o estimador de máxima verossimilhança de μ_i.

Assim como a distribuição binomial, a binomial negativa é também adequada para dados de contagem. Considere um modelo de regressão com resposta binomial negativa e $Y_1, Y_2, ..., Y_n$ variáveis aleatórias independentes tais que $Y_i \sim \text{BN}(\mu_i, \phi)$. A função de

Tabela 2.1: Funções de ligação para o modelo binomial

Logit	**Probit**	**Log-log**
$\log\left(\frac{\pi}{1-\pi}\right) = \frac{e^{\boldsymbol{x_i^T\beta}}}{1+e^{\boldsymbol{x_i^T\beta}}}$	$\Phi^{-1}(\pi_i) = \Phi(\boldsymbol{x_i^T\beta})$	$\log\left(-\log\left(1-\pi_i\right)\right) = 1 - e^{-e^{\boldsymbol{x_i\beta}}}$

Fonte: Sousa (2022)

probabilidade de Y_i é dada por (Paula, 2004):

$$f(y_i;\mu_i,k) = \frac{\Gamma(k+y_i)}{\Gamma(y_i+1)\Gamma(k)}\left(\frac{\mu_i}{\mu_i+k}\right)^{y_i}\left(\frac{k}{\mu_i+k}\right)^{k} \tag{2.4.4}$$

onde $y_i = 0, 1, 2, \ldots$; μ e $k > 0$ são, respectivamente, a média e o parâmetro de precisão.

As distribuições Poisson e binomial negativa são utilizadas para representar dados de contagem, porém, a última também incorpora a sobredispersão por meio do parâmetro $\alpha = 1/k$. A média de Y_i é dada por $E(Y_i) = \mu_i$ e sua variância por $Var(Y_i) = \mu_i + \mu_i^2/\phi$. Conclui-se, portanto, que $Var(Y_i) > E(Y_i)$, caracterizando a sobredispersão dessa distribuição de probabilidade (Paula, 2004).

É possível utilizar uma variável *offset* $\boldsymbol{t}$ que permite ajustar o modelo para diferenças nas unidades de observação, mas sem incluir um coeficiente estimado para essa variável, tal que:

$$\log\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k}\right) = \boldsymbol{X\beta} + \log\left(\boldsymbol{t}\right) \tag{2.4.5}$$

Considerando a função de ligação canônica, o logaritmo da função de verossimilhança é dado por:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{k},\boldsymbol{y}) = \sum_{i=1}^{n}\left[\log\left\{\frac{\Gamma(k+y_i)}{\Gamma(k)\Gamma(1+y_i)}\right\} + k\log(k) + y_i\log(e^{\boldsymbol{X_i\beta}}) - (y_i+k)\log(k+e^{\boldsymbol{X_i\beta}})\right] \tag{2.4.6}$$

onde $\boldsymbol{y} = (y_1, ..., y_n)^T$ é a amostra de valores observados para Y e $\Gamma(z) = (z-1)!$, para z inteiro positivo, é a função gama.

A função desvio sob o modelo binomial negativo fica expressa por:

$$D(\boldsymbol{y},\hat{\boldsymbol{\mu}},\boldsymbol{k}) = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu_i}}\right) - (y_i+k)\log\left(\frac{1+\frac{1}{k}y_i}{1+\frac{1}{k}\mu_i}\right)\right] \tag{2.4.7}$$

Vale mencionar que a distribuição Poisson é um caso limite da binomial negativa

quando $k \to \infty$ (ou $\alpha \to 0$). Ademais, ao contrário das regressões clássica e de Poisson, o modelo Binomial Negativo regularmente não faz uso da função de ligação canônica descrita em (2.4.5). Existem diferentes ligações que podem ser utilizadas, mas no modelo tradicional de regressão Binomial Negativo denominado NB-2, é utilizada a função de ligação logarítmica $g(\mu) = \theta = \log \mu$ (Hilbe, 2011). Como o modelo Binomial Negativo é geralmente aplicado quando o modelo Poisson não é adequado, o uso da mesma função de ligação da regressão de Poisson, cuja canônica é a logarítmica, permite uma comparação direta entre eles, facilitando a avaliação dos benefícios da modelagem NB-2.

Para que os erros padrão possam ser calculados com base na matriz de informação, o Método Escore de Fisher sofre alterações (Hilbe, 2011). Primeiramente, é definida a matriz diagonal A_0 com elementos dados por:

$$a_{i0} = \frac{1}{V(\boldsymbol{\mu})}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^2 + (\boldsymbol{y} - \boldsymbol{\mu})\frac{V(\boldsymbol{\mu})g''(\boldsymbol{\mu}) + V'(\boldsymbol{\mu})g'(\boldsymbol{\mu})}{V(\boldsymbol{\mu})^2 g'(\boldsymbol{\mu})^3} \tag{2.4.8}$$

sendo $V(\boldsymbol{\mu})$ a função de variância, $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\top$ e $i = 1, ..., n$.

Assim, a matriz $\boldsymbol{A}$ é representada da seguinte forma:

$$\boldsymbol{A} = \sum_{i=1}^{n} \frac{k e^{\boldsymbol{x_i\beta}}}{k + e^{\boldsymbol{x_i\beta}}}\left(\frac{y_i - e^{\boldsymbol{x_i\beta}}}{k + e^{\boldsymbol{x_i\beta}}} + 1\right) \tag{2.4.9}$$

Já a segunda derivada do logaritmo da função de verossimilhança, a qual pode ser usada para o cálculo da variância de α, é:

$$H = \sum_{i=1}^{n}\left(\psi'(k + y_i) - \psi'(k) + \frac{1}{k} - \frac{2}{(k + \mu_i)} + \frac{(y_i + k)}{[(k + \mu_i)(k + \mu_i)]}\right) \tag{2.4.10}$$

onde $\psi(z) = \frac{\partial \log \Gamma(z)}{\partial z}$ é a função digama e $\psi'(z) = \frac{\partial \psi(z)}{\partial z} = \frac{\partial^2 \log \Gamma(z)}{\partial z^2}$ a função trigama.

Considerando a aplicação de algum algoritmo de otimização, como Newton-Raphson, para a estimação dos parâmetros do modelo binomial negativo, tem-se:

$$\hat{\alpha} = \frac{1}{\hat{k}} \tag{2.4.11}$$

$$Var(\hat{\alpha}) = -\frac{1}{H\hat{k}} \tag{2.4.12}$$

em que $\hat{\alpha}$ e $\hat{k}$ são os estimadores de α e k, respectivamente, e $Var(\hat{k}) = -\frac{1}{H}$.

Tabela 2.2: Parâmetros das distribuições binomial, Poisson e binomial negativa

Distribuição	ϕ	θ	$b(\theta)$	μ	$V(\mu)$	$c(y)$
Binomial	1	$\log\left(\frac{\mu}{1-\mu}\right)$	$m\log(1+e^{\theta})$	$\frac{me^{\theta}}{1+e^{\theta}}$	$\frac{\mu}{m}(m-\mu)$	$\log\binom{m}{y}$
Poisson	1	$\log(\mu)$	$\mu=\exp(\theta)$	$b'(\theta)=\exp(\theta)$	$b''(\theta)=exp(\theta)=\mu$	$-\log(y!)$
Binomial Negativa	1	$\log\left(\frac{\mu}{\mu+k}\right)$	$-\log\left(\frac{k}{\mu+k}\right)$	$b'(\theta)=\frac{\mu}{k}$	$b''(\theta)=\frac{\mu(\mu+k)}{k^2}$	$\log\left[\frac{\Gamma(k+y)}{\Gamma(y+1)\Gamma(k)}\right]$

2.5 Modelo Binomial Negativo Inflacionado de Zeros

O modelo binomial negativo inflacionado de zeros tem como base os modelos já apresentados, mas com a inclusão de um aspecto a mais: o excesso de zeros na distribuição. Sua função densidade é da forma (Yau et al., 2003):

$$f(y,p,\mu,k)=\begin{cases}p+(1-p)\left(\frac{k}{k+\mu}\right)^{k}, & y=0\\ (1-p)\frac{\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{y}, & y>0\end{cases} \tag{2.5.1}$$

onde p agora representa a probabilidade de zeros inflacionados, μ é a média da binomial negativa e k é o inverso do parâmetro de superdispersão α, sendo $k>0$.

A distribuição Binomial Negativa Inflacionada de Zeros (BNIZ) apresenta o estado zero para modelar contagens nulas e o estado binomial negativo para contagens não nulas (Garay et al., 2011). Além disso, assim como a distribuição Poisson é um caso limite da binomial negativa, a Poisson inflacionada de zeros (PIZ) é um caso limite da BNIZ. Quando $k\to\infty$, a binomial negativa inflacionada de zeros resulta em uma PIZ (Yau et al., 2003).

Sendo Y uma variável aleatória com distribuição BNIZ, sua média é dada por $E(Y)=(1-p)\mu$ e sua variância por $Var(Y)=(1-p)(1+\frac{\mu}{k}+p\mu)\mu$. Quanto à função de ligação, são usadas duas, cada uma associada a um parâmetro da distribuição, conforme a Tabela 2.3.

Tabela 2.3: Funções de ligação para o modelo binomial negativo inflacionado de zeros

Canônica	**Logit**
Parte não inflacionada de zeros	Parte inflacionada de zeros
$\log(\boldsymbol{\mu})=\boldsymbol{X\beta}$	$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right)=\boldsymbol{G\gamma}$

Tem-se que $\boldsymbol{X}$ é a matriz de covariáveis da distribuição binomial negativa de dimensões n x L e $\boldsymbol{G}$ a matriz com os valores fixados das covariáveis associadas à inflação

de zeros, de dimensões n x M. Já $\boldsymbol{\beta}$ é o vetor de parâmetros do primeiro modelo e $\boldsymbol{\gamma}$ do segundo.

Para estimar os parâmetros do modelo binomial negativo inflacionado de zeros, o algoritmo EM é utilizado, havendo a introdução de $\boldsymbol{z}$, um conjunto de variáveis indicadoras de inflação de zeros (Fumes, 2009; Garay et al., 2011).

Com isso, o logaritmo da função de verossimilhança é expresso por:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{y}, k) = \sum_{i=1}^{n} \Bigg\{ & z_i \boldsymbol{G_i}\boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{G_i}\boldsymbol{\gamma}}) + \\ & (1 - z_i) \log \left(\frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left[\frac{e^{\boldsymbol{X_i}\boldsymbol{\beta}}}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}} \right]^{y_i} \left[\frac{k}{k + e^{\boldsymbol{X_i}\boldsymbol{\beta}}} \right]^{k} \right) \Bigg\}\end{aligned} \tag{2.5.2}$$

onde $\boldsymbol{G_i}$ e $\boldsymbol{X_i}$ são as i-ésimas linhas de $\boldsymbol{G}$ e $\boldsymbol{X}$ respectivamente. Se $\boldsymbol{\gamma} = 0$, o modelo se resume a uma binomial negativa, reiterando o poder de generalização da distribuição BNIZ (Sousa, 2022).

A matriz de informação observada é dada por:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}, k) = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix} \tag{2.5.3}$$

E sua inversa é dada por:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}, k)^{-1} = \begin{pmatrix} Var(\hat{\boldsymbol{\beta}}) & Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) & Cov(\hat{\boldsymbol{\beta}}, \hat{k}) \\ Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) & Var(\hat{\boldsymbol{\gamma}}) & Cov(\hat{k}, \hat{\boldsymbol{\gamma}}) \\ Cov(\hat{\boldsymbol{\beta}}, \hat{k}) & Cov(\hat{k}, \hat{\boldsymbol{\gamma}}) & Var(\hat{k}) \end{pmatrix} \tag{2.5.4}$$

Mais detalhes sobre os modelos e as equações podem ser vistos em Sousa (2022).

Capítulo 3

Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

3.1 Introdução

As correlações entre variáveis podem ser, em alguns casos, fenômenos espacialmente não-estacionários, ou seja, podem variar em intensidade (e até em sentido) a depender da localização geográfica. Quando isso ocorre, é mais interessante utilizar regressão espacial para modelar os dados, de modo a levar em consideração essa heterogeneidade espacial.

Por conseguinte, a técnica da regressão geograficamente ponderada surgiu como uma aplicação de modelos locais, em que se tem parâmetros calculados para cada localização e a calibração é feita por meio de pesos que variam com a distância geográfica. Assim, este Capítulo tem por objetivo a apresentação das principais ideias da RGP e, em especial, da sua versão para a distribuição binomial negativa inflacionada de zeros.

3.2 Especificações da Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada

A técnica RGP se baseia, originalmente, na distribuição Gaussiana. A calibração é feita de forma local para cada ponto com base nas observações mais próximas, fazendo com que os coeficientes da regressão formem uma superfície contínua (Fotheringham et al.,

2002). Esses parâmetros são denominados β_{il}, para a i-ésima localização e l-ésima covariável.

Um dos conceitos principais da RGP é o da matriz de pesos espaciais $\boldsymbol{W}(u_i, v_i) = \boldsymbol{W}(i)$ com dimensões n x n para o i-ésimo local:

$$\mathbf{W}(i) = \begin{pmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{pmatrix} \tag{3.2.1}$$

Também denominada como matriz de proximidade espacial, $\boldsymbol{W}$ auxilia na representação da estrutura espacial nas áreas em estudo e é utilizada no cálculo de estatísticas de autocorrelação espacial. Existem algumas opções quanto ao formato da função de ponderação, que é a que determina os pesos w_{ij} da matriz $\boldsymbol{W}(i)$. Um exemplo que considera a distância d_{ij} entre os pontos i e j é:

$$w_{ij} = \begin{cases} 1, & d_{ij} < d \\ 0, & \text{caso contrário} \end{cases} \tag{3.2.2}$$

Na situação em questão, d é fixado de forma arbitrária. Valores grandes de d, que podem incluir todas as observações na estimação dos β_{il}, atribuindo-lhes, portanto, peso unitário, resultam essencialmente em estimar por Mínimos Quadrados Ordinários (MQO). Se d é excessivamente pequeno, isso pode resultar na utilização de apenas o próprio ponto na estimação. Além disso, na aplicação dessa função, ao se aproximar da distância d, ocorre uma transição abrupta, uma vez que não há decaimento da função para maiores distâncias.

Uma alternativa para evitar essa descontinuidade seria empregar uma função contínua exponencial quadrática do tipo:

$$w_{ij} = e^{\frac{-d_{ij}^2}{2b^2}} \tag{3.2.3}$$

onde b é o parâmetro de suavização (ou *bandwidth*), cuja escolha de valor é um passo crítico, de acordo com Dempster et al. (2009). Esse é um elemento da função de ponderação que determina a taxa de decaimento dos pesos à medida que a distância entre os pontos aumenta.

Um valor alto para b significa que a área de influência na calibração do ponto

é grande, ou seja, o decaimento dos pesos é suave. Já se o parâmetro de suavização for pequeno, a área de influência é menor, indicando um decaimento mais acentuado. Dessa forma, a função apresentada na Equação (3.2.3), conhecida como *kernel* gaussiano, assume valores decrescentes quanto maior for a distância d_{ij}, na forma da distribução normal. Para reduzir o trabalho computacional gerado por essa expressão, já que a função de ponderação associa um peso para cada ponto de interesse, indica-se a utilização de uma mistura das Equações (3.2.2) e (3.2.3):

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2, & d_{ij} < b \\ 0, & \text{caso contrário} \end{cases} \tag{3.2.4}$$

Há situações em que os dados não estão distribuídos de maneira uniforme na região ou se agrupam em áreas de tamanhos distintos. Nessas circunstâncias, é aconselhável que o parâmetro de suavização possa ser ajustado de acordo com a disposição dos dados observados.

Nesse sentido, existe a ideia de definir b como o número de vizinhos, ou seja, a quantidade de pontos próximos ao local i que se deseja incluir na calibração desse ponto. Nessa técnica, o parâmetro de suavização recebe o nome de adaptável, enquanto no caso anterior, é dito fixo. A Figura 3.1 ilustra a diferença desses dois casos, exemplificando com duas localizações.

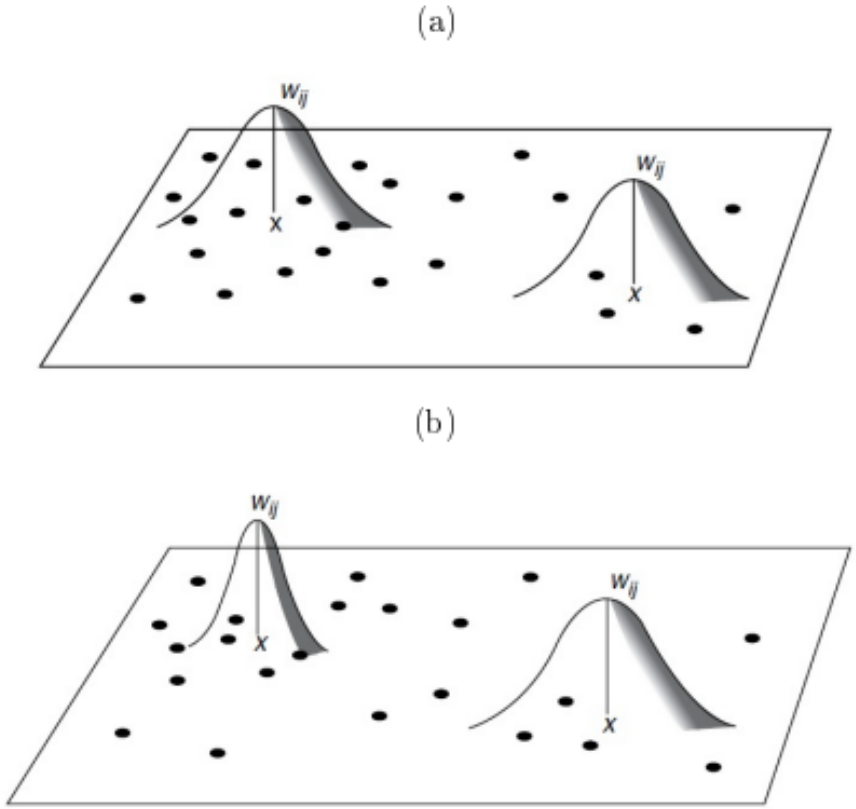


Figura 3.1: Funções de ponderação espacial (a) Fixas (b) Adaptáveis

Fonte: Fotheringham et al. (2002)

Os processos de escolha de função de ponderação, do valor para o parâmetro de suavização e da montagem da matriz de pesos são necessários para que se possa estimar os parâmetros da RBNIZGP: β_{il} são os parâmetros da parte não inflacionada de zeros do modelo, γ_{im} são os parâmetros da parte inflacionada e $k_i = \frac{1}{\alpha_i}$ é o parâmetro de precisão da distribuição binomial negativa, onde i se refere a cada local presente no conjunto de dados, l se refere a cada covariável da parte não inflacionada e m é referente a cada variável explicativa da parte inflacionada de zeros.

O modelo de RBNIZGP proposto em Da Silva e De Sousa (2023) é uma aplicação da RGP baseada na distribuição BNIZ, adequando-se a dados de contagem com super-dispersão e excesso de zeros e que variam espacialmente. Considerando $\boldsymbol{\beta_i}$ e $\boldsymbol{\gamma_i}$ os vetores dos parâmetros β_{il} e γ_{im} para o i-ésimo local, respectivamente, suas estimativas, juntamente com as estimativas do parâmetro k_i, podem ser calculadas a partir do algoritmo Expectativa-Maximização (EM) representado no Algoritmo 1.

Algoritmo 1: Fonte: Da Silva e De Sousa (2023)

Entrada: $\boldsymbol{\beta_i}$,$\boldsymbol{\gamma_i}$

Estimar $\boldsymbol{\beta_i}$ e k_i para $y > 0$, por meio da RBNGP. [1]

$\boldsymbol{\gamma_0} = \sum_{j=1}^{n} I_{(y_j=0)} - \sum_{j=1}^{n}(k_i/(\mu_j + k_i))^{k_i})/n$

Estimar $\boldsymbol{\gamma_i} = \log(\boldsymbol{\gamma_0}/(1 - \boldsymbol{\gamma_0}))$

$DiffD_i = 1$, $OldD_i = 0$

enquanto $(abs(DiffD_i) > 10^{-6})$ **faça**

  Passo E: Estimar a esperança condicional z_i

  se $y_i = 0$ **então**

    $$z_j = \left(1 + e^{-\boldsymbol{G_j\gamma_i}}\left[\frac{\hat{k_i}}{e^{\boldsymbol{X_j\hat{\beta}_i}}+\hat{k_i}}\right]^{\hat{k_i}}\right)^{-1}$$

  fim

  senão

    0

  fim

  Passo M para $\boldsymbol{\beta_i}$ e k_i: Estimação da RBNGP ponderado por $(1 - z_j)$, minimizando o desvio (D_1):

  $\{\boldsymbol{\eta} = \boldsymbol{X\beta_i} + offset$

  $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$

  $$\boldsymbol{M} = \frac{\Gamma(k_i+\boldsymbol{y})}{\Gamma(k_i)\Gamma(\boldsymbol{y}+1)}\left(\frac{\boldsymbol{\mu}}{\boldsymbol{\mu}+k_i}\right)^{\boldsymbol{y}}\left(\frac{k_i}{\boldsymbol{\mu}+k_i}\right)^{k_i}$$

  $D_1 = \sum_{j=1}^{n}[(1 - z_j)(\log(M_j))]$

  $\}$

  Passo M para $\boldsymbol{\gamma_i}$: Estimação RLGP [2] não ponderada, utilizando z_j como variável resposta, minimizando o desvio (D_2):

  $\{\ \boldsymbol{\eta} = \boldsymbol{G\gamma_i}$

  $$D_2 = \sum_{j=1}^{n}\left(z_j\eta_j - \sum_{j=1}^{n}\log(1 + \exp(\eta_j))\right)$$

  $\}$

  Maximização: $OldD_i = D_i$

  $D_i = D_1 + D_2$

  $DiffD_i = OldD_i - D_i$

fim

Fonte: Da Silva e De Sousa (2023)

Antes da inicialização de um ajuste desse modelo, o primeiro passo a ser tomado é a análise do comportamento dos dados, a fim de verificar se seguem, de fato, uma distribuição binomial negativa inflacionada de zeros. Para isso, deve-se observar a quantidade de zeros na variável resposta. Conforme proposto por Lambert (1992) para a Poisson inflacionada de zeros, pode-se utilizar a probabilidade de excesso de zeros média observada,

dada por $\hat{p_0}$, para isso.

$$\hat{p}_0 = \frac{\#\left(y_i = 0\right) - \sum_{i=1}^{n}\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^k}{n} \tag{3.2.5}$$

onde $\left(\frac{k}{e^{\boldsymbol{X_i\beta}}+k}\right)^k$ é a quantidade esperada de zeros em uma binomial negativa. Quando $k \to \infty$, esse valor converge para $\exp(-e^{\boldsymbol{X\beta}})$, ou seja, se torna equivalente ao da Poisson inflacionada de zeros (Lambert, 1992). Quando não houver presença de zeros ou quando a quantidade de zeros for menor do que aquela esperada pelo modelo binomial negativo, então $\gamma = 0$ e o modelo apropriado é a regressão binomial negativa geograficamente ponderada.

Seguindo a ideia da regressão global de que o número de observações deve ser maior do que o número de variáveis para que se possa fazer as estimações, na RBNIZGP deve-se ter $\sum_{j=1}^{n} w_{ij} > (L+M)$, onde L é o tamanho do vetor $\boldsymbol{\beta}$ e M é o tamanho do vetor $\boldsymbol{\gamma}$.

Em relação à escolha de valores iniciais para as estimativas locais k_i, uma possibilidade é o uso da estimativa k do modelo binomial negativo global. Já os erros padrão das estimativas de $\boldsymbol{\gamma_i}$, $\boldsymbol{\beta_i}$ e k_i são calculadas de forma conjunta. Assim, a matriz de covariância estimada é dada por:

$$\begin{aligned}&\boldsymbol{C}\left(u_i,v_i\right)\boldsymbol{A}^{-1}\left(u_i,v_i\right)\boldsymbol{C^T}\left(u_i,v_i\right) = \left(\boldsymbol{X}^T\boldsymbol{W}\left(u_i,v_i\right)\boldsymbol{A}\left(u_i,v_i\right)\boldsymbol{X}\right)^{-1}\\ &\left(\boldsymbol{X}^T\boldsymbol{W}\left(u_i,v_i\right)\boldsymbol{A}\left(u_i,v_i\right)\boldsymbol{W}\left(u_i,v_i\right)\boldsymbol{X}\right)\times\left(\boldsymbol{X}^T\boldsymbol{W}\left(u_i,v_i\right)\boldsymbol{A}\left(u_i,v_i\right)\boldsymbol{X}\right)^{-1}\end{aligned} \tag{3.2.6}$$

Em (3.2.6), a matriz de pesos $\boldsymbol{W}$ aparece duas vezes no termo que não está invertido. Isso impacta o processo de estimação de parâmetros que, pelo método Escore de Fisher, tem a matriz de covariância de $\hat{\boldsymbol{\beta}}$, considerando $n \to \infty$ dada por:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X^T}\hat{\boldsymbol{A}}\boldsymbol{X}^{-1} \tag{3.2.7}$$

em que $\hat{\boldsymbol{A}}$ é uma matriz de pesos $\boldsymbol{A}$ avaliada em $\hat{\boldsymbol{\beta}}$. Se $w_{ij} = 1, \forall i, j$, isto é, fazendo de um modelo local um modelo global, então (3.2.6) torna-se equivalente ao que está disposto em (3.2.7).

Dessa forma, para que a variância dos parâmetros $\boldsymbol{\beta_i}$ da RBNIZGP seja equivalente a RBNGP, a partir das segundas derivadas e quando $\boldsymbol{\gamma_i} = 0$, basta dizer:

$$I_{11} = -(\mathbf{X}^T\mathbf{d}_{bb}\mathbf{X})(\boldsymbol{X^T d_{bb}A_x^{-1}d_{bb}^T X})^{-1}(\boldsymbol{X^T d_{bb}X}) \tag{3.2.8}$$

em que

$$d_{bb} = \boldsymbol{W}(i)\boldsymbol{A_x}(i) \tag{3.2.9}$$

onde d_{bb} é derivada de segunda ordem. Dado que a matriz de covariância em (2.5.4) é a inversa da matriz de informação de Fisher em (2.5.3) correspondente ao modelo binomial negativo inflacionado de zeros, torna-se o termo igual a:

$$Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}) = (\boldsymbol{X^T d_{bb} X})^{-1}(\boldsymbol{X^T d_{bb} A^{-1} d_{bb}^T X})(\boldsymbol{X^T d_{bb} X})^{-1} \tag{3.2.10}$$

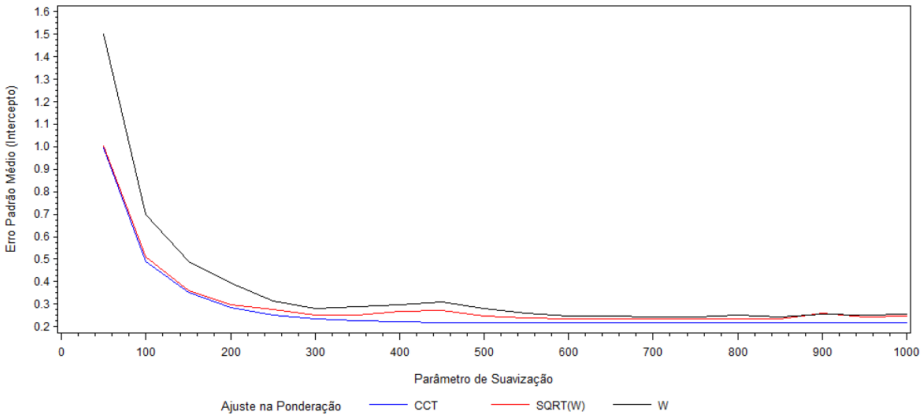
equivalente a (3.2.6). Essa transformação faz-se necessária porque como $\boldsymbol{W}(u_i, v_i)$ está ao quadrado no termo não invertido de (3.2.6), isso faz com que a variância de $\boldsymbol{\beta_i}$ seja tanto menor quanto for o parâmetro de suavização, quando comparado ao caso de haver $\boldsymbol{W}(i)$ apenas uma vez. Para ajustar uma RBNIZGP, em que $\boldsymbol{\gamma_i} > 0$, uma possível aproximação para essa expressão seria utilizar $\sqrt{\boldsymbol{W}(u_i, v_i)}$ no lugar de $\boldsymbol{W}(u_i, v_i)$:

$$I_{11} = -(\mathbf{X}^T \mathbf{d}_{sbb} \mathbf{X}) \tag{3.2.11}$$

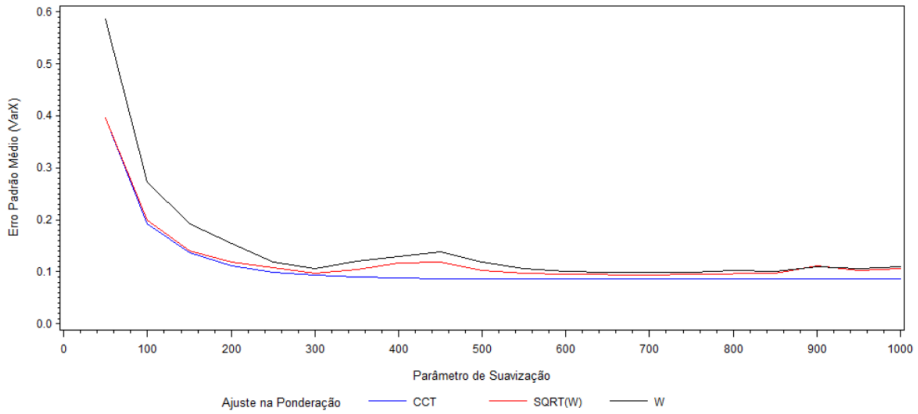
onde:

$$\mathbf{d}_{sbb} = \sqrt{\mathbf{W}(u_i, v_i)}\mathbf{A}(u_i, v_i) \tag{3.2.12}$$

A Figura 3.2 mostra o efeito do tipo da ponderação nos erros padrão de $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ em dados simulados, observado para o intercepto e para a covariável VarX, sendo CCT a forma da matriz de covariâncias estimada para o modelo RBNGP.



(a)



(b)

Figura 3.2: Efeito do ajuste na matriz de ponderação espacial

Fonte: Da Silva e De Sousa (2023)

Em relação à função desvio, expressa no Algoritmo 1, o valor $\frac{\Gamma(k+y)}{\Gamma(k)\Gamma(y+1)}$ pode ser omitido do cálculo, de forma a evitar problemas de convergência que podem resultar de um valor muito alto de k. Além disso, tendo estimadores locais e erros padrão, é possível realizar testes de significância para cada estimativa. Esses são pseudo testes t com estatística:

$$t_k(u_i, v_i) = \frac{\hat{\beta}_l(u_i, v_i)}{EP[\hat{\beta}_l(u_i, v_i)]} \tag{3.2.13}$$

com distribuição próxima da normal.

É importante mencionar que Da Silva e Fotheringham (2016) propuseram uma correção para o nível de significância α desses testes:

$$\alpha = \frac{p'}{p_e}\xi_m = \frac{\xi_m}{\frac{p_e}{p'}} \tag{3.2.14}$$

onde $p_e = 2tr(\boldsymbol{S}) - tr(\boldsymbol{S^T S})$ é o número efetivo de parâmetros independentes, $\boldsymbol{S}$ é a matriz de projeção (*hat matrix*), p' representa o número de parâmetros do modelo e ξ_m é o nível de significância desejado sem considerar a correlação espacial.

Sendo o AICc[3] uma medida comparativa entre modelos, seu cálculo foi desenvolvido por Hurvich e Tsai (1989):

$$AIC_c = -2\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{\gamma}) + 2r + \frac{2r(r+1)}{n-r-1} \tag{3.2.15}$$

em que $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{\gamma})$ representa o logaritmo da função de verossimilhança da RBNIZGP e $r = tr(\boldsymbol{R}) + tr(\boldsymbol{S}) + \frac{p_e}{p'}$ é o número efetivo de parâmetros do modelo. Nesse caso, $\boldsymbol{R}$ e $\boldsymbol{S}$ são as matrizes de projeção das partes não inflacionada e inflacionada, respectivamente.

Da Silva e Rodrigues (2014) discutiram o problema na estimação de r_2, o número de parâmetros efetivos de superdispersão no modelo e propuseram o uso de um modelo RBNGPg, ou seja, mantendo o parâmetro de superdispersão como global ($r_2 = 1$). Uma outra solução seria utilizar $r_2 = \frac{p_e}{p'} \geq 1$ (Da Silva e Fotheringham, 2016), resultando no número efetivo de parâmetros sendo representado por $r = p_e + \frac{p_e}{p'}$.

Retornando à determinação do parâmetro de suavização, a validação cruzada *CV* (*Cross Validation*) é comumente utilizada como critério de escolha, conforme sugerido por Cleveland (1979). Para isso, o CV é calculado como:

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2 \tag{3.2.16}$$

onde $\hat{y}_{\neq i}(b)$ é o valor ajustado para i excluindo-se da calibração o próprio i-ésimo ponto. Esse processo de minimização do CV é usualmente feito pelo algoritmo de otimização por seção áurea (*Golden Section Search*).

Uma importante característica da distribuição binomial negativa inflacionada de zeros é que esta tem a distribuição binomial negativa como caso particular e as distribuições Poisson e Poisson inflacionada de zeros como casos limites. Essa relação entre distribuções é retratada na Figura 3.3. O fluxograma parte do modelo mais geral para casos particulares a fim de que facilite a experiência do usuário da seleção do modelo.

Da mesma forma, o algoritmo da RBNIZGP permite a estimação dos modelos de Regressão Poisson Inflacionada de Zeros Geograficamente Ponderada (RPIZGP), regressão Poisson geograficamente ponderada e regressão binomial negativa geograficamente ponderada. Partindo dessa regressão mais geral, o algoritmo acomoda a estrutura

[3]critério AIC corrigido para pequenas amostras

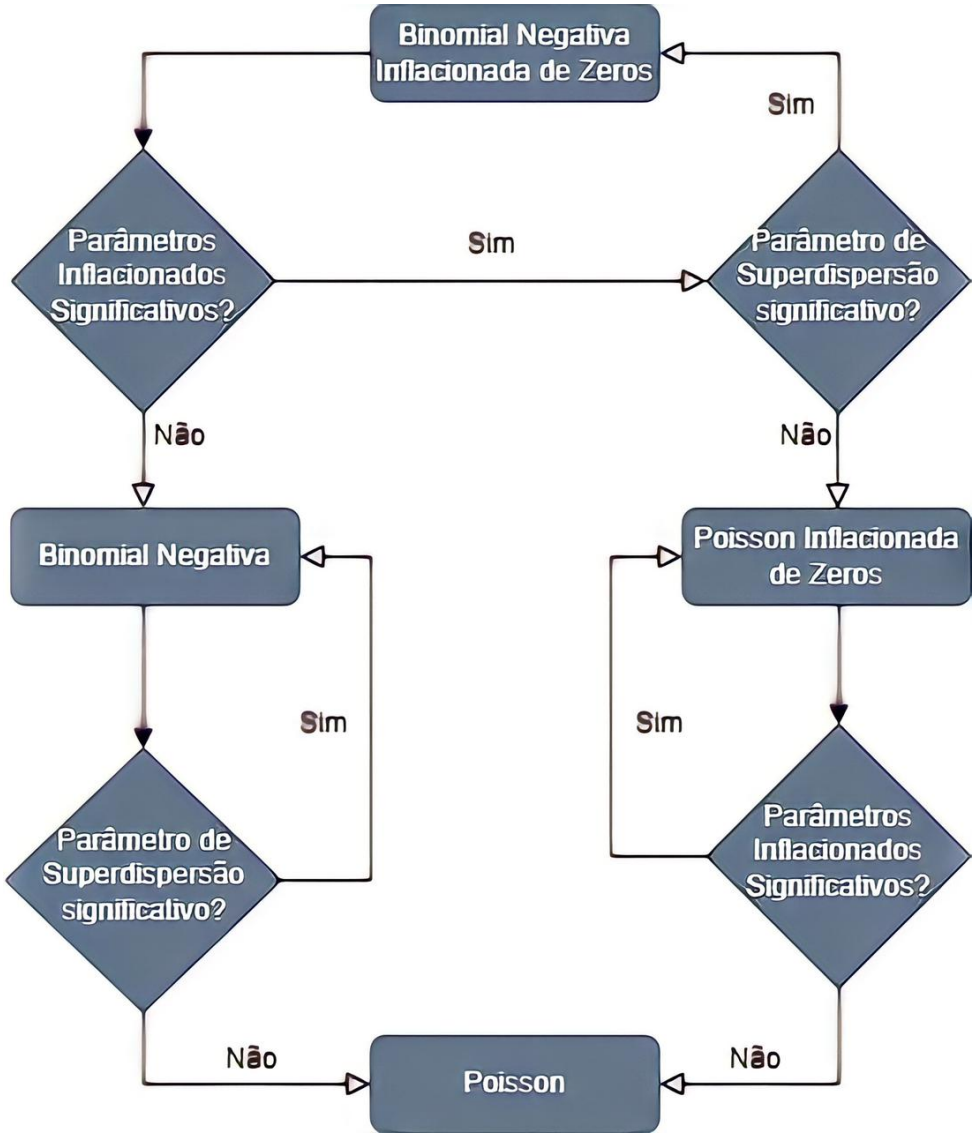


Figura 3.3: Relação entre os modelos binomial negativo inflacionado de zeros, Poisson inflacionado de zeros, binomial negativo e Poisson

Fonte: Da Silva e De Sousa (2023)

dos dados a partir de testes de significância para os parâmetros inflacionados de zero e de superdispersão.

Tabela 3.1: Modificações na RBNIZGP que resultam em outros modelos de regressão

Parâmetros da RBNIZGP	**Modelo gerado**
$\gamma = 0$	RBNGP
$\alpha = 0$	RPIZGP
$\alpha = 0$, $\gamma = 0$	RPGP[4]
$\alpha = 0$, $\beta = 0$	RLGP

Fonte: Da Silva e De Sousa (2023), com alterações

A função desvio para a binomial negativa inflacionada de zeros é dada por (Martin e Hall, 2016):

$$D = 2\sum_{i=1}^{n}[\mathcal{L}(\hat{z}, \mathbf{y}, \hat{k}; \mathbf{y}) - \mathcal{L}(\hat{\pi}, \hat{\mu}, \hat{k}; \mathbf{y})] \tag{3.2.17}$$

onde $\hat{\boldsymbol{\pi}} = \frac{e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}}}{1+e^{\boldsymbol{G}\hat{\boldsymbol{\gamma}}}}$, $\hat{\boldsymbol{\mu}} = e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}$ e $\boldsymbol{G}$ é a matriz de covariáveis que representam o excesso de zeros, vindas da distribuição binomial. Além disso:

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) = &-\sum_{i=1}^{n}\log\left(1+e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}}\right) + \sum_{y_i>0}\left[k\log\left(\frac{k}{k+\hat{\mu}_i}\right) + y_i\log\left(\frac{\hat{\mu}_i}{k+\hat{\mu}_i}\right)\right] + \\ &\sum_{y_i=0}\log\left[e^{\boldsymbol{G_i}\hat{\boldsymbol{\gamma}}} + \left(\frac{k}{k+\hat{\mu}_i}\right)^k\right] + \\ &\sum_{y_i>0}\left[-\log\Gamma\left(y_i+1\right) - \log\Gamma(k) + \log\Gamma\left(y_i+k\right)\right]\end{aligned} \tag{3.2.18}$$

e

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) = &-\sum_{i=1}^{n}\log\left(1+z_i\right) + \sum_{y_i>0}\left[k\log\left(\frac{k}{k+y_i}\right) + y_i\log\left(\frac{y_i}{k+y_i}\right)\right] + \\ &\sum_{y_i=0}\log\left[z_i + \left(\frac{k}{k+y_i}\right)^k\right] + \\ &\sum_{y_i>0}\left[-\log\Gamma\left(y_i+1\right) - \log\Gamma(k) + \log\Gamma\left(y_i+k\right)\right]\end{aligned} \tag{3.2.19}$$

A expressão da medida R^2 é dada por (Martin e Hall, 2016):

$$R^2_{ZINB} = 1 - \frac{\mathcal{L}(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y})}{\mathcal{L}(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - \mathcal{L}(0, \bar{y}, \hat{k}; \boldsymbol{y})} \tag{3.2.20}$$

com

$$\begin{aligned}\mathcal{L}(0, \bar{y}, \hat{k}; \boldsymbol{y}) = &-\sum_{i=1}^{n}\log(1+0) + \sum_{y_i>0}\left[k\log\left(\frac{k}{k+\bar{y}}\right) + y_i\log\left(\frac{\bar{y}}{k+\bar{y}}\right)\right] + \\ &\sum_{y_i=0}\log\left[0 + \left(\frac{k}{k+\bar{y}}\right)^k\right] + \sum_{y_i>0}\left[-\log\Gamma\left(y_i+1\right) - \log\Gamma(k) + \log\Gamma\left(y_i+k\right)\right]\end{aligned} \tag{3.2.21}$$

E sua versão ajustada é da forma (Martin e Hall, 2016):

$$R^2_{ZINB,adj} = 1 - \frac{\mathcal{L}(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - \mathcal{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{k}; \boldsymbol{y}) + L + M + 1,5}{\mathcal{L}(\hat{\boldsymbol{z}}, \boldsymbol{y}, \hat{k}; \boldsymbol{y}) - \mathcal{L}(0, \bar{y}, \hat{k}; \boldsymbol{y})} \tag{3.2.22}$$

[2]Regressão Poisson Geograficamente Ponderada

onde L e M são as dimensões de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente.

Capítulo 4

Materiais e Métodos

4.1 Introdução

Com o objetivo de realizar a implementação computacional da RBNIZGP no *software* `R` e comparar os resultados desse algoritmo com aqueles já obtidos previamente em SAS (Da Silva e De Sousa, 2023), serão utilizados os dados de Weinstein et al. (2021) sobre os casos de COVID-19 na Coreia do Sul a fim de verificar a qualidade da adaptação.

4.2 Materiais

A exemplo do estudo de caso realizado em Da Silva e De Sousa (2023), este trabalho utilizará dados de ocorrência de COVID-19 de 20 de janeiro de 2020 a 20 de março de 2020, divulgados pelo Centro para Controle e Prevenção de Doenças da Coreia do Sul (*Korea Centers for Disease Control and Prevention, KCDC*). O conjunto de dados é formado por 244 observações, e a variável resposta utilizada na modelagem é o número de casos da doença na fase inicial da pandemia (pré-quarentena). Já as variáveis explicativas estão descritas na Tabela 4.1.

Tabela 4.1: Descrição das variáveis preditoras de dados da COVID-19 do *KCDC*

Variável	**Descrição**
Morbidity	Comorbidade
high_sch_p	Proporção de pessoas com 2º grau completo
Healthcare_access	Acesso à saúde
diff_sd	Dificuldade de distanciamento social
Crowding	Aglomeração
Migration	Migração
Health_behavior	Comportamento de saúde, baseada em medidas de obesidade, alcoolismo e tabagismo

A história da pandemia de COVID-19 tem início em dezembro de 2019 quando, na cidade chinesa de Wuhan, província de Hubei, a Organização Mundial da Saúde (OMS) foi alertada sobre vários casos de pneumonia ocorrendo no local. Desde então, o número de ocorrências foi crescendo rapidamente até que, em janeiro de 2020, o surto do novo coronavírus tornou-se uma Emergência de Saúde Pública e, posteriormente, em março daquele mesmo ano, passou a ser caracterizada como pandemia. [1]

Distante pouco mais de dois mil quilômetros da China, a Coreia do Sul logo se tornou um dos países mais afetados pela propagação do vírus, especialmente em regiões populosas e próximas a grandes aeroportos, como a capital Seul e a cidade de Daegu, como pode ser visto na Figura 4.1. Entretanto, políticas adotadas pelo governo coreano para combater o avanço da doença mantiveram controlados os números de mortes. [2]

[1] Organização Pan-Americana da Saúde (OPAS). Histórico da Pandemia COVID-19. Disponível em: `https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19`. Acesso em: 23 de novembro de 2023

[2] BBC. Coronavírus: o que é o COVID-19 e como está se espalhando pelo mundo. Disponível em: `https://www.bbc.com/portuguese/internacional-51877262`. Acesso em: 23 de novembro de 2023

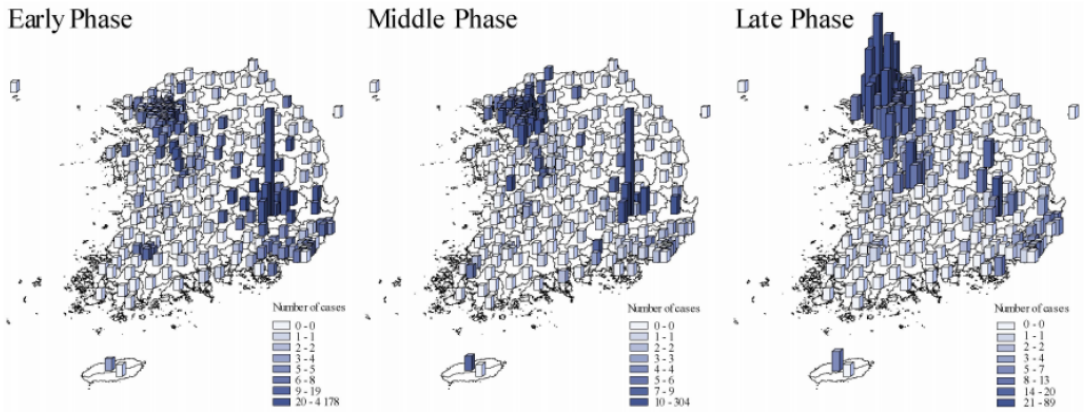


Figura 4.1: Distribuição espacial das ocorrências de COVID-19 em diferentes fases de pandemia na Coreia do Sul, 2020

Fonte: Da Silva e De Sousa (2023)

Diante do cenário apresentado, o modelo RBNIZGP será ajustado a esses dados, buscando obter, tanto através do SAS como do `R`, as melhores medidas de ajuste para os modelos gerados por esse algoritmo. Além disso, busca-se analisar os resultados obtidos por meio desse ajuste, com o apoio de visualizações como tabelas e mapas.

4.3 Métodos

A aplicação do modelo RBNIZGP consiste em três passos principais: escolha do valor para o parâmetro de suavização b, estimação dos parâmetros $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ e k (ou α) do modelo, e cálculo dos erros padrão dessas estimativas.

Primeiramente, encontra-se o valor de b que minimize algum critério como o CV ou o AIC. Isso é feito, em geral, a partir da *Golden Section Search* (GSS), que é uma técnica de otimização de funções.

A ideia da GSS é partir de um intervalo (h_0, h_3) inicial que contenha o mínimo da função. Depois encontram-se dois pontos h_1 e h_2 que estão a uma distância $d = GR(h_3 - h_0)$ de h_0 e h_3, onde $GR = \frac{\sqrt{5}-1}{2} \approx 0,618$ é um valor conhecido como *Golden Ratio*. Em seguida, são comparados os valores da função nesses dois pontos a fim de escolher (h_0, h_2) ou (h_1, h_3) como o novo intervalo. Esse processo se repete até que o intervalo encontrado seja pequeno o suficiente, daí toma-se seu ponto médio como sendo o mínimo procurado.

A estimação dos parâmetros do modelo RBNIZGP é feita conforme descrito no Algoritmo 1. E, por fim, os erros padrão são achados a partir da raiz quadrada das

variâncias dos estimadores, as quais são calculadas dentro da matriz de covariâncias, dada pela Equação (3.2.6).

Capítulo 5

Pacote `gwzinbr` no `R`

5.1 Introdução

Os pacotes do `R` são conjuntos de funções acerca de um determinado tema ou com um próposito em comum. Quaisquer dessas bibliotecas criadas em linguagem `R`, para serem considerados pacotes oficiais, devem ser submetidas e aceitas na plataforma CRAN (*The Comprehensive R Archive Network*). Sendo assim, o código criado deve seguir determinados padrões e ter um nível de qualidade mínimo para ser aceito e publicado no CRAN.

O pacote `gwzinbr` foi criado para a execução do modelo RBNIZGP e recebeu seu nome por causa do nome do modelo em inglês *Geographically Weighted Zero Inflated Negative Binomial Regression* (GWZINBR). Esse pacote pode ser instalado no `RStudio` e utilizado em qualquer lugar do mundo. Sua documentação pode ser acessada tanto no *help* do `RStudio` como no CRAN.

Este Capítulo tem por objetivo descrever o processo de criação do código para o pacote, incluindo dificuldades enfrentadas e diferenças entre o SAS e o `R`, além de apresentar as funções que compõem o pacote, detalhando seus parâmetros de entrada e as saídas que podem ser acessadas.

5.2 Aspectos Computacionais

O pacote elaborado neste trabalho teve como base a macro GWZINBR desenvolvida em SAS, cuja fundamentação teórica está apresentada no Capítulo 3. O código original, que serviu de inspiração para a implementação em `R`, realiza um conjunto de instruções definidas uma vez e invocadas para a execução do programa, encapsulado em

macros.

Em resumo, as etapas de ambos os algoritmos se iniciam na construção da função/ macro `Golden`. Primeiramente, é definida a rotina para o cálculo das estimativas dos parâmetros do modelo global, seguida da definição dos parâmetros da GSS e, por fim, a impressão do valor ótimo para o parâmetro de suavização. Ao longo dessas etapas, são feitas verificações relacionadas a alguns dos argumentos passados para a função, como a distribuição de probabilidade dos dados, o tipo do parâmetro de suavização e o critério para a escolha do seu valor ideal. Mais detalhes podem ser encontrados na documentação do pacote e no seu código fonte.

Em seguida, é construída a função `gwzinbr`, que opera com parâmetros muito semelhantes aos requeridos na `Golden`, conforme a Seção 5.3. Seu algoritmo também se inicia com a leitura dos dados e o cômputo das estimativas globais. O procedimento segue com o cálculo das distâncias; a seleção do modelo; a obtenção das estimativas dos parâmetros e de suas variâncias; a realização de um teste de não-estacionariedade, quando necessário; a obtenção dos erros padrão e outras medidas para o modelo global; e a apresentação dos resultados.

É importante ressaltar que a implementação exata dessas funções em `R` encontrou algumas limitações. Isso porque existem algumas distinções no funcionamento dos dois *softwares*, das quais decorrem variações estruturais dos códigos e pequenas diferenças nos resultados apresentados no Capítulo 6. Um exemplo que não afeta os resultados mas que requer adaptações no código é a característica conhecida como *case sensitive* que está presente no `R`, mas não no SAS. Isso significa que o `R` faz distinção entre termos que estão em letra maiúscula ou em letra minúscula.

Outro exemplo é a forma de capturar e manipular as informações obtidas pelas chamadas das funções. Em SAS, a leitura dos dados ocorre de forma mais direta e, no caso da macro GWZINBR mencionada, conta com o suporte da `PROC IML` para ler as variáveis diretamente do conjunto de dados especificado. Já em `R`, o código captura a chamada da função, ajusta seus argumentos e os avalia para criar um quadro de dados do modelo a partir de uma fórmula fornecida. Disso, é extraída a variável resposta e criada uma matriz com as covariáveis especificadas na fórmula.

Também é essencial levar em consideração a diferença entre vetores e matrizes ao realizar operações de multiplicação no `R`. No SAS IML (*Interactive Matrix Language*), o mesmo comando é usado para a criação de vetores e matrizes, enquanto no `R` existe até mesmo uma distinção entre matriz coluna e vetor, apesar de ambas as estruturas serem

equivalentes na álgebra linear.

Um desafio enfrentado durante a construção do algoritmo em `R` diz respeito a diferenças no armazenamento de números decimais entre as duas linguagens. Em um primeiro momento, tais diferenças podem parecer irrelevantes, pois acontecem a partir de casas decimais distantes. Entretanto, como a execução do algoritmo envolve a realização de centenas de cálculos, divergências desse tipo tendem a se propagar, principalmente quando estão dentro de laços aninhados.

Além disso, algumas funções nativas do `R` e do SAS, apesar de serem consideradas equivalentes entre os *softwares*, possuem comportamentos distintos. Um exemplo, o qual gerou grande impacto na implementação da função `gwzinbr`, é a função `det`. Esse é um comando para cálculo de determinantes de matrizes e é existente tanto no SAS como no `R`. Em alguns casos de teste em que o determinante é um valor muito próximo de zero, essa função retornou valores distintos: resultou em zero no SAS e em valores menores do que 10^{-30} no `R` ou vice-versa.

Ademais, o SAS e o `R` diferem em suas linhas de corte para arredondamentos para zero. Ao computar 10^{-308}, o SAS considera o valor como um número inválido e o converte em zero para algumas operações, enquanto no `R` este continua apenas a ser um valor muito pequeno. O `R` considera como zero apenas valores a partir de 10^{-324}. Essa distinção acaba prejudicando algumas verificações que são feitas em condicionais ao longo do algoritmo trabalhado, em especial as que checam a inversibilidade de matrizes a partir do valor do seu determinante.

Como a parte inflacionada de zeros do algoritmo da RBNIZGP envolve laços aninhados, cálculos de determinantes, inversão de matrizes, armazenamento de números decimais e valores muito pequenos, além de cálculos extensos com matrizes e vetores, essa etapa do modelo foi bastante impactada durante a implementação em `R`. Apesar de que a grande maioria das estimativas dessa parte do modelo não apresentu problemas, alguns pontos específicos dos dados divergiram em relação ao SAS.

As características dos *softwares* que foram apontadas nesta Seção são intrínsecas de cada linguagem, sendo difícil contornar diferenças causadas por elas. Um exemplo é o caso das funções nativas diferirem: como o SAS e o `R` não disponibilizam o código fonte desses comandos, não é possível entender, de fato, onde as funções se distinguem e nem propor soluções para aproximar seus comportamentos. Porém, as diferenças encontradas nos resultados não comprometem a qualidade do modelo e nem as conclusões retiradas.

5.3 Funções

Foram criadas duas funções no pacote `gwzinbr`, sendo a primeira delas chamada `Golden`. Essa função executa o algoritmo de *Golden Section Search* (GSS) explicado na Seção 4.3. Seus parâmetros estão listados e descritos na Tabela 5.1.

Tabela 5.1: Parâmetros da função `Golden()`

Parâmetro	Descrição
`data`	nome do conjunto de dados
`formula`	fórmula do modelo de regressão
`xvarinf`	nome das variáveis que explicam o comportamento inflacionado de zeros, o valor padrão é `NULL`
`weight`	nome da variável que contém os pesos amostrais, o valor padrão é `NULL`
`lat`	nome da variável que contém as latitudes no conjunto de dados
`long`	nome da variável que contém as longitudes no conjunto de dados
`globalmin`	indica se o usuário quer garantir um mínimo global no processo de otimização, o valor padrão é `TRUE`
`method`	indica o método a ser usado para o cálculo do parâmetro de suavização (`adaptive_bsq`, `fixed_g`)
`model`	indica a distribuição de probabilidade a ser usada na regressão (`zinb`, `zip`, `negbin`, `poisson`), o valor padrão é `"zinb"`
`bandwidth`	indica o critério a ser usado para a escolha do parâmetro de suavização (`cv`, `aic`), o valor padrão é `"cv"`
`offset`	nome da variável que contém os valores de *offset*, o valor padrão é `NULL`
`force`	indica se o usuário quer que a função de probabilidade passada como argumento para `model` seja forçadamente usada. Caso contrário, o modelo se adapta aos dados. O valor padrão é `FALSE`
`maxg`	inteiro indicando o número máximo de iterações para a parte inflacionada de zeros, o valor padrão é `100`
`distancekm`	indica se o cálculo das distâncias deve ser feito em quilômetros, o valor padrão é `FALSE`

A partir dessa entrada de argumentos, a função executa o algoritmo de otimização para minimizar o critério indicado e encontrar o valor ideal para o parâmetro de suavização. A saída da função `Golden` consiste em uma lista com quatro valores, os quais são listados e explicados na Tabela 5.2.

A segunda função do pacote é chamada `gwzinbr` e é a que faz efetivamente o ajuste do modelo, retornando estimativas dos parâmetros, entre outras saídas. Seus parâmetros de entrada são praticamente os mesmos apresentados na Tabela 5.1, com

Tabela 5.2: Valores de saída da função `Golden()`

Elemento da saída	Descrição
`h_values`	valores iniciais testados para o parâmetro de suavização
`iterations`	valores testados para o parâmetro de suavização em cada iteração e respectivos valores de CV/ AIC calculados
`gss_results`	valor mínimo encontrado para CV/ AIC e respectivo valor ideal do parâmetro de suavização, para fácil acesso
`min_bandwidth`	valor ideal do parâmetro de suavização

exceção do `globalmin` e do `bandwidth`, que são parâmetros da função `Golden`, mas não da `gwzinbr`.

Além dos parâmetros já explicados, a função `gwzinbr` requer três argumentos a mais: o valor de `grid` é um conjunto de dados a ser usado para as coordenadas dos locais, com padrão `NULL`, o valor de `int_inf` indica se a parte inflacionada do modelo deve incluir um intercepto (seu padrão é `TRUE`), e o `h` é o valor do parâmetro de suavização (podendo-se usar o valor encontrado através da `Golden`).

Vale mencionar que, no pacote oficial `gwzinbr` publicado no CRAN, as saídas das duas funções não são exibidas automaticamente após a sua execução. Os únicos resultados que aparecem de forma automática são os *NOTES*, que consistem em mensagens para o usuário. Para acessar os demais resultados, é preciso salvar o retorno da função em um novo objeto do `R` e depois utilizar o operador $ para exibir cada elemento da saída. No entanto, as ilustrações do Capítulo 6 foram feitas utilizando a exibição automática para facilitar a exposição desses resultados no relatório.

5.4 Tempo de Execução

Ao longo deste trabalho, diversas combinações de argumentos foram testadas nas funções `Golden` e `gwzinbr` a fim de que o código em `R` fosse capaz de reproduzir resultados corretos para quaisquer situações. Os tempos de processamento em SAS e `R`, realizados na mesma máquina, para alguns desses testes (com `bandwidth = "aic"` e `method = "adaptive_bsq"`) podem ser conferidos nas Tabelas 5.4 e 5.5. O código para os demais testes pode ser encontrado no GitHub, juntamente com outros arquivos auxiliares que foram utilizados ao longo da implementação computacional desenvolvida.

Tabela 5.3: Valores de saída da função `gwzinbr()`

Elemento da saída	Descrição
`bandwidth`	valor do parâmetro de suavização
`measures`	estatísticas de ajustamento do modelo e outras medidas
`qntls_gwr_param_estimates`	quantis das estimativas dos parâmetros
`descript_stats_gwr_param_estimates`	estatísticas descritivas das estimativas dos parâmetros
`t_test_gwr_param_estimates`	resultados para os testes t de significância dos parâmetros
`qntls_gwr_se`	quantis dos erros padrão
`descript_stats_gwr_se`	estatísticas descritivas dos erros padrão
`qntls_gwr_zero_infl_param_estimates`	quantis das estimativas dos parâmetros inflacionados de zero
`descript_stats_gwr_zero_infl_param_estimates`	estatísticas descritivas das estimativas dos parâmetros inflacionados de zero
`t_test_gwr_zero_infl_param_estimates`	resultados para os testes t de significância dos parâmetros inflacionados de zero
`qntls_gwr_zero_infl_se`	quantis dos erros padrão inflacionados de zero
`descript_stats_gwr_zero_infl_se`	estatísticas descritivas dos erros padrão inflacionados de zero
`non_stationary_test`	resultados do teste de não-estacionariedade para as estimativas dos parâmetros
`non_stationary_test_zero_infl`	resultados do teste de não-estacionariedade para as estimativas dos parâmetros inflacionados de zero
`global_param_estimates`	estimativas dos parâmetros do modelo global
`analysis_max_like_zero_infl_param_estimated`	análise de máxima verossimilhança para as estimativas dos parâmetros inflacionados de zero
`analysis_max_like_gof_measures`	medidas de ajustamento para a as estimativas dos parâmetros análise de máxima verossimilhança
`variance_covariance_matrix`	matriz de variância-covariância
`residuals`	resíduos do modelo
`param_estimates_grid`	estimativas dos parâmetros do modelo de RGP usando os dados de `grid`
`alpha_estimates`	estimativas para o parâmetro `alpha` de superdispersão
`gwr_param_estimates`	estimativas para os parâmetros de RGP

Tabela 5.4: Tempo de processamento da função `Golden` por modelo no SAS e no `R`.

Distribuição	***Software***	
	SAS	**R**
Poisson	8 segundos	14 segundos
Binomial Negativa	19 segundos	38 segundos
Poisson Inflacionado de Zeros (PIZ)	3 minutos e 15 segundos	8 minutos e 42 segundos
Binomial Negativo Inflacionado de Zeros (BNIZ)	6 minutos e 46 segundos	21 minutos e 54 segundos

Tabela 5.5: Tempo de processamento da função `gwzinbr` por modelo no SAS e no `R`.

Distribuição	***Software***	
	SAS	**R**
Poisson	3 segundos	5 segundos
Binomial Negativa	3 segundos	5 segundos
Poisson Inflacionado de Zeros (PIZ)	38 segundos	2 minutos e 12 segundos
Binomial Negativa Inflacionado de Zeros (BNIZ)	39 segundos	1 minuto e 24 segundos

Em geral, o código em SAS é mais rápido do que o do `R`. Essa diferença se acentua para a função `Golden`, cujo custo computacional é maior. Para as distribuições mais simples, Poisson e binomial negativa, o tempo de execução é quase o dobro no `R`, para ambas as funções. Já para as outras distribuições, a `gwzinbr` leva praticamente o mesmo tempo para processar a PIZ e a BNIZ no SAS, enquanto no `R` a segunda é mais rápida. Por fim, a `Golden` leva mais do que o dobro de tempo para executar o modelo RBNIZGP comparado ao modelo RPIZGP em ambos os *softwares*, sendo que o tempo do SAS representa menos da metade do tempo do `R`.

Capítulo 6

Resultados

6.1 Introdução

Para ilustrar o uso das funções criadas no pacote `gwzinbr`, foram utilizados os dados de COVID-19 na Coréia do Sul, apresentados na Seção 4.2. A variável resposta é chamada de `n_covid1` e representa o número de casos da doença na primeira fase da pandemia, enquanto as variáveis explicativas são aquelas listadas na Tabela 4.1.

Como exemplo, utilizou-se a distribuição BNIZ (`model = "zinb"`), que é o foco principal do trabalho, com critério AIC (`bandwidth = "aic"`) e parâmetro de suavização adaptável (`method = "adaptive_bsq"`). A escolha das covariáveis foi feita com base nas suas significâncias e no ajuste dos modelos (Sousa, 2022).

A partir dos resultados obtidos para as estimativas do modelo executado, serão apresentadas também algumas visualizações em mapas, permitindo um melhor entendimento da distribuição espacial das variáveis, sua variabilidade e sua associação com a resposta.

6.2 Golden

A chamada da função `Golden` para o teste mencionado é da seguinte forma:

```
> Golden(data = korea_base_artigo,formula = n_covid1~Morbidity+
high_sch_p+Healthcare_access+diff_sd+Crowding+Migration+
Health_behavior, xvarinf = c("Healthcare_access", "Crowding"),
long = "x", lat = "y", offset = "ln_total", model = "zinb",
method = "adaptive_bsq", bandwidth = "aic", globalmin = FALSE,
distancekm = TRUE, force = TRUE)
```

As saídas no `R` e no SAS são apresentadas nas Figuras 6.1 e 6.2, respectivamente.

h0	h1	h2	h3
5	96.289876	152.71012	244

golden	xmin	npar
1470.5637	82	58.189721

	GMY	H1	CV1	H2	CV2
1	1	96.28987639	1501.7852222	152.71012361	1704.4195055
2	1	61.420246552	1545.1450237	96.289877058	1501.7852222
3	1	96.289876645	1501.7852222	117.84049352	1644.2733371
4	1	82.970863168	1470.5636814	96.2898769	1501.7852222
5	1	74.739260126	1508.8302439	82.970863326	1470.5636814
6	1	82.970863229	1470.5636814	88.058273798	1478.490366
7	1	79.826670636	1484.6565231	82.970863289	1470.5636814

Figura 6.1: Saída da função `Golden` no SAS.

```
Bandwidth: 82
$h_values
  h0       h1       h2  h3
1  5 96.28988 152.7101 244

$iterations
  GSS_count       h1     aic1       h2     aic2
1         1 96.28988 1501.616 152.71012 1689.979
2         1 61.42025 1544.541  96.28988 1501.616
3         1 96.28988 1501.616 117.84049 1637.559
4         1 82.97086 1476.690  96.28988 1501.616
5         1 74.73926 1501.553  82.97086 1476.690
6         1 82.97086 1476.690  88.05827 1478.694
7         1 79.82667 1484.349  82.97086 1476.690

$gss_results
      aic bandwidth
1 1476.69        82
```

Figura 6.2: Saída da função `Golden` no `R`.

Os resultados do `R` foram equivalentes aos obtidos usando a macro original do SAS. Foram necessárias sete iterações para se obter o valor ideal para o parâmetro de suavização. Como `globalmin` recebeu `FALSE` como argumento, o algoritmo da *Golden Section Search* é executado apenas uma vez (`GSS_count=1`), partindo do intervalo inicial $(h_0; h_3) = (5; 244)$. O valor ótimo encontrado para o *bandwidth* foi de 82, pois resulta no mínimo de 1476,69 para a estatística AIC. Em geral, os valores do AIC não são idênticos aos do SAS (resultado final de 1470,56), por questões de arredondamento nas iterações,

mas isso não compromete a convergência do algoritmo, nem o *bandwidth* encontrado.

Além do exemplo apresentado, foram executados diversos outros testes para verificar o funcionamento adequado do código no `R`, sempre comparando-se aos valores obtidos no SAS. Para a função `Golden`, alguns desses resultados estão resumidos na Tabela 6.1.

Tabela 6.1: Valores encontados pela função `Golden` para o parâmetro de suavização, por modelo, método e critério.

Modelo	Fixo		Adaptável	
	CV	AIC	CV	AIC
BNIZ	199,96	199,96	230	82
PIZ	733,70	36,98	230	56
Binomial Negativa	189,74	156,67	230	82
Poisson	733,70	47,21	48	79

Olhando apenas para os valores de *bandwidth* encontrados, não se pode concluir sobre o melhor modelo a ser utilizado. No entanto, a partir desses diferentes resultados, pode-se testar a execução da `gwzinbr` e tirar a conclusão a partir das medidas de ajustamento.

Em geral, o *bandwidth* adaptável é mais versátil do que o fixo, pois funciona bem mesmo quando há diferenças na densidade dos dados ao longo das localizações, conforme descrito na Seção 3.2. Quanto ao critério de minimização, o AIC é interessante por servir como medida comparativa entre modelos. E, por fim, a distribuição BNIZ pode ser usada como ponto de partida com o argumento `force = FALSE`, de forma que o algoritmo acomode a distribuição aos dados.

Ao se comparar o ajuste dos diferentes modelos possíveis, os melhores resultados foram os das distribuições binomial negativa e binomial negativa inflacionada de zeros, com parâmetro de suavização igual a 82 (Da Silva e De Sousa, 2023), encontrado para o método adaptável com critério AIC, conforme a Tabela 6.1.

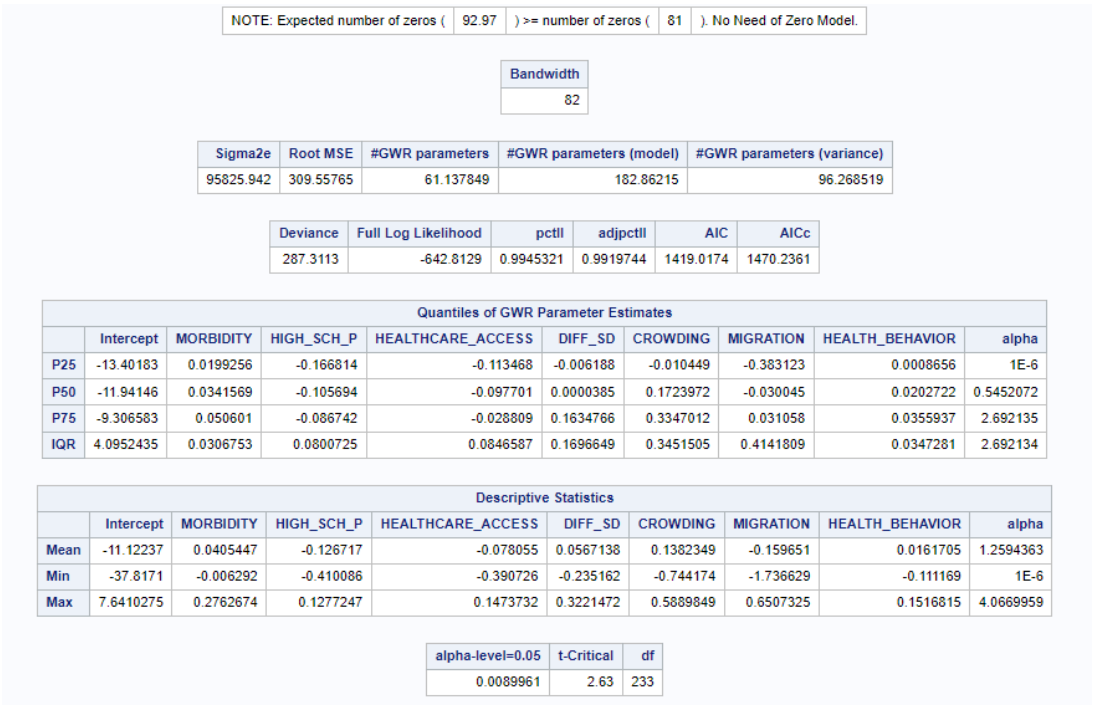
6.3 gwzinbr

Tendo selecionado o valor para o parâmetro de suavização, o próximo passo é o ajuste do modelo através da função `gwzinbr`:

```
> gwzinbr(data = korea_base_artigo,formula = n_covid1~Morbidity+
high_sch_p+ Healthcare_access+diff_sd+Crowding+Migration+
Health_behavior, xvarinf = c("Healthcare_access", "Crowding"),
long = "x", lat = "y", offset = "ln_total",
method = "adaptive_bsq",  model = "zinb", distancekm = TRUE,
```

```
h= 82, force = TRUE)
```

Algumas das saídas resultantes são apresentadas nas Figuras 6.3 até 6.12, alternando os *outputs* do SAS e do R.



NOTE: Expected number of zeros (92.97) >= number of zeros (81). No Need of Zero Model.

Bandwidth
82

Sigma2e	Root MSE	#GWR parameters	#GWR parameters (model)	#GWR parameters (variance)
95825.942	309.55765	61.137849	182.86215	96.268519

Deviance	Full Log Likelihood	pctll	adjpctll	AIC	AICc
287.3113	-642.8129	0.9945321	0.9919744	1419.0174	1470.2361

Quantiles of GWR Parameter Estimates

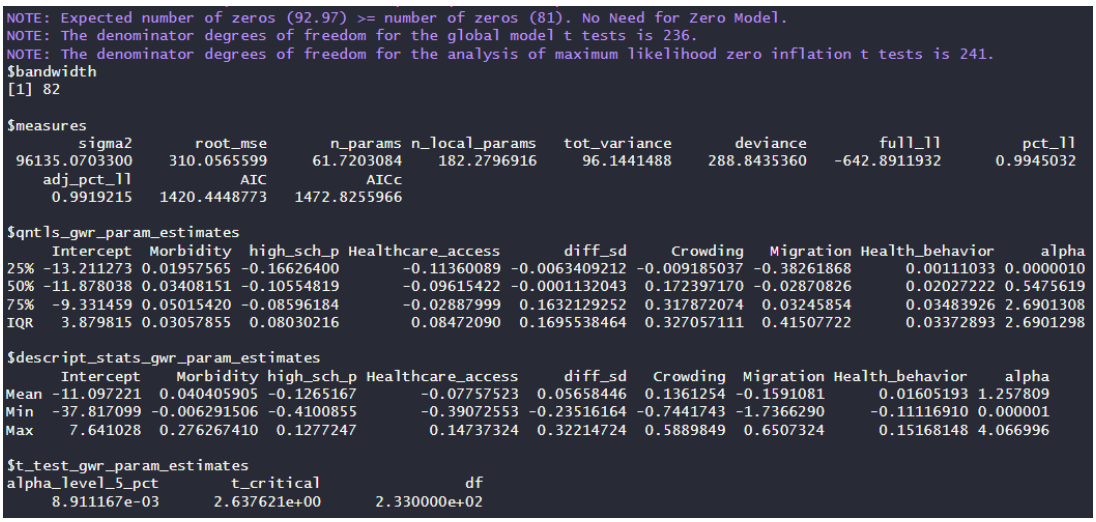
	Intercept	MORBIDITY	HIGH_SCH_P	HEALTHCARE_ACCESS	DIFF_SD	CROWDING	MIGRATION	HEALTH_BEHAVIOR	alpha
P25	-13.40183	0.0199256	-0.166814	-0.113468	-0.006188	-0.010449	-0.383123	0.0008656	1E-6
P50	-11.94146	0.0341569	-0.105694	-0.097701	0.0000385	0.1723972	-0.030045	0.0202722	0.5452072
P75	-9.306583	0.050601	-0.086742	-0.028809	0.1634766	0.3347012	0.031058	0.0355937	2.692135
IQR	4.0952435	0.0306753	0.0800725	0.0846587	0.1696649	0.3451505	0.4141809	0.0347281	2.692134

Descriptive Statistics

	Intercept	MORBIDITY	HIGH_SCH_P	HEALTHCARE_ACCESS	DIFF_SD	CROWDING	MIGRATION	HEALTH_BEHAVIOR	alpha
Mean	-11.12237	0.0405447	-0.126717	-0.078055	0.0567138	0.1382349	-0.159651	0.0161705	1.2594363
Min	-37.8171	-0.006292	-0.410086	-0.390726	-0.235162	-0.744174	-1.736629	-0.111169	1E-6
Max	7.6410275	0.2762674	0.1277247	0.1473732	0.3221472	0.5889849	0.6507325	0.1516815	4.0669959

alpha-level=0.05	t-Critical	df
0.0089961	2.63	233

Figura 6.3: Medidas de ajustamento e medidas resumo das estimativas dos parâmetros do modelo RBNIZGP no SAS.



```
NOTE: Expected number of zeros (92.97) >= number of zeros (81). No Need for Zero Model.
NOTE: The denominator degrees of freedom for the global model t tests is 236.
NOTE: The denominator degrees of freedom for the analysis of maximum likelihood zero inflation t tests is 241.
$bandwidth
[1] 82

$measures
        sigma2       root_mse       n_params n_local_params   tot_variance       deviance        full_ll         pct_ll
 96135.0703300    310.0565599     61.7203084    182.2796916     96.1441488    288.8435360   -642.8911932      0.9945032
    adj_pct_ll            AIC           AICc
     0.9919215   1420.4448773   1472.8255966

$qntls_gwr_param_estimates
     Intercept  Morbidity  high_sch_p Healthcare_access       diff_sd     Crowding   Migration Health_behavior      alpha
25% -13.211273 0.01957565 -0.16626400       -0.11360089 -0.0063409212 -0.009185037 -0.38261868      0.00111033 0.0000010
50% -11.878038 0.03408151 -0.10554819       -0.09615422 -0.0001132043  0.172397170 -0.02870826      0.02027222 0.5475619
75%  -9.331459 0.05015420 -0.08596184       -0.02887999  0.1632129252  0.317872074  0.03245854      0.03483926 2.6901308
IQR   3.879815 0.03057855  0.08030216        0.08472090  0.1695538464  0.327057111  0.41507722      0.03372893 2.6901298

$descript_stats_gwr_param_estimates
      Intercept    Morbidity high_sch_p Healthcare_access      diff_sd   Crowding  Migration Health_behavior     alpha
Mean -11.097221  0.040405905 -0.1265167       -0.07757523  0.05658446  0.1361254 -0.1591081      0.01605193 1.257809
Min  -37.817099 -0.006291506 -0.4100855       -0.39072553 -0.23516164 -0.7441743 -1.7366290     -0.11116910 0.000001
Max    7.641028  0.276267410  0.1277247        0.14737324  0.32214724  0.5889849  0.6507324      0.15168148 4.066996

$t_test_gwr_param_estimates
alpha_level_5_pct        t_critical                df
     8.911167e-03      2.637621e+00      2.330000e+02
```

Figura 6.4: Medidas de ajustamento e medidas resumo das estimativas dos parâmetros do modelo RBNIZGP no R.

As diferenças que aparecem nos resultados do R (Figura 6.4), em relação aos do SAS (Figura 6.3), estão, em sua maioria, presentes apenas nas casas decimais. Conforme explicado na Seção 5.2, isso ocorre por causa de diferenças intrínsecas dos dois *softwares*,

mas não compromete a interpretação dos resultados e as conclusões retiradas. O mesmo vale para os demais resultados apresentados nesta Seção.

A primeira mensagem (*NOTE*) presente nas Figuras 6.3 e 6.4 afirma que o número de zeros existentes na variável resposta é menor do que a quantidade de zeros esperada, indicando que não há necessidade de se usar um modelo inflacionado de zeros. De acordo com Da Silva e De Sousa (2023), a distribuição binomial negativa tem um bom ajuste e seria suficiente para representar esses dados. No entanto, optou-se por apresentar os resultados da BNIZ neste trabalho, pois com ela ganha-se informações adicionais acerca do elevado número de zeros em algumas localidades.

Em geral, o modelo apresentado teve um bom ajuste, com medidas de pseudo-R^2 próximas de 1 (`pct_ll` e `adj_pct_ll`). Em relação às estimativas dos parâmetros, seus 244 valores locais podem ser acessados na saída da função, mas apenas alguns quantis e algumas estatísticas descritivas são apresentadas nas Figuras 6.3 e 6.4, para se ter uma visão geral desses resultados.

Pode-se observar que para os parâmetros *high_sch_p*, *Healthcare_access* e *Migration*, os valores estimados são principalmente negativos, indicando que o aumento na proporção de pessoas com 2° grau completo, no acesso à saúde e na migração são aspectos que costumam estar associados a uma diminuição nos casos de COVID-19. Já as covariáveis referentes a comorbidades, dificuldade no distânciamento social, aglomeração e comportamento da saúde costumam ter uma relação positiva com a quantidade de casos.

Vale mencionar que para todas as covariáveis, as estimativas dos parâmetros têm valores mínimos negativos e valores máximos positivos, ressaltando a importância de se trabalhar com modelos locais para se observar as diferenças nas relações entre essas variáveis e a resposta em diferentes regiões. As significâncias locais dessas associações podem ser acessadas juntamente com as estimativas completas.

Quantiles of GWR Standard Errors

	Intercept	MORBIDITY	HIGH_SCH_P	HEALTHCARE_ACCESS	DIFF_SD	CROWDING	MIGRATION	HEALTH_BEHAVIOR	alpha
P25	1.9260429	0.0166207	0.0334753	0.0421432	0.0194396	0.1503689	0.0794469	0.0238547	0.0000708
P50	2.563735	0.018037	0.0428585	0.0509359	0.0392808	0.1988465	0.1630393	0.0273525	0.1306174
P75	3.8164543	0.0202785	0.0703887	0.0788653	0.0920356	0.2444133	0.3581445	0.0414738	0.6147293
IQR	1.8904114	0.0036578	0.0369135	0.036722	0.072596	0.0940444	0.2786976	0.0176191	0.6146585

Descriptive Statistics of Standard Errors

	Intercept	MORBIDITY	HIGH_SCH_P	HEALTHCARE_ACCESS	DIFF_SD	CROWDING	MIGRATION	HEALTH_BEHAVIOR	alpha
Mean	2.8557101	0.0189092	0.0545982	0.0571459	0.05493	0.1949737	0.2142526	0.032645	0.3372178
Min	0.545767	0.004699	0.0087553	0.0037658	0.0069349	0.0230722	0.0333633	0.0069548	7.8096E-7
Max	9.0820737	0.0813235	0.1441028	0.1736015	0.1649103	0.3961816	0.6613982	0.1102241	5.5830578

Figura 6.5: Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros do modelo RBNIZGP no SAS.

```
$qntls_gwr_se
    Intercept   Morbidity high_sch_p Healthcare_access    diff_sd
25%  1.830891 0.016250807 0.03340113        0.03873080 0.01932156
50%  2.426821 0.017974318 0.04281006        0.04984204 0.03900117
75%  3.671036 0.019937748 0.06827175        0.07620065 0.09183948
IQR  1.840145 0.003686942 0.03487062        0.03746985 0.07251792
     Crowding Migration Health_behavior        alpha
25% 0.1358059 0.0792293      0.02386183 6.945393e-05
50% 0.1797057 0.1630393      0.02822986 1.274790e-01
75% 0.2412560 0.3549930      0.04092498 5.276189e-01
IQR 0.1054501 0.2757637      0.01706315 5.275495e-01

$descript_stats_gwr_se
     Intercept   Morbidity  high_sch_p Healthcare_access     diff_sd
Mean 2.6728269 0.018639896 0.052621958        0.05475769 0.052963161
Min  0.2279687 0.003622368 0.008755345        0.00376583 0.006934877
Max  9.0820737 0.081323517 0.144102811        0.17360148 0.164910325
      Crowding  Migration Health_behavior        alpha
Mean 0.1804235 0.20611966     0.032025605 2.750471e-01
Min  0.0230722 0.03336335     0.006954777 7.809611e-07
Max  0.3240767 0.66139824     0.111123880 5.583057e+00
```

Figura 6.6: Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros do modelo RBNIZGP no `R`.

Observando as médias dos erros padrão nas Figuras 6.5 e 6.6, é possível perceber que os valores são muito similares entre os dois *softwares*, mas que o `R` resultou em valores um pouco menores. Como exemplos, tem-se uma média de 2,86 para os erros padrão do intercepto no SAS e de 2,67 no `R`; para a variável *Crowding* esse valor foi de 0,19 no SAS e 0,18 no R; e para o parâmetro de superdispersão `alpha` os valores médios obtidos foram de 0,34 e 0,28 para o SAS e o `R`, respectivamente.

Quantiles of GWR Zero Inflation Parameter Estimates

	Intercept	HEALTHCARE_ACCESS	CROWDING
P25	-33.2412	0.1114549	0.3250423
P50	-19.46976	0.1477504	0.7348632
P75	-15.82026	0.3949313	1.1811416
IQR	17.42094	0.2834763	0.8560993

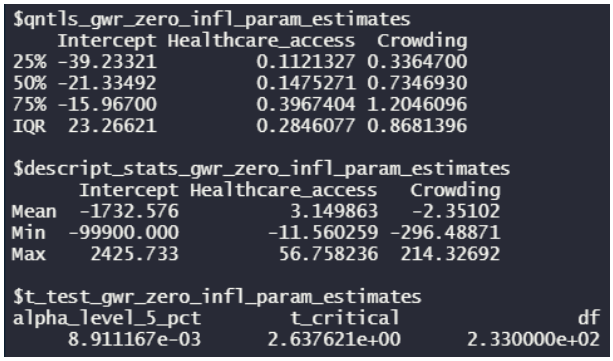
Descriptive Statistics

	Intercept	HEALTHCARE_ACCESS	CROWDING
Mean	-502.9967	3.1317015	-2.29754
Min	-99900	-11.56026	-296.4887
Max	2425.7326	56.74222	214.26621

alpha-level=0.05	t-Critical	df
0.0089961	2.63	233

Figura 6.7: Medidas para as estimativas dos parâmetros inflacionados de zero para o modelo RBNIZGP no SAS.

Uma das partes mais computacionalmente complexas do algoritmo é a do cálculo das estimativas dos parâmetros inflacionados de zeros. Por esse motivo, esses valores



```
$qntls_gwr_zero_infl_param_estimates
    Intercept Healthcare_access  Crowding
25% -39.23321         0.1121327 0.3364700
50% -21.33492         0.1475271 0.7346930
75% -15.96700         0.3967404 1.2046096
IQR  23.26621         0.2846077 0.8681396

$descript_stats_gwr_zero_infl_param_estimates
      Intercept Healthcare_access   Crowding
Mean  -1732.576          3.149863   -2.35102
Min  -99900.000        -11.560259 -296.48871
Max    2425.733         56.758236  214.32692

$t_test_gwr_zero_infl_param_estimates
alpha_level_5_pct        t_critical                df
     8.911167e-03      2.637621e+00      2.330000e+02
```

Figura 6.8: Medidas para as estimativas dos parâmetros inflacionados de zero para o modelo RBNIZGP no R.

estimados e os respectivos erros padrão sofreram algumas diferenças em seus resultados do R quando comparados ao SAS. Além disso, como essas medidas são apresentadas de forma resumida por meio de quantis e estatísticas descritivas, é importante comentar que as funções de cálculo de quantis são estruturalmente diferentes no SAS e no R e costumam resultar em valores diferentes mesmo para cálculos mais diretos e em dados reduzidos.

Tomando o intercepto como exemplo, os quantis das estimativas foram bastante próximos nos dois programas, sendo a maior diferença entre os valores do primeiro quartil, que difere em 6 unidades, conforme as Figuras 6.7 e 6.8. Os valores mínimo e máximo coincidem nas duas saídas, enquanto a média apresenta uma grande diferença (-503,00 no SAS e -1732,58 no R) devido a alguns valores extremos que aparecem como consequência das características dos *softwares* explicadas na Seção 5.2.

Já para as duas variáveis explicativas da parte inflacionada de zeros (*Heathcare_acces* e *Crowding*), as medidas resumo são muito similares no SAS e no R, existindo diferenças apenas na parte decimal dos valores. O teste de significância dos parâmetros também apresentou resultados equivalentes nos dois *softwares*.

Ainda nas Figuras 6.7 e 6.8, os valores dos quartis para as estimativas do parâmetro *Crowding* são maiores do que os respectivos resultados para *Healthcare_access*, indicando que a associação entre o nível de aglomeração e a probabilidade de se ter COVID-19 é mais forte do que a associação entre o acesso à saúde e a probabilidade de se ter essa doença. Essa observação está de acordo com o que foi concluído por Sousa (2022), por meio de testes de significância.

Pelas Figuras 6.9 e 6.10, é possível notar que os erros padrão da parte inflacionada de zeros foram, em geral, maiores no R. Comparando do SAS para o R, respectivamente, as médias do intercepto foram de magnitude 10^{19} e 10^{20}, as do parâmetro *Crowding* foram da magnitude de 10^{18} e 10^{19} e as de *Healthcare_access* tiveram a mesma magnitude de

Quantiles of GWR Zero Inflation Standard Errors			
	Intercept	HEALTHCARE_ACCESS	CROWDING
P25	7.6322511	0.2187123	0.3732483
P50	16.553474	0.4046865	0.8178866
P75	1113.9621	21.377397	107.75207
IQR	1106.3299	21.158685	107.37882

Descriptive Statistics of Zero Inflation Standard Errors			
	Intercept	HEALTHCARE_ACCESS	CROWDING
Mean	7.7316E19	2.4197E18	4.7708E18
Min	0	0	0
Max	9.5321E21	4.2693E20	7.536E20

Figura 6.9: Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros inflacionados de zero do modelo RBNIZGP no SAS.

```
$qntls_gwr_zero_infl_se
       Intercept Healthcare_access     Crowding
25%     8.568618          0.2411661    0.5017304
50%    22.109869          0.4944940    1.0215027
75% 32019.031744        470.9809885 1776.2627975
IQR 32010.463126        470.7398224 1775.7610671

$descript_stats_gwr_zero_infl_se
        Intercept Healthcare_access     Crowding
Mean 1.895142e+20      7.736350e+18 1.089555e+19
Min  7.059613e-21      1.192503e-22 1.529135e-22
Max  9.532354e+21      4.500386e+20 7.536176e+20
```

Figura 6.10: Quantis e estatísticas descritivas para os erros padrão das estimativas dos parâmetros inflacionados de zero do modelo RBNIZGP no `R`.

10^{18}, mas com maior valor no `R` ($2,42 < 7,73$).

As saídas para o modelo global, nas Figuras 6.11 e 6.12, foram idênticas entre SAS e `R`, considerando as casas decimais disponíveis para visualização. Além disso, as estimativas dos parâmetros condizem com a distribuição geral de estimativas apresentada anteriormente para os modelos locais: intercepto negativo; parâmetro de superdispersão próximo de 3; comorbidade, dificuldade de distanciamento social, aglomeração e comportamento da saúde sendo aspectos que se relacionam de forma positiva com o número de casos de COVID-19; enquanto proporção de pessoas com ensino médio completo, nível de acesso à saúde e migração possuem correlação negativa com a resposta.

A parte inflacionada de zeros também resultou em valores positivos para as estimativas dos parâmetros *Healthcare_access* e *Crowding*, indicado que o aumento nessas variáveis é acompanhado de aumento na probabilidade de se ter COVID-19, estando de acordo com o que foi observado nos modelos locais. Além disso, os erros padrão não apresentaram valores muito elevados, e os testes t para todas as covariáveis resultaram em p-valores pequenos, representando a significância dos parâmetros na explicação sobre

Global Parameter Estimates

	Par. Est.	Std Error	t Value	Pr > |t|
Intercept	-15.1281	1.8349091	-8.24	<.0001
MORBIDITY	0.0470072	0.008186	5.74	<.0001
HIGH_SCH_P	-0.110767	0.0359583	-3.08	0.0023
HEALTHCARE_ACCESS	-0.144631	0.0326653	-4.43	<.0001
DIFF_SD	0.0680944	0.0293187	2.32	0.0211
CROWDING	0.4354038	0.1250349	3.48	0.0006
MIGRATION	-0.367194	0.0849756	-4.32	<.0001
HEALTH_BEHAVIOR	0.051316	0.0192384	2.67	0.0082
alpha	3.1147571	0.2049051	15.20	<.0001

NOTE: The denominator degrees of freedom for the t tests is	236	.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimate

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-13.864379	3.619402	-3.83	0.0002
Healthcare_access	0.130669	0.060252	2.17	0.0311
Crowding	0.569421	0.222802	2.56	0.0112

NOTE: The denominator degrees of freedom for the t tests is	241	.

Deviance	Full Log Likelihood	pctll	adjpctll	AIC	AICc
286.623	-748.2946	0.3788865	0.3247114	1520.5892	1521.9399

Figura 6.11: Estimativas dos parâmetros e medidas de ajustamento para a versão global do modelo RBNIZGP no SAS.

```
$global_param_estimates
                    Par. Est.   Std Error   t Value     Pr > |t|
Intercept         -15.12809973 1.834909135 -8.244604 1.154632e-14
Morbidity           0.04700723 0.008186009  5.742386 2.858860e-08
high_sch_p         -0.11076695 0.035958326 -3.080426 2.312151e-03
Healthcare_access  -0.14463148 0.032665290 -4.427681 1.457036e-05
diff_sd             0.06809439 0.029318707  2.322558 2.105441e-02
Crowding            0.43540381 0.125034866  3.482259 5.923225e-04
Migration          -0.36719446 0.084975607 -4.321175 2.288365e-05
Health_behavior     0.05131596 0.019238443  2.667365 8.174228e-03
alpha               3.11475715 0.204905075 15.200976 0.000000e+00

$analysis_max_like_zero_infl_param_estimates
                    Estimate  Std Error   t Value     Pr > |t|
Intercept         -13.8643794 3.61940226 -3.830572 0.0001631042
Healthcare_access   0.1306688 0.06025153  2.168722 0.0310822946
Crowding            0.5694210 0.22280241  2.555722 0.0112117546

$analysis_max_like_gof_measures
    deviance      full_ll       pct_ll   adj_pct_ll          AIC
 286.6229976 -748.2946208    0.3788865    0.3247114 1520.5892417
        AICc
1521.9398910
```

Figura 6.12: Estimativas dos parâmetros e medidas de ajustamento para a versão global do modelo RBNIZGP no SAS.

a quantidade de casos da doença.

Em geral, as medidas de ajustamento indicam que o modelo global teve ajuste inferior ao RBNIZGP. Os indicadores de pseudo-R^2 que eram próximos de 1 para o modelo

mencionado, estão entre 0,3 e 0,4 para a versão global. Ademais, tanto o AIC como o AICc aumentaram quando comparados aos seus valores no modelo RBNIZGP ($1420,44 < 1520,59$ e $1472,83 < 1521,94$).

6.4 Visualização dos Resultados

Após o ajuste do modelo RBNIZGP através da função `gwzinbr`, é interessante observar como as variáveis envolvidas e as estimativas do modelo se distribuem espacialmente ao longo do país. Para isso, foram construídos mapas no `R` através do pacote leaflet, com versões interativas disponíveis no RPubs. Entre as covariáveis utilizadas no ajuste, a variável de aglomeração, *Crowding*, foi escolhida para exemplificar essas distribuições espaciais, por ser de grande relevância para a explicação do fenômeno e por estar incluída tanto na parte inflacionada de zeros como na parte não inflacionada do modelo.

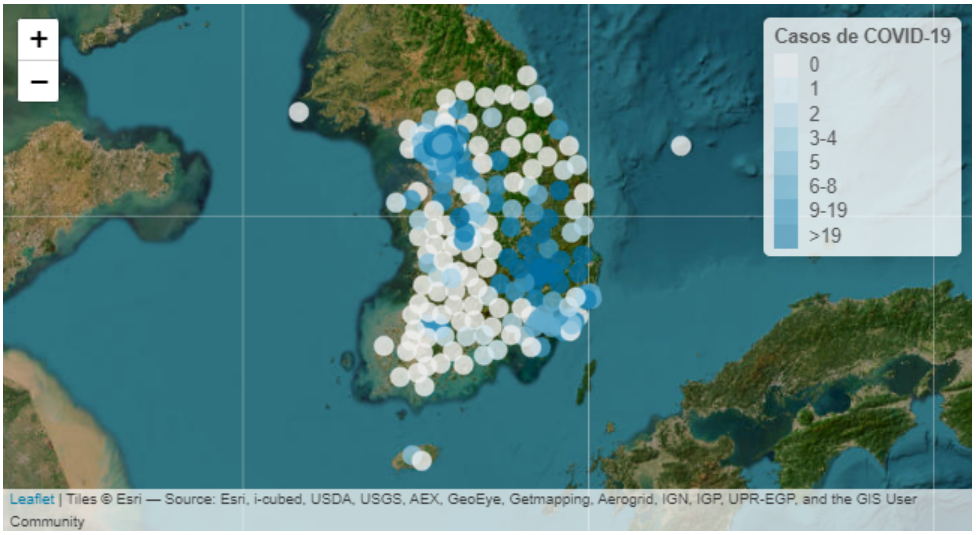


Figura 6.13: Distribuição espacial do número de casos de COVID-19 na Coreia do Sul durante fase inicial da pandemia

A Figura 6.13 apresenta a distribuição espacial dos casos de COVID-19 pela Coreia do Sul na fase inicial da pandemia, em que se observa uma maior concentração de zeros. Isso reforça a ideia de um comportamento inflacionado de zeros e que, complementado com a superdispersão identificada nos dados (os valores de casos de COVID-19 podem chegar até o máximo de 4155), caracteriza uma provável distribuição binomial negativa inflacionada de zeros.

É interessante observar que as regiões com maiores números de casos são também aquelas que figuram entre as principais do país: Seul (capital), Busan (onde está localizado o porto mais movimentado do país) e Daejeon (polo tecnológico). As localizações dessas cidades podem ser identificadas na Figura 6.14.

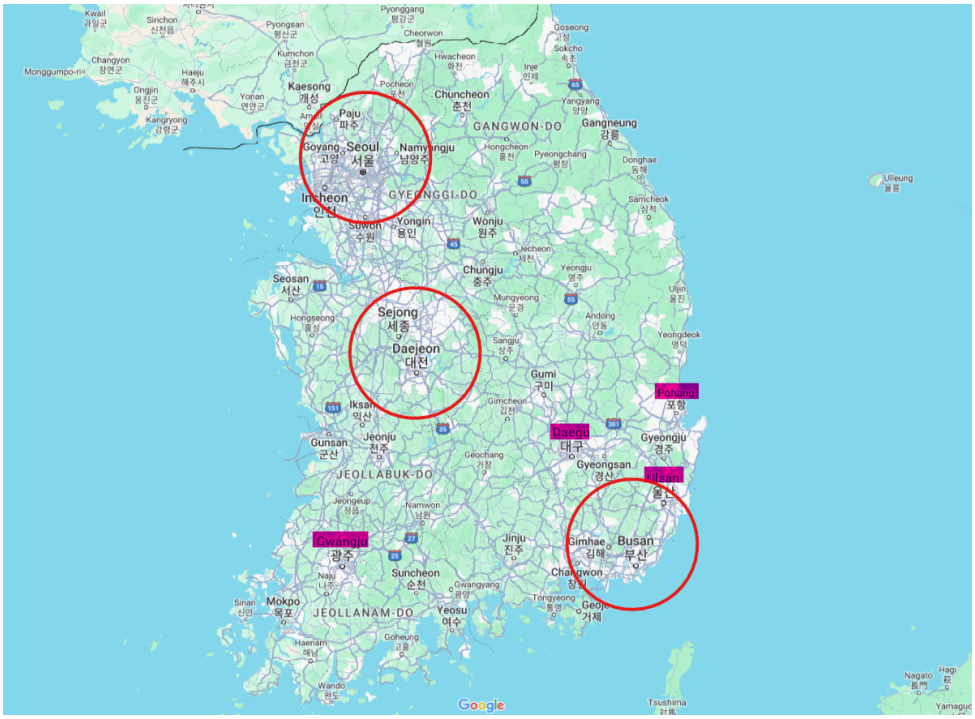


Figura 6.14: Mapa da Coreia do Sul

Fonte: Google Maps, com adaptações

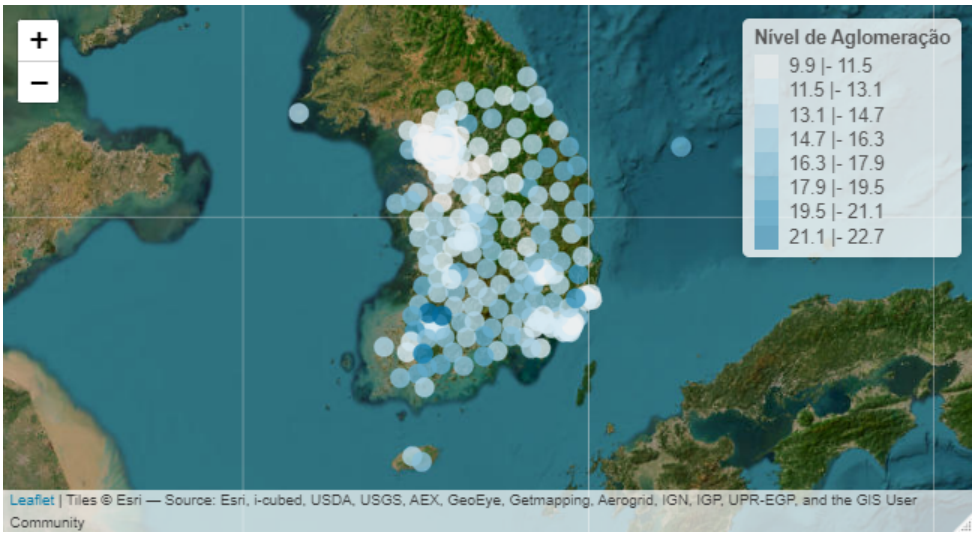


Figura 6.15: Distribuição espacial do nível de aglomeração (*Crowding*) na Coreia do Sul

Já na Figura 6.15 observa-se o comportamento espacial do nível de aglomeração (variável explicativa *Crowding*). São níveis que variam de moderado a baixo, e que traduzem uma política local ostensiva de combate à doença desde a fase inicial da pandemia.[1]

As estimativas para o parâmetro *Crowding* e suas distribuições no espaço podem

[1]BBC. Coronavírus: o que está por trás do sucesso da Coreia do Sul para salvar vidas em meio à pandemia. Disponível em: `https://www.bbc.com/portuguese/internacional-51877262`. Acesso em: 02 de junho de 2024

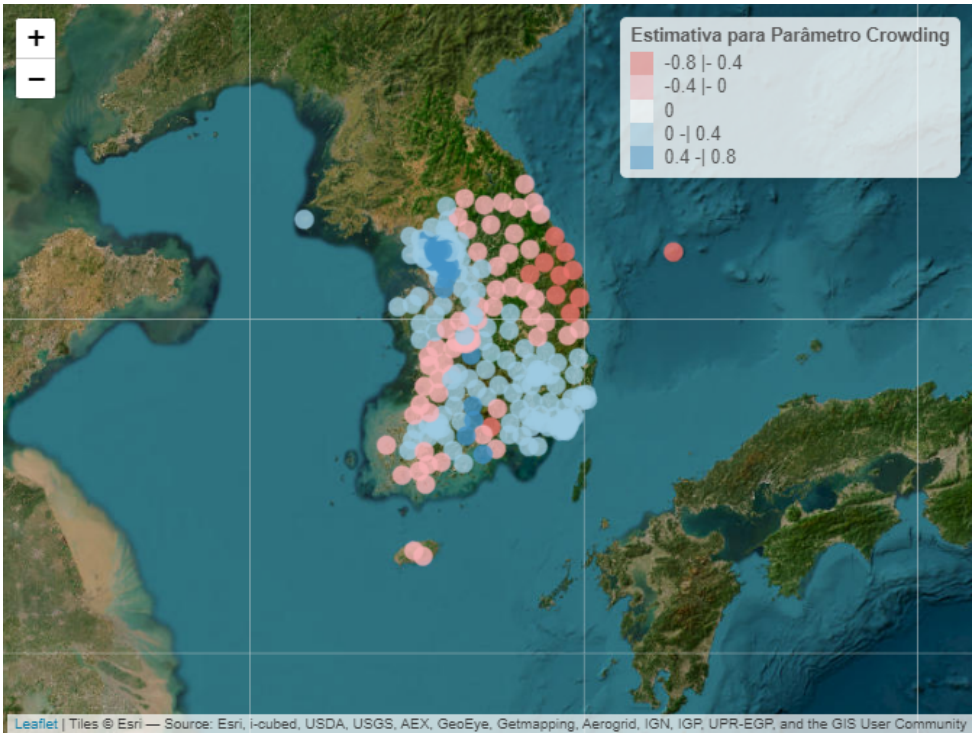


Figura 6.16: Distribuição espacial das estimativas do parâmetro *Crowding* no modelo RBNIZGP na Coreia do Sul

ser conferidas na Figura 6.16. A visualização auxilia no entendimento inicial de como essa variável pode impactar o número de contaminações pela doença. As cores mais claras no mapa representam valores mais próximos de zero, sugerindo que a aglomeração não tenha uma associação tão forte com a quantidade de casos naquelas regiões.

Uma pequena porção da região litorânea do país (em tons escuros de vermelho) indica uma associação negativa com a resposta, ou seja, o aumento no nível de aglomeração é acompanhado de um decréscimo no número de casos nesses locais. Em sentido contrário, porções em tons de azul mais escuro, localizadas principalmente próximo da capital, sugerem uma correlação positiva.

A Figura 6.17 apresenta as estimativas calculadas para o parâmetro *Crowding* na parte inflacionada de zeros do modelo. Ou seja, os valores dos coeficientes, nesse caso, representam a relação entre o nível de aglomeração e a ocorrência ou não de COVID-19. Assim, observa-se que a região mais nordeste do mapa contém alguns valores negativos para representar essa relação. Porém, de forma geral, na maior parte do país a aglomeração é uma característica que está associada positivamente com a probabilidade de se ter a doença.

Além disso, alguns locais próximos a Seul apresentaram valores nulos para a estimativa do coeficiente inflacionado de zeros, indicando que a aglomeração não têm

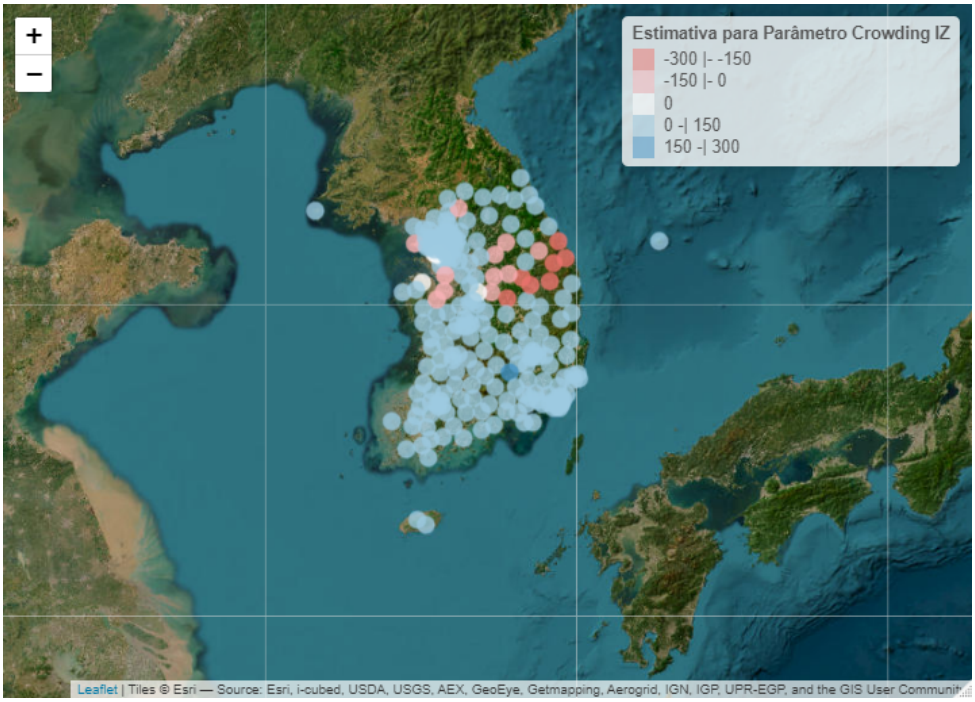


Figura 6.17: Distribuição espacial das estimativas do parâmetro *Crowding* inflacionado de zeros no modelo RBNIZGP na Coreia do Sul

significância na explicação da ocorrência de COVID-19 nessas localizações. E o único local onde o valor estimado foi muito elevado foi no condado de Hapcheon, mais ao sul do país, representado em azul escuro.

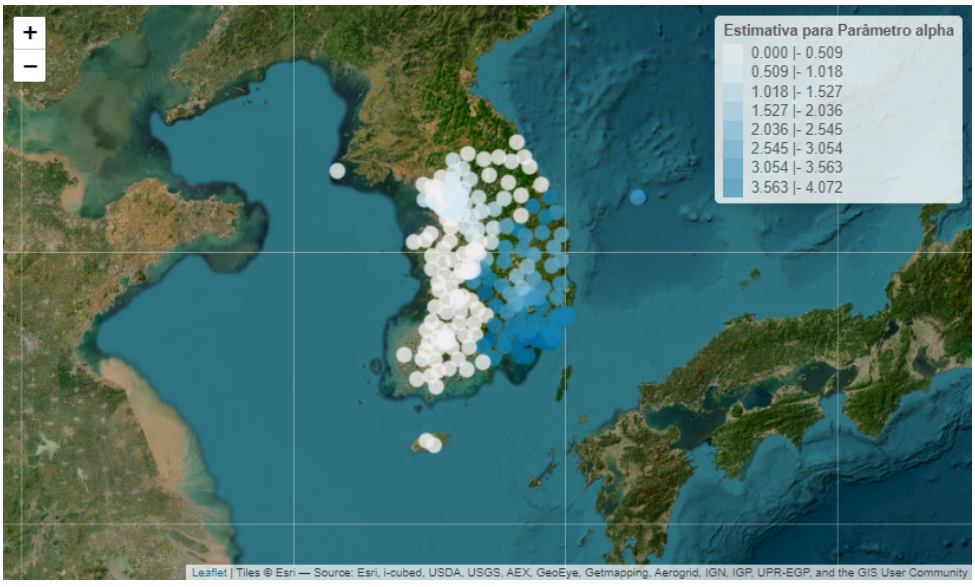


Figura 6.18: Distribuição espacial das estimativas do parâmetro de superdispersão no modelo RBNIZGP na Coreia do Sul

Por fim, a Figura 6.18 mostra a distribuição espacial das estimativas para o parâmetro de superdispersão α. A partir desse mapa, é possível ter uma ideia inicial de quais distribuições são mais adequadas para representar os dados de resposta em lo-

calidades específicas. Valores maiores para o parâmetro sugerem que distribuições que acomodam esse comportamento de alta variabilidade (como a binomial negativa e a binomial negativa inflacionada de zeros) são as mais indicadas para a modelagem. Isso ocorre para o leste do país, enquanto que para o oeste distribuições como a Poisson ou a Poisson inflacionada de zeros seriam o suficiente para descrever o número de casos.

Capítulo 7

Conclusões

O presente trabalho alcançou seu objetivo principal ao implementar um pacote `R` (atualmente já publicado no CRAN) para o algoritmo de Regressão Binomial Negativa Inflacionada de Zeros Geograficamente Ponderada (RBNIZGP), proposto por Da Silva e De Sousa (2023) e previamente implementado em SAS pelos referidos autores. O pacote `gwzinbr`, ainda que focado na RBNIZGP, é capaz de ajustar espacialmente dados de contagem que apresentem comportamentos compatíveis com pelo menos outras três distribuições conhecidas (Poisson, Binomial Negativa e Poisson Inflacionada de Zeros).

A comparação dos resultados obtidos por ambos os *softwares* a partir de um estudo de caso realizado pelos autores da teoria possibilitou não somente a compreensão do método pelas realizadoras deste relatório, como também a experiência de desenvolvimento e gerenciamento de pacotes no `R` e um maior entendimento sobre o SAS e seus aspectos computacionais. Ainda que algumas saídas não tenham resultado em uma replicação exata, o pacote desenvolvido em `R` se mostrou capaz de alcançar os resultados esperados, sendo uma alternativa acessível para pesquisadores.

Como as diferenças nos tempos de execução entre SAS e `R` mostraram que o pacote `gwzinbr` ainda pode ser melhorado em termos de eficiência computacional, espera-se que contribuições futuras possam otimizar o código. Uma sugestão seria o uso de processamento pararelo, o qual permite o aproveitamento de recursos computacionais físicos para que diferentes núcleos de uma máquina possam realizar diferentes cálculos ao mesmo tempo. Também é interessante a aplicação futura desse algoritmo em novos conjuntos de dados e a realização de estudos de simulação para melhor comparação entre as funções em SAS e `R`.

Referências Bibliográficas

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Da Silva, A. R. & De Sousa, M. D. R. (2023). Geographically weighted zero-inflated negative binomial regression: A general case for count data. *Spatial Statistics*, 58:100790.

Da Silva, A. R. & Fotheringham, A. S. (2016). The multiple testing issue in geographically weighted regression. *Geographical Analysis*, 48:233–247.

Da Silva, A. R. & Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression - incorporating overdispersion. *Statistics and Computing*, 24(5):769–783.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (2009). Geographically weighted regression. White paper, National Centre for Geocomputation. National University of Ireland Maynooth.

Dobson, A. J. & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*, (3rd ed.). Chapman and Hall/CRC.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Fumes, G. (2009). Uso de modelos inflacionados de zeros na análise de questionários de frequência alimentar. Master's thesis, Universidade Estadual Paulista Júlio de Mesquita Filho.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., & Lachis, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56:1030–1039.

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

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

Lambert, D. (1992). Zero-inflated poisson regression. *Technometrics*, 34:1–14.

Martin, J. & Hall, D. B. (2016). R^2 measures for zero-inflated regression models for count data with excess zeros. *Journal of Statistical Computation and Simulation*, 86(18):3777–3790.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Paula, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. Editora IME-USP.

Sousa, M. D. R. (2022). Regressão binomial negativa inflacionada de zeros geograficamente ponderada. Dissertação de Mestrado em Estatística, Universidade de Brasília.

Weinstein, B., Da Silva, A. R., Kouzoukas, D. E., Bose, T., Kim, G. J., Correa, P. A., Pondugula, S., Lee, Y., Kim, J., & Carpenter, D. O. (2021). Precision mapping of covid-19 vulnerable locales by epidemiological and socioeconomic risk factors, developed using south korean data. *International Journal of Environmental Research and Public Health*, 18:1–14.

Yau, K. K. W., Wang, K., & Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.