
UFAM

# MODELANDO DADOS DE CONTAGEM COM INFLAÇÃO DE ZEROS, SOBREDISPERSÃO E DEPENDÊNCIA ESPACIAL

Carla Zeline Rodrigues Bandeira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: Max Sousa de Lima

Manaus
Setembro de 2015

MODELANDO DADOS DE CONTAGEM COM INFLAÇÃO DE ZEROS, SOBREDISPERSÃO E DEPENDÊNCIA ESPACIAL

Carla Zeline Rodrigues Bandeira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinada por:

Prof. Max Sousa de Lima, D.Sc.

Prof. Fábio Nogueira Demarqui, Ph.D.

Prof. James Dean Oliveira dos Santos Junior, D.Sc.

MANAUS, AM – BRASIL
ABRIL DE 2015

Rodrigues Bandeira, Carla Zeline

MODELANDO DADOS DE CONTAGEM COM INFLAÇÃO DE ZEROS, SOBREDISPERSÃO E DEPENDÊNCIA ESPACIAL/Carla Zeline Rodrigues Bandeira. – Manaus: UFAM/ICE, 2015.

XIII, 67 p.: il.; 29, 7cm.

Orientador: Max Sousa de Lima

Dissertação (mestrado) – UFAM/ICE, Área de Concentração: Estatística, 2015.

Referências Bibliográficas: p. 64 – 67.

1. Inflação de Zeros. 2. Sobredispersão. 3. Dependência Espacial. 4. Quase Verossimilhança. 5. Equações de Estimação Generalizadas. 6. Algoritmo Expectation-Solution. 7. Inferência Bootstrap. I. Sousa de Lima, Max. II. Universidade Federal do Amazonas, UFAM, Área de Concentração: Estatística. III. Título.

*Este trabalho dedico à Deus, pois*
*sem Ele nada existiria.*

# Agradecimentos

Agradeço primeiramente à Deus, criador de todas as coisas, pelo dom da vida, por conceder-me saúde, perseverança e capacidade para o desenvolvimento deste trabalho.

À toda minha família pelo amor, carinho e companheirismo cultivado entre nós. Em especial, agradeço aos meus pais, Zeneide e Carlos, e ao meu padrasto, Aldenir, por todo ensinamento, amor e educação dados à mim ao longo da vida, e por sempre me incentivarem nos estudos. À minha avó, Zulmira, por todo ensinamento de vida e fé. Às minhas irmãs, Caroline e Camila, e sobrinha (filha), Ana Celine, por todo o amor, companheirismo, carinho e felicidade que me proporcionam todos os dias. Aos meus afilhados, Adrielle e Alberto, por existirem em minha vida. Aos meus primos, Lucélia, Jean, Eduarda e Haroldo, Alberto e Edna, pelos momentos de alegria e descontração. Aos meus tios, em especial à tia Nilda, por entender esses três anos de ausência da sua casa, à tia Carminha por sempre orar por mim e ao tio Nazareno, pelo amor e carinho que tem por mim como uma filha.

Aos membros da banca examinadora dessa defesa de dissertação, por aceitarem o convite para avaliar este trabalho. Ao meu orientador, professor Max Sousa de Lima, pela paciência, dedicação, confiança e incetivo na busca por conhecimento, contribuindo significativamente com o desenvolvimento deste e minha formação acadêmica.

Aos meus amigos e professores do Departamento de Estatística, que contribuíram direta e indiretamente na minha formação acadêmica, pelo apoio e pela consideração. Em especial, aos que foram meus professores do mestrado: James Dean, José Raimundo, Max Lima e Celso Rômulo, por todo conhecimento disseminado, ao professor José Cardoso, pela amizade e conselhos, aos professores e amigos Nelson Filho, Diego Souza, Camila Pinheiro, Carina Coelho, Márcia Brandão e Jocely Lopes, e amigas do mestrado, Renan, Vanessa, Renata e Regina, pelos momentos de descontração e trocas de conhecimento.

Finalmente, mas não menos importante, agradeço aos meus amigos Alessandra, Adriana, Adriano, Diana, Geane, Marcos, Patrícia e Raquel, por todos os momentos de alegria e diversão vividos, e por entenderem todas as minhas ausências.

À CAPES, pelo apoio financeiro em 10 meses de estudos.

*”Com efeito, de tal modo Deus amou o mundo, que lhe deu seu Filho único, para que todo o que Nele crer não pereça, mas tenha a vida eterna.”* (João 3:16).

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

MODELANDO DADOS DE CONTAGEM COM INFLAÇÃO DE ZEROS, SOBREDISPERSÃO E DEPENDÊNCIA ESPACIAL

Carla Zeline Rodrigues Bandeira

Setembro/2015

Orientador: Max Sousa de Lima

Área de Concentração : Estatística

Neste trabalho foi proposto um novo modelo para dados de contagem com excesso de zeros, sobredispersão e dependência espacial. Para acomodar simultaneamente essas características, utilizou-se uma quase verossimilhança inflacionada de zeros (**QIZ**), onde a dependência espacial foi incorporada no processo de estimação através das equações de estimação generalizadas (**GEE**). O algoritmo de estimação usado nesse processo foi o **ES** (*Expectation-Solution*); os intervalos de confiança para os parâmetros foram obtidos via Inferência Bootstrap. Estudos de simulação foram realizados considerando-se vários cenários. Finalmente, o método proposto foi ilustrado usando dados de casos de Hanseníase no Estado do Amazonas.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

MODELING COUNT DATA WITH ZEROS INFLATION, OVERDISPERSION AND SPATIAL DEPENDENCE.

Carla Zeline Rodrigues Bandeira

September/2015

Advisor: Max Sousa de Lima

Research area: Statistics

This work proposes a new model for count data with excess zeros, overdispersion and spatial dependence. To accommodate these characteristics simultaneously, we used an zero-inflated quasi-likelihood (**QIZ**), where the spatial dependence is incorporated in the estimation process through generalized estimating equations (**GEE**). The estimation algorithm used in this process was the **ES** (*Expectation-Solution*); confidence intervals for the parameters were obtained via Bootstrap Inference. Simulation studies have been performed in various scenarios. Finally, the method is illustrated using data of leprosy cases in the State of Amazonas.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

## 1.1 Aspectos Gerais

Na época atual, em que a tecnologia se torna a cada dia mais avançada, o surgimento de dados com comportamentos mais complexos têm requerido modelos estatísticos mais robustos, que consigam adequar-se a eles, ou seja, que modelem esses dados de forma correta. Com isso, a demanda por métodos mais sofisticados de análise e interpretação de dados com características mais completas crescem. Dentro desses novos tipos de dados, encontram-se os que estão geograficamente referenciados e correlacionados, que inspiraram a criação de novas técnicas para análise e modelagem, formando um campo da estatística, conhecido como análise de dados espaciais ou estatística espacial.

Em estatística espacial, os dados de contagem geralmente são modelados através de distribuições convencionais, como a Poisson ou a Binomial, o que talvez não seja adequado em muitos cenários. Por exemplo, em áreas como medicina, saúde pública ou epidemiologia é comum, devido a heterogeneidade da população, a contagem de casos de doenças apresentar maior variabilidade do que a prevista pela distribuição usual, pois as contagens em determinadas áreas são bem maiores do que a predita pelo modelo. Esse excesso de variabilidade é chamado de sobredispersão, cuja variância é maior do que a média, e tem sido amplamente considerado na literatura (Fahrmeir & Tutz (2001); Lima *et al.* (2013)). A falta de modelagem para a sobredispersão existente pode levar a uma subestimação do erro padrão e com isso ocasionar em inferências distorcidas para os parâmetros do modelo (ver Zhang *et al.* (2012)).

Outro problema comum em dados de contagens é que muitas vezes estes apresentam um número excessivo de zeros que não são esperados pelo modelo usual. Por exemplo, em um ambiente de vigilância epidemiológica, onde se realiza a contagem de novos casos de pessoas com hanseníase, pequenas áreas podem apresentar um menor número de casos de infectados em relação ao valor esperado predito, em decorrência da distância desses lugares em relação aos estabelecimentos de saúde. Além disso, a subnotificação de novos casos pode

ocorrer, em regiões subdesenvolvidas, devido à coleta de dados ineficiente ou à dificuldade de acesso a lugares remotos. Esses fatores geram contagens de casos com excesso de zeros, fazendo com que haja heterogeneidade no processo. Perumean-Chaney *et al.* (2013) verificaram em seu estudo que, ignorando o excesso de zeros nos dados, as estimativas para o modelo usual Poisson são equivocadas, pois há uma violação no modelo estatístico usual e, por consequência, em problemas de teste de hipóteses, o erro tipo I é inflado, acarretando em perdas de resultados estatisticamente significativos.

Modelos para dados inflacionados de zeros (**ZI**), têm sido usados em diversas áreas (Hall (2000); Cheung (2002); Yau *et al.* (2004); Warton (2005)). Parâmetros estimados usando **ZI** podem, também, ser severamente viciados se as contagens positivas forem substancialmente dispersas, ou seja, se houver a sobredispersão na parte positiva dos dados. Simultaneamente, dados de contagem podem apresentar essas duas fontes independentes de efeitos de sobredispersão. Se a sobredispersão é causada pela inflação de zeros, então o modelo Poisson inflacionado de zeros **ZIP**, introduzido por Lambert (1992), pode fornecer um ajuste suficiente para os dados. Uma vez modelada a inflação de zeros, se os dados continuam a sugerir sobredispersão adicional, devemos considerar um modelo de contagem que acomode também a sobredispersão nos valores positivos.

A não modelagem simultânea da sobredispersão e inflação de zeros pode causar uma inferência enganosa. Por exemplo, em um estudo simulado, Perumean-Chaney *et al.* (2013) verificaram que, quando a inflação de zeros nos dados for ignorada, as estimativas para o modelo Poisson são equivocadas e os resultados estatisticamente significativos podem ser perdidos. Quando a sobredispersão dentro do modelo inflacionado de zeros for ignorada, a estimativa do erro Tipo I é inflada. Nestes casos, os modelos inflacionados de zeros Poisson Generalizado (**ZIGP**), Poisson Duplo (**ZIDP**) ou o Binomial Negativo (**ZINB**) podem ser boas alternativas para a modelagem conjunta da inflação de zeros e sobredispersão nos dados (Lima & Duczmal (2014)).

Em processos de contagem geograficamente referenciados, os modelos **ZIGP**, **ZIDP** e **ZINB** são flexíveis para incorporar a inflação de zeros, a sobredispersão e o ajuste por covariáveis, mas ao mesmo tempo são limitados por não assumirem dependência ou existência de correlação espacial, o que sempre ocorre em problemas dessa natureza, pois dados coletados em áreas vizinhas tendem a ser mais similares (ou correlacionados) do que os obtidos em áreas mais distantes geograficamente. Um exemplo comum desta situação, ocorre na área de saúde pública, onde epidemiologistas estudam a variação geográfica dos casos de doenças para gerar e refinar hipóteses testáveis sobre a sua etiologia. Neste contexto, modelos hierárquicos têm sido propostos para utilizar localizações espaciais e seus vizinhos como substitutos para fatores de riscos desconhecidos ou não mensuráveis na análise dos casos da doença.

Para acomodar simultaneamente os problemas de sobredispersão, inflação de zeros e

dependência espacial em processos espaciais de contagem, propomos uma nova modelagem utilizando uma quase verossimilhança inflacionada de zeros (**QIZ**). A função de quase verossimilhança (**Q**) ou, mais precisamente, a função de quase log-verossimilhança, foi proposta por Wedderburn (1974) e reexaminada por McCullagh & Nelder (1983). Essa função pode ser usada para estimação de forma semelhante à função de verossimilhança. Sua grande vantagem é necessitar apenas da especificação da relação entre a média e a variância das observações, enquanto que na verossimilhança precisa-se especificar também a forma correta da distribuição das observações. A quase verossimilhança foi estendida por Nelder & Pregibon (1987) para incluir termos da variância, comparar diferentes funções de variância e, ainda, a possibilidade de modelar a dispersão (ou a sobredispersão) como uma função de covariáveis. A estimação dos parâmetros, neste caso, é realizada sobre a suposição de independência estatística entre as observações.

Nesta dissertação, propomos que o excesso de zeros e a sobredispersão sejam modelados, respectivamente, por uma distribuição de Bernoulli e pela quase verossimilhança estendida. O resultado é uma mistura de modelos representado por uma quase verossimilhança estendida inflacionada de zeros (**QIZ**). Para incorporar a dependência espacial, utilizamos no processo de estimação as equações de estimação generalizadas (**GEE**), propostas por Liang & Zeger (1986) , que construíram funções de estimação para os parâmetros de interesse na ausência da verossimilhança totalmente especificada e presença de correlação, que é exatamente o nosso caso.

O algoritmo de estimação utilizado nesse processo foi o **ES** (Expectation-Solution) (ver Elashoff & Ryan (2004)), que consiste na substituição do passo de maximização (**M**) no algoritmo **EM** por um passo que requer a solução (**S**) de uma equação de estimação generalizada. No contexto de mistura de GLM's, Rosen *et al.* (2000) mostraram que se o algoritmo **ES** convergir, ele convergirá para um estimador não-viciado, consistente e assintoticamente Normal, sob suaves condições de regularidade.

## 1.2 Justificativa e Importância do Trabalho

Devido ao avanço tecnológico, surgiram várias estruturas de dados mais complexas, incluindo estes que englobam excesso de zeros, sobredispersão e dependência espacial. Essas novas estruturas exigem modelos compatíveis com seus comportamentos, que consigam modelar os dados da melhor forma, nos dando estimativas e interpretações corretas dos dados. A grande importância deste trabalho é por haver pouquíssimos trabalhos nessa área de estatística espacial, que englobe ao mesmo tempo aos dados de contagem o excesso de zeros, sobredispersão e dependência espacial, além da grande aplicabilidade do modelo em dados com essas características, como por exemplo dados epidemiológicos, de saúde pública e de criminalidade, podendo ajudar na vigilância epidemiológica e na análise de incidência de

crimes.

## 1.3 Objetivos

Este trabalho teve como principal objetivo a modelagem de processos de contagem, com excesso de zeros, sobredispersão e dependência espacial. Como metas e objetivos específicos tivemos:

1) A construção de modelos para processos de contagem distribuídos no espaço com excesso de zeros, sobredispersão e correlação espacial;

2) O desenvolvimento e implementação do algoritmo **ES** para a estimação de parâmetros dos modelos propostos;

3) A realização da análise dos modelos propostos com dados reais;

4) A realização de estudos com dados simulados para avaliar o desempenho do método proposto em vários cenários.

## 1.4 Estrutura do Trabalho

A dissertação está estruturada como segue: no Capítulo 2 descreveremos as principais abordagens utilizadas no desenvolvimento deste trabalho, que são o modelo para dados de contagem inflacionados de zeros (**ZI**), com sua representação estocástica e seu processo de estimação via algoritmo EM, as funções de quase verossimilhança (**Q**) e quase verossimilhança estendida $Q^+$, com suas características e propriedades, as equações de estimação generalizadas **GEE**, o algoritmo **ES** e intervalos de confiança *Bootstrap*. No Capítulo 3 introduziremos o modelo proposto, de quase-verossimilhança inflacionada de zeros (**QIZ**), para dados sem e com dependência espacial, suas caracterização, propriedades e estimação via algoritmo **ES**, e também apresentaremos a inferência *Bootstrap*, com a definição de intervalos de confiança gerados por este método. Estudos de simulação comparativos, realizados em diversos cenários, a fim de comparar o desempenho do modelo são discutidos no Capítulo 4. Uma aplicação do modelo proposto em dados reais de hanseníase, da região norte do Brasil, é apresentada no Capítulo 5. Finalmente, no Capítulo 6 discutimos os resultados obtidos, as principais conclusões e propostas de trabalhos futuros.

# Capítulo 2

# Fundamentação Teórica

Neste capítulo faremos um breve levantamento da teoria utilizada no desenvolvimento deste trabalho. "Os principais tópicos abordados"são os modelos inflacionados de zeros, quase-verossimilhança, equações de estimação generalizada, modelos de regressão espacial, algoritmo *Expectation-Solution* (**ES**) e intervalos de confiança *Bootstrap*, com suas respectivas definições e propriedades.

## 2.1 O Modelo Inflacionado de Zeros

Eventualmente, é bem comum dados de contagem apresentarem um número excessivo de zeros. Esses zeros podem ser ocasionados por diferentes processos inerentes aos dados. Em diversas áreas, como por exemplo vigilância epidemiológica, saúde pública, biologia, sociologia, engenharia, agricultura e criminalidade, dados com essa característica surgem facilmente. Um exemplo, em vigilância epidemiológica, seria a contagem de novos casos de hanseníase, em determinada localização (cidade, estado, região, país, etc.), apresentar uma quantidade de zeros acima do predito pelo modelo probabilístico proposto para os dados, como por exemplo os modelos de Poisson ou Binomial. Esse excesso de zeros pode ter ocorrido, por exemplo, pela subnotificação de casos, devido a dificuldade de acesso a lugares remotos para o registro de novos casos, ou pela não ocorrência de novos casos nessa localidade. Para esse problema, existem diversos modelos de regressão inflacionados de zeros (**ZI**) que podem perfeitamente modelar esses tipos de dados.

Uma forma bastante simples de verificar se dados de contagem possuem ou não um excesso de zeros é através da quantidade: $zi = \widehat{p}_0 - \widehat{P}(Y = 0)$, a qual chamaremos de índice de inflação de zeros, em que $Y$ é uma variável aleatória discreta, $\widehat{p}_0$ é a proporção de zeros nos dados e $\widehat{P}(Y = 0)$ é a probabilidade de ocorrer o zero segundo o modelo de contagem proposto. Se o valor de $zi <= 0$, diz-se que a variável aleatória Y segue a distribuição de contagem usual proposta, e se $zi > 0$, modelos **ZI** que acomodem a inflação de zeros são mais adequados para os dados Lambert (1992).

Atualmente, existem diversos modelos (**ZI**), que são bastante utilizados na literatura para modelar dados de contagem com excesso de zeros, como por exemplo os modelos: Poisson Inflacionado de Zeros (**ZIP**), Binomial Inflacionado de Zeros (**ZIB**), Poisson Generalizado Inflacionado de Zeros (**ZIGP**) e Conway-Maxwell Inflacionado de Zeros (**ZICM**). Johnson & Kotz (1969) desenvolveram o modelo (**ZIP**) sem efeito de covariáveis, Lambert (1992) adicionou ao modelo **ZIP** o efeito de covariáveis, aplicando esse modelo em dados de contagem de defeitos de manufatura. Hall (2000) adaptou o modelo **ZIP** de Lambert (1992) e desenvolveu o **ZIB** incorporando efeitos mistos e sobredispersão, aplicando-o em dados de horticultura. Podemos encontrar, ainda, aplicações dos modelos **ZI** nos contextos de estatística espacial (Cancado *et al.* (2011); Lima *et al.* (2013)), séries temporais (Yang (2012)), nas áreas de medicina (Van den Broek (1995)), de biologia (Nie *et al.* (2006)) e em construções de novos modelos, como por exemplo o modelo de regressão Poisson inflacionado de zeros multinível (Lee *et al.* (2006) e modelos marginais para dados agrupados inflacionados de zero Hall & Zhang (2004)).

A teoria sugere que os zeros excedentes são gerados por um processo separado dos valores da contagem, e que estes podem ser modelados de maneira independente. Dessa forma, os modelos **ZI** são representados por uma mistura de duas distribuições. Seja $\mathbf{Y} = (Y_1, ..., Y_n)^T$ um vetor de observações de contagens independentes, dizemos que $Y_i$ segue um modelo inflacionado de zeros $\mathbf{ZI}(p_i, \theta)$, se sua distribuição é da forma:

$$Y_i \sim p_i \, \mathbb{I}_{\{y=0\}} + (1-p_i)\mathbb{P}_\theta \tag{2.1}$$

em que $p_i$ é a probabilidade de ocorrer o zero estrutural, $\mathbb{I}$ é a função indicadora, $(1-p_i)$ é a probabilidade de $Y_i$ seguir uma distribuição de contagem $\mathbb{P}_\theta$, parametrizada pelo vetor $\theta$. A função de probabilidade de $Y_i$ é dada por:

$$f(y_i; p_i, \theta) = \begin{cases} p_i + (1-p_i)P_\theta(Y_i = y_i), & \text{se } y_i = 0, \\ (1-p_i)P_\theta(Y_i = y_i), & \text{se } y_i > 0. \end{cases} \tag{2.2}$$

Estocasticamente, o modelo admite a seguinte representação:

$$Y_i | U_i = (1-U_i)Z_i \tag{2.3}$$

em que $U_i$ é uma a variável latente, seguindo a distribuição de Bernoulli$(p_i)$ e $Z_i$ segue uma distribuição de contagem $P_\theta$, com média $\mu_i$, variância $\sigma_i^2$ e $\theta = (\mu_i, \sigma_i^2)$, com $U_i$ e $Z_i$ independentes. É fácil mostrar que, marginalmente, $Y_i$ segue um modelo **ZI** com valor esperado e variância, respectivamente, dados por:

$$\mathbb{E}(Y_i) = (1-p_i)\mu_i \quad \text{e} \quad Var(Y_i) = (1-p_i)\sigma_i^2 + p_i(1-p_i)\mu_i^2. \tag{2.4}$$

Por exemplo, se $P_\theta$ é uma Poisson, então o valor esperado e a variância desse modelo são da forma:

$$\mathbb{E}(Y_i) = (1-p_i)\mu_i \quad \text{e} \quad Var(Y_i) = (1-p_i)\mu_i + p_i(1-p_i)\mu_i^2,$$

com função de probabilidade

$$f(y_i) = \begin{cases} 0, & \text{com probabilidade } p_i + (1-p_i)e^{-\mu_i}, \\ y_i, & \text{com probabilidade } (1-p_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 1,2,\ldots. \end{cases}$$

Usando a representação estocástica (2.3), a função de verossimilhança aumentada, $f(\mathbf{y},\mathbf{u};\mathbf{p},\theta)$, com dados observados $\mathbf{y} = (y_1,\ldots,y_n)$ e não observados $\mathbf{u} = (u_1,\ldots,u_n)$, é descrita da seguinte forma:

$$f(\mathbf{y},\mathbf{u};\mathbf{p},\theta) = \prod_{i=1}^{n} p_i^{u_i}[(1-p_i)P_\theta(Y_i = y_i)]^{1-u_i}. \tag{2.5}$$

Então, a log-verossimilhança completa do modelo **ZI** é dada por:

$$\begin{aligned} l^c(\mathbf{p},\theta;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{n}[u_i \log p_i + (1-u_i)\log(1-p_i)] + \sum_{i=1}^{n}(1-u_i)\log P_\theta(Y_i = y_i) \\ &= l^c(\mathbf{p};\mathbf{u}) + l^c(\theta;\mathbf{y},\mathbf{u}) \end{aligned} \tag{2.6}$$

Note que em (2.6) temos a soma de duas log-verossimilhanças completas, uma $l^c(\mathbf{p};\mathbf{u})$ que depende de dados não observados $\mathbf{u}$ e do vetor de parâmetros $\mathbf{p}$ e outra $l^c(\theta;\mathbf{y},\mathbf{u})$ que depende dos dados observados $\mathbf{y}$ e não observados $\mathbf{u}$, e do vetor de parâmetros $\theta$. No contexto de modelos lineares generalizados (**GLM**) isso pode ser interpretado como uma mistura de dois **GLM**'s Nelder & Wedderburn (1972).

Um **GLM** (ver Nelder & Wedderburn (1972)) é formado por três componentes, como segue:

(i) **Componente Aleatório**: composto pela variável resposta $Y_i$, que é assumida pertencer à família exponencial com função de probabilidade ou função densidade de probabilidade $f(y_i;\theta_i,\phi)$, em que $\phi$ é o parâmetro de dispersão, fixo e conhecido, $\theta_i$ é o parâmetro que caracteriza a distribuição e $Y_i$'s independentes.

(ii) **Componente Sistemático**: composto por p covariáveis $\mathbf{X} = (X_1,...,X_p)$ e por parâmetros de regressão desconhecidos $\beta = (\beta_1,...\beta_p)$. Assim, podemos expressar a média como uma função de $\eta_i = \mathbf{X}_i^T\beta$. O parâmetro $\eta$ é chamado de preditor linear.

(iii) **Função de Ligação**: é uma função monotônica diferenciável que associa o componente aleatório ao componente sistemático. Então, ao preditor linear teremos associ-

ada uma função da média $g(\mu_i) = \eta_i$, em que o valor esperado de $Y_i$ é representado por $\mu_i$ e sua variância por $\phi V(\mu_i)$. A quantidade $V(\mu_i)$ é chamada de função de variância do modelo.

Sendo assim, no modelo **ZI**, teremos dois **GLM**'s dados da seguinte forma:

No primeiro **GLM**(**p**), para a modelagem do excesso de zeros, considerando **G** uma matriz de covariáveis, $\gamma$ um vetor de parâmetros de regressão e a função $logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$, teremos:

$$U_i = u_i, \quad \tau_i = \mathbf{G}_i^T \gamma \quad \text{e} \quad g(p_i) = logit(p_i) = \tau_i \tag{2.7}$$

E no segundo **GLM**($\theta$), para a modelagem da média, considerando **B** uma matriz de covariáveis e $\beta$ um vetor de parâmetros de regressão, teremos:

$$Y_i = y_i, \quad \eta_i = \mathbf{B}_i^T \beta \quad \text{e} \quad g(\mu_i) = log(\theta_i) = \eta_i \tag{2.8}$$

A partir da descrição dos modelos, é preciso estimar seus parâmetros e avaliar a precisão das estimativas. Na mistura de **GLM**'s, a estimação dos parâmetros é realizada via algoritmo **EM**.

### 2.1.1 Estimação via Algoritmo EM

Para encontrar os estimadores dos parâmetros desconhecidos **p** e $\theta$, utiliza-se o método de máxima verossimilhança via algoritmo **EM** (Dempster *et al.* (1977)), o qual é indicado quando o conjunto de dados é incompleto ou envolve quantidades não observáveis (variáveis latentes). No último caso, as variáveis latentes podem ser incorporadas ao modelo propositalmente para facilitar a estimação dos parâmetros de interesse, tendo em vista que com sua inserção no modelo, a log-verossimilhança completa pode ser reescrita como a soma de duas log-verossimilhanças completas, como visto em (2.6). Em cada iteração, o algoritmo **EM** alterna entre as operações de Esperança (passo **E**) e de Maximização (passo **M**).

Considere $\vartheta = (\mathbf{p}, \theta)$ o vetor de parâmetros do modelo. O algoritmo **EM** maximiza a função log-verossimilhança $l(\vartheta; \mathbf{y}, \mathbf{u})$ usando a verossimilhança completa (2.5) e a distribuição condicional $f(\mathbf{u}|\mathbf{y}, \vartheta)$ de **u** dado **y** e $\vartheta$. Assim, a maximização de (2.6) via algoritmo **EM** ocorre em dois passos.

**Passo E**: Inicialize o processo iterativo com $\vartheta^{(0)} = (\mathbf{p}^{(0)}, \theta^{(0)})$ e na (k+1)-ésima iteração a estimativa de $u_i^{(k)}$ é a esperança condicional sobre **y** e a estimativa corrente $\vartheta^{(k)}$. Isto é, compute $\mathbb{E}\left\{ l^c(\mathbf{p}, \theta; \mathbf{y}, \mathbf{u}) | \mathbf{y}, \vartheta^{(k)} \right\}$ com respeito a distribuição condicional de **u**. Como a $l^c$ é linear em **u**, a esperança condicional é dada por $l^c(\vartheta^{(k)}; \mathbf{y}, \mathbf{u}^{(k)}) = l^c(\mathbf{p}, \theta; \mathbf{y}, \mathbf{u}^{(k)})$, em

que na k-ésima iteração faremos $\mathbf{u}^{(k)} = \mathbb{E}(\mathbf{U}|\mathbf{y}, \vartheta^{(k)})$ , com i-ésimo elemento

$$u_i^{(k)} = P(u_i = 1|y_i, \vartheta^{(k)}) = \frac{P(Y_i = y_i|u_i = 1, \vartheta^{(k)})p_i^{(k)}}{P(Y_i = y_i|u_i = 1, \vartheta^{(k)})p_i^{(k)} + P(Y_i = y_i|u_i = 0, \vartheta^{(k)})(1 - p_i^{(k)})}.$$

Usando (2.7) e (2.8), podemos encontrar a seguinte expressão para $u_i^{(k)}$:

$$u_i^{(k)} = \begin{cases} \left(1 + exp\{-logit(p_i^{(k)}) + l(\theta^{(k)}; y_i)\}\right)^{-1} & \text{se} \quad y_i = 0 \\ 0 & \text{se} \quad y_i > 0 \end{cases}$$

**Passo M**: Como temos uma mistura de **GLM**'s, maximizar $l^c(\vartheta^{(k)}; \mathbf{y}, \mathbf{u}^{(k)})$ é equivalente a maximizar cada **GLM**, separadamente, em relação aos seus respectivos parâmetros, da seguinte maneira:

(i) Passo M para $\mathbf{p}$: na $(k+1)$-ésima iteração, maximizar a $l^c(\mathbf{p}, \theta; \mathbf{y}, \mathbf{u}^{(k)})$ com relação a $\mathbf{p}$ é equivalente a maximizar $l^c(\mathbf{p}; \mathbf{u})$, considerando $\mathbf{u} = \mathbf{u}^{(k)}$.

(ii) Passo M para $\theta$: na $(k+1)$-ésima iteração, maximizar a $l^c(\mathbf{p}, \theta; \mathbf{y}, \mathbf{u}^{(k)})$ com relação a $\theta$ é equivalente a maximizar $l^c(\theta; \mathbf{y}, \mathbf{u})$, considerando $\mathbf{u} = \mathbf{u}^{(k)}$.

A Informação de Fisher para a mistura de **GLM**'s é dada por:

$$\mathscr{I}(\vartheta) = -E\left(\frac{\partial S(\vartheta)}{\partial \vartheta}\right),$$

em que as funções escore, para cada **GLM**, são escritas da seguinte forma:

$$S(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu}{\partial \beta}\right)^T [\phi V(\mu_i)]^{-1}(y_i - \mu_i), \tag{2.9}$$

e

$$S(\gamma) = \sum_{i=1}^{n} \left(\frac{\partial \mathbf{p}}{\partial \gamma}\right)^T [\phi V(p_i)]^{-1}(u_i - p_i), \tag{2.10}$$

podemos maximizar a $\mathbb{E}\left\{l^c(\mathbf{p}, \theta; \mathbf{y}, \mathbf{u})|\mathbf{y}, \vartheta^{(k)}\right\}$ através do algoritmo de Newton Raphson-Scoring de Fisher (**NR-SF**). De acordo com os seguinte passos:

**1.** Inicializar a iteração com o valor de $\vartheta^{(0)}$;

**2.** Para $k \longrightarrow k+1$ atualizar o valor de $\vartheta$, via $\vartheta^{(k+1)} = \vartheta^{(k)} + (\mathscr{I}^{(k)})^{-1} S(\vartheta^{(k)})$.

**3.** Repetir o passo **2** até que $\|\vartheta^{(k+1)} - \vartheta^{(k)}\| < \varepsilon$, ou seja, até se obter a convergência.

Quando a função de ligação $g(.)$ é canônica, teremos a seguinte expressão

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial \eta_i}{\partial \mu_i} = V^{-1}(\mu_i).$$

Então, reescrevendo os **GLM's** a partir dessa suposição, temos:

1. No **GLM(p)**, fixando $\phi = 1$, em que $\mu = p$ e a função de variância $V(p_i) = p_i(1-p_i)$. Definindo as seguintes expressões de forma matricial:

$$W_i = V^{-1}(p_i), \quad \Delta = diag(g'(p_i), ..., g'(p_n)), \quad \mathbf{h} = \eta + \Delta(\mathbf{u} - p),$$

em que a matriz de covariâncias $\mathbf{W} = diag(W_1, ..., W_n)$ é diagonal $(n \times n)$, e o valor de $u_i$ também obtido no passo **E** por

$$u_i^{(k)} = \left(1 + exp\{-logit(p_i(\gamma^{(k)})) + l_i(\theta_i(\beta^{(k)}); 0)\}\right)^{-1} I_{\{y_i=0\}},$$

o estimador para o parâmetro $\gamma$ será obtido, também, via algoritmo **NR-SF**, pela expressão:

$$\gamma^{(k+1)} = (\mathbf{G}^T \mathbf{W}^{(k)} \mathbf{G})^{-1} \mathbf{G} \mathbf{W}^{(k)} \mathbf{h}^{(k)}.$$

2. Da mesma forma, no **GLM($\theta$)** definimos as expressões matriciais $\widetilde{W}_i = (1 - u_i)^2 V^{-1}(\mu_i)$, $\widetilde{\triangle} = diag(g'(\mu_1), ..., g'(\mu_n))$, $\widetilde{\mathbf{h}} = \eta + \triangle(\mathbf{y} - \mu)$ e $\widetilde{\mathbf{W}} = diag(\widetilde{W}_1, ..., \widetilde{W}_n)$. Assim, o estimador para o parâmetro $\beta$ será obtido, via algoritmo **NR-SF**, através da expressão:

$$\beta^{(k+1)} = (\mathbf{F}^T \widetilde{\mathbf{W}}^{(k)} \mathbf{F})^{-1} \mathbf{F} \widetilde{\mathbf{W}}^{(k)} \widetilde{\mathbf{h}}^{(k)}.$$

Note que para se construir uma função de verossimilhança é necessário pressupor um modelo probabilístico, a partir do qual se especifica a função de probabilidade e define-se os intervalos de valores dos parâmetros do modelo. Essa especificação implica em que se detenha o conhecimento prévio do modelo, ou seja, saber através de qual mecanismo os dados foram gerados, ou basear-se em experiências anteriores significativas sobre dados semelhantes. No entanto, algumas vezes, não queremos ou não podemos assumir previamente algum modelo probabilístico para os dados. Neste caso, uma abordagem via função de quase verossimilhança é mais adequada.

## 2.2 Quase Verossimilhança

Um conceito muito importante ao longo deste trabalho é o de quase verossi-milhança, que pode ser utilizado quando não queremos ou não podemos assumir um modelo probabilístico para os dados. Na função de verossimilhança é necessário pressupor o modelo probabilístico para os dados. Na quase verossimilhança, por outro lado, somente o primeiro e

segundo momentos da distribuição dos dados precisam ser definidos, além disso, a variância de cada observação é especificada como sendo igual ou proporcional a alguma função da média. A função de quase verossimilhança para modelos lineares generalizados, foi proposta por Wedderburn (1974), reexaminada por McCullagh & Nelder (1983) e é definida a seguir.

**Definição 1.** *Considerando $Y_i, i = 1,...,n$, variáveis aleatórias independentes, com média $\mathbb{E}(Y_i) = \mu_i$ e variância $Var(Y_i) = a(\phi)V(\mu_i)$, onde $V$ é alguma função conhecida, denominada função de variância, e $a(\phi)$, que mede a dispersão do modelo, pode ser desconhecida. Suponha que cada $\mu_i$ é uma função conhecida de um conjunto de parâmetros $\beta_1,...,\beta_p$. E ainda, suponha que $a(\phi)$ é uma constante, que não depende de $\beta_1,...,\beta_p$. Então, para cada observação, definimos a função de quase verossimilhança $Q(y_i,\mu_i)$ pela relação:*

$$\frac{\partial Q(y_i,\mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{Var(Y_i)} \tag{2.11}$$

*ou de forma equivalente,*

$$Q(\mathbf{y},\mu) = \sum_{i=1}^{n} Q(y_i,\mu_i) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{a(\phi)V(t)} dt \tag{2.12}$$

Sob a suposição de independência dos componentes do vetor resposta $\mathbf{Y}$, a matriz $V(\mu)$ deve ser diagonal. Assim, podemos escrevê-la da seguinte forma:

$$V(\mu) = diag\{V_1(\mu),...,V_n(\mu)\}$$

Uma hipótese relevante sobre a função $V_i(\mu)$ é que ela deve depender apenas da i-ésima componente de $\mu$.

Por analogia, a função quase-desvio ($D$), que mede a discrepância entre as observações e seus valores esperados, é obtida de forma análoga à estatística da razão de log-verossimilhanças. Assim, para a i-ésima observação correspondente, essa função é escrita da forma:

$$D(y_i;\mu_i) = 2a(\phi)Q(\mu_i;y_i) = -2\int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt \tag{2.13}$$

que é uma função estritamente positiva, exceto em $y_i = \mu_i$. O desvio total $D(y;\mu)$ é uma função que não depende de $a(\phi)$, mas de $\mathbf{y}$ e de $\mu$ somente. Essa função é obtida de forma análoga à estatística da razão de log-verossimilhanças.

A função Q tem muitas propriedades em comum com a função de log-verossimilhança. De forma particular, se a distribuição de Y pertencer à família exponencial uniparamétrica, podemos mostrar que Q é a função de log-verossimilhança. Em seu artigo, Wedderburn (1974) demonstrou que a quase-verossimilhança tem as seguintes propriedades, semelhantes as da log-verossimilhança:

(i) $\mathbb{E}\left(\frac{\partial Q}{\partial \mu}\right) = 0$;

(ii) $\mathbb{E}\left(\frac{\partial Q}{\partial \beta_i}\right) = 0$;

(iii) $\mathbb{E}\left(\frac{\partial Q}{\partial \mu}\right)^2 = -\mathbb{E}\left(\frac{\partial^2 Q}{\partial \mu^2}\right) = \frac{1}{a(\phi)V(\mu)}$;

(iv) $\mathbb{E}\left(\frac{\partial Q}{\partial \beta_i}\frac{\partial Q}{\partial \beta_j}\right) = -\mathbb{E}\left(\frac{\partial^2 Q}{\partial \beta_i \partial \beta_j}\right) = \frac{1}{a(\phi)V(\mu)}\frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j}$;

Assim, para estimarmos o parâmetro $\beta$, utilizaremos a função quase-escore (2.14), ou seja, resolveremos a seguinte equação $S(\beta) = 0$, sendo que

$$\begin{aligned}
\frac{\partial Q}{\partial \beta} = S(\beta) &= \sum_{i=1}^{n} S_i(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T Var^{-1}(Y_i;\beta,\phi)(y_i - \mu_i(\beta)) \\
&= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T [a(\phi_i)V(\mu_i)]^{-1}(y_i - \mu_i(\beta)) \\
&= \sum_{i=1}^{n} (H_i)^T (V_i)^{-1}(y_i - \mu_i(\beta)) \\
&= \mathbf{H}^T \mathbf{V}^{-1}(\mathbf{Y} - \mu), \qquad (2.14)
\end{aligned}$$

em que o componente $\mathbf{V} = diag(a(\phi)V(\mu_i))$ é uma matriz diagonal $(n \times n)$ e $\mathbf{H}$ é uma matriz de derivadas $(n \times p)$, em que cada elemento dela corresponde à derivada $(\frac{\partial \mu_i}{\partial \beta_j})$, com $i = 1,...,n$ e $j = 1,...,p$.

A função quase escore (2.14) é um caso bastante especial, pois ela tem a forma de uma equação de estimação generalizada (**GEE**) de Liang & Zeger (1986), sob a suposição de independência. Uma **GEE**, denotada por $g(y;\theta)$, é uma função dos dados $y$ e dos parâmetros $\theta$, tendo média zero para todo o espaço paramétrico de $\theta$, de forma que a $E[g(y;\theta)] = \mathbf{0}$ (McCullagh & Nelder (1989)).

### 2.2.1 Estimação de Parâmetros

Considerando a função quase escore (2.14), podemos definir sua matriz de covariâncias, que é equivalente ao negativo do valor esperado da derivada de (2.14), dada por:

$$i_\beta = H^T V^{-1} H. \qquad (2.15)$$

No contexto de quase verossimilhança, essa matriz $i_\beta$ desempenha o mesmo papel que a informação de Fisher nas funções de verossimilhança comuns. Então, se iniciarmos o processo de estimação com um valor arbitrário de $\widehat{\beta}_0$, suficientemente próximo de $\widehat{\beta}$, a

sequência de estimativas dos parâmetros, gerados pelo método **NR-SF**, obtidas pela iteração até a ocorrência de convergência, é dada por:

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + (\mathbf{H^T V^{-1} H})^{-1}\mathbf{H^T V^{-1}}(\mathbf{y} - \mu) \tag{2.16}$$

onde $\widehat{\beta}^{(0)} = \widehat{\beta}_0$.

Wedderburn (1974) e McCullagh (1983) mostraram que as funções de quase verossimilhança e seus estimadores de máxima quase verossimilhança (**EMQV**) têm muitas propriedades análogas às da verossimilhança e seus estimadores de máxima verossimilhança (**EMV**). Em particular, o **EMQV** $\widehat{\beta}$ é não-viesado e assintoticamente normal, com média $\beta$. E as matrizes de covariância assintóticas podem ser derivadas de forma usual da matriz de derivadas de segunda ordem de $Q$.

Um dos problemas da quase verossimilhança consiste na comparação de diferentes funções de variância no mesmo conjunto de dados. Nelder & Pregibon (1987) notaram que uma distribuição com determinada função de variância pode existir, mas sem pertencer à classe das distribuições necessárias para um modelo linear generalizado adequado. Então, para avaliar diferentes funções de variância, eles desenvolveram a quase verossimilhança estendida, que é uma generalização natural da quase verossimilhança e permite uma estimação ou modelagem do parâmetro de dispersão $a(\phi)$ ou parâmetros não lineares na variância.

## 2.3 Quase Verossimilhança Estendida

A quase verossimilhança foi estendida por Nelder & Pregibon (1987) para incluir termos da variância, comparar diferentes funções de variância, preditores lineares e funções de ligação, e a possibilidade de modelar a dispersão como uma função de covariáveis, em que essa última abordagem nos interessa no desenvolvimento do modelo proposto. Fazendo $a(\phi) = \phi$, a quase verossimilhança estendida $Q^+$, definida por McCullagh & Nelder (1989), é dada pela expressão:

$$Q^+(\mu_i, \phi; y_i) = -\frac{1}{2\phi}D(y_i, u_i) - \frac{1}{2}\log(\phi), \tag{2.17}$$

em que

$$D(y_i, \mu_i) = -2\phi\{Q(\mu_i, y_i) - Q(y_i, y_i)\} = -2\int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)}dt \tag{2.18}$$

é chamada de função quase desvio do modelo estendido.

Uma família bastante útil é obtida considerando potências de $\mu_i$ (Nelder & Pregibon

(1987)):

$$V_\lambda(\mu_i) = \mu_i^\lambda, \tag{2.19}$$

em que $\lambda$ é conhecido e assume valores positivos como $0, 1, 2, 3$, que correspondem a funções de variância associadas com as distribuições Normal, Poisson, Gama e Normal Inversa, respectivamente. Para a família de funções de variância (2.19), podemos escrever a função quase desvio como:

$$D(y_i;u_i) = \begin{cases} 2\left\{y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i)\right\} & \text{se} \quad \lambda = 1 \\ 2\left\{\frac{y_i}{\mu_i}\log(\frac{y_i}{\mu_i}) - 1\right\} & \text{se} \quad \lambda = 2 \\ \frac{2\left\{y_i^{2-\lambda} - (2-\lambda)y_i\mu_i^{1-\lambda} + (1-\lambda)\mu_i^{2-\lambda}\right\}}{(1-\lambda)(2-\lambda)} & \text{caso contrário.} \end{cases}$$

Essas são algumas formas para a função de variância com suas respectivas funções quase desvio. Neste caso, é possível definir uma função de ligação h, tal que $h(\phi_i) = Z_i\lambda$, em que Z representa a estratificação de variáveis ou covariáveis afetando somente a dispersão. Os estimadores dos parâmetros $\beta$ obtidos pela maximização de $Q^+$ são similares aos obtidos pela maximização de $Q$, que são os **EMQV**. Isso acontece, porque $Q^+$ é uma função linear de $Q$ com coeficientes independentes de $\beta$. O estimador de $\phi$ obtido da maximização de $Q^+$ é $\widehat{\phi} = \sum_{i=1} nD(y_i;\widehat{\mu}_i)/n$, que é o quase-desvio médio.

E, derivando a $Q^+$, com relação aos parâmetros $\beta$ e $\phi$, obtemos os seguintes resultados:

**1.** $\frac{\partial Q^+(\mu_i,\phi,y_i)}{\partial\beta} = [\phi V(\mu_i)]^{-1}(y_i - \mu_i)\left(\frac{\partial\mu_i}{\partial\beta}\right)$.

**2.** $\frac{\partial Q^+(\mu_i,\phi,y_i)}{\partial\phi} = \frac{1}{2\phi^2}D(y_i,\mu_i) - \frac{1}{2\phi}$.

**3.** Se existe uma função de ligação $h^+$ e a covariável $\mathbf{Z}_i$ tais que: $h^+(\phi_i) = \mathbf{Z}_i^T\lambda$. Então, a função quase escore, em relação a $\lambda$, é dada como segue:

$$\frac{\partial Q^+(\mu_i,y_i)}{\partial\lambda} = \frac{D(\mu_i,y_i) - \phi_i}{2\phi_i^2}\left(\frac{\partial\phi_i}{\partial\lambda}\right)$$

Nesse caso, o estimador do parâmetro de dispersão, desenvolvido por Nelder & Pregibon (1987), tem a forma:

$$\widehat{a(\phi)} = \frac{1}{n-p}\sum_i \frac{(Y_i - \widehat{\mu}_i)^2}{V_i(\widehat{\mu}_i)} = \frac{\chi^2}{n-p}. \tag{2.20}$$

A utilização da quase verossimilhança estendida $Q^+$ nos permite a comparação de modelos, nos quais o componente aleatório é especificado somente em relação aos seus dois primeiros momentos. Assim, as técnicas padrão para o ajuste e a comparação de modelos

podem, então, ser aplicadas nessa classe flexível de modelos, que são as quase verossimilhanças estendidas. Daí, o motivo de utilizarmos a $Q^+$ na modelagem proposta por esse trabalho.

## 2.4 Equações de Estimação Generalizadas

Na teoria **GLM**, a suposição de independência entre os indivíduos deve ser satisfeita para que, a partir daí, seja realizado o tratamento dos dados. Neste caso, esses modelos se tornam limitados em estudos georreferenciados que levam em consideração uma estrutura de dependência espacial natural entre indivíduos de um mesmo grupo ou que estejam em localizações mais próximas (vizinhas). Para acomodar a dependência na estrutura de um **GLM**, Zeger & Liang (1986) e Liang & Zeger (1986) desenvolveram as equações de estimação generalizadas (**GEE**). As **GEE**'s nada mais são que uma extensão dos **GLM**'s, com a inclusão de uma estrutura de correlação no processo de estimação. Em uma **GEE**, não é necessário assumir que a distribuição da variável resposta pertença à família exponencial de distribuições, porém basta assumir que a média e a variância estejam caracterizadas como em um **GLM**. A abordagem **GEE** foi aplicada inicialmente no contexto de dados longitudinais e medidas repetidas, mas pode ser considerada também para dados georreferenciados (Monod (2007)).

Usando a mesma notação de Liang & Zeger (1986), considere um vetor resposta $\mathbf{Y_i} = (Y_{i1}, ..., Y_{im_i})^T$, de dimensão $(m_i \times 1)$, e matriz de covariáveis $\mathbf{X}_i = (X_{i1}, ..., X_{im_i})^T$, de dimensão $(m_i \times p)$, para o $i$-ésimo indivíduo, $i = 1, ..., n$, da seguinte forma:

$$\widetilde{\mathbf{V}}_i(\rho) = V_i^{1/2} R_i(\rho) V_i^{1/2} \phi^{-1},$$

em que, $V_i$ é uma matriz diagonal $(m_i \times m_i)$ com elementos da diagonal iguais a $V(\mu_i)$ que é a função de variância definida na quase verossimilhança, $R_i(\rho)$ é uma matriz de correlação simétrica $(m_i \times m_i)$ chamada de matriz de correlação de trabalho e $\rho$ é um escalar e/ou vetor de parâmetros desconhecidos que caracterizam completamente $R_i(\rho)$. Se $R_i(\rho)$ é de fato a verdadeira matriz de correlação para $\mathbf{Y_i}$, então define-se as **GEE**'s, como:

$$\sum_{i=1}^{n} \mathbf{H}_i^T \widetilde{\mathbf{V}}_i^{-1}(\rho)(\mathbf{y}_i - \mu_i) = \mathbf{0}, \tag{2.21}$$

em que $\mathbf{H}_i = \frac{\partial \mu_i}{\partial \beta_j}$ é o vetor de derivadas e $\widetilde{\mathbf{V}}_i(\rho)$ é a matriz de covariâncias para o $i$-ésimo indivíduo. O vetor $\widehat{\beta}$ solução da equação (2.21) é o estimador de $\beta$ obtido através da abordagem **GEE**. Esses estimadores são consistentes, sob fracas suposições, e o modelo é robusto a erros, na especificação da estrutura da matriz de correlação $\mathbf{R}_i(\rho)$. Não é difícil perceber que

a função quase escore dada em (2.14) é um caso particular de uma **GEE** com $\mathbf{R}_i(\rho)$ igual a matriz identidade. Por isso, nós utilizamos neste trabalho a abordagem de Liang & Zeger (1986) sob o cenário espacial, em que os dados serão modelados através de uma mistura de quase-verossimilhanças.

## 2.5 Dependência Espacial em Dados de Contagens

Nesse trabalho assumimos que $\{Y_i \equiv Y(\mathbb{S}_i), \mathbb{S}_i \subset \mathbb{S}, i = 1,2,\ldots,L\}$ é a realização de processo espacial de contagens na área $\mathbb{S}_i$ com $\mathbb{S} \subset \mathscr{R}^2$. Mais especificamente, vamos assumir que $\mathbb{S}$ é um mapa particionado em $n$ áreas $\mathbb{S}_i$, com $y_i$ representando uma contagem observada em $\mathbb{S}_i$. A principal suposição deste tipo de processo é que as contagens observadas em áreas mais próximas tendem a ser mais similares, isto é possuem uma correlação ou dependência espacial maior do que contagens observadas em áreas mais distantes entre si. Formalmente essa dependência pode ser expressa em termos do argumento que a esperança condicional de $Y_i$ dados todos $Y_j$'s, dependem apenas dos $Y_j$'s que ocorrem em áreas vizinhas de $\mathbb{S}_i$ Besag (1974), ou seja,

$$\mathbb{E}(Y_i|Y_j, j \neq i) = \mathbb{E}(Y_i|Y_j, j \sim i), \tag{2.22}$$

em que $j \sim i$ denota o conjunto de todos os vizinhos da área $\mathbb{S}_i$. Uma forma de acomodar a correlação espacial neste tipo de processo, é através da matriz de similaridade ou proximidade espacial, também chamada matriz de vizinhança e denotada por $\mathbf{W}$, em que cada elemento $w_{ij}$ representa uma medida de proximidade entre $\mathbb{S}_i$ e $\mathbb{S}_j$. Esta medida de proximidade pode, por exemplo, ser calculada a partir de um dos seguintes critérios:

- $w_{ij} = 1$, se o centroide de $\mathbb{S}_i$ está a uma determinada distância de $\mathbb{S}_j$ e caso contrário, $w_{ij} = 0$;
- $w_{ij} = 1$ se $\mathbb{S}_i$ e $\mathbb{S}_j$ compartilham fronteiras e $w_{ij} = 0$ caso contrário.

Agora considere que temos um **GLM** espacial e queremos predizer os resíduos, $r_i = Y_i - \mathbb{E}(Y_i)$, dado todos os outros $r_j$'s. Se assumirmos que os $r_i$'s são independentes, então a média global dos resíduos, $\mathbb{E}(r_i) = 0$ é o melhor preditor de $r_i$. No entanto, se queremos utilizar as características locais do processo, devemos assumir que os $r_i$'s vizinhos são similares, de modo que uma média ponderada dos $r_j, j \sim i$, $(\sum_{j\sim i} w_{ij} r_j / w_{i+})$, pode predizer melhor $r_i$, em que $w_{i+} = \sum_{j\sim i} w_{ij}$. Dessa forma, combinando com (2.22) teremos que o preditor de $r_i$ dado todos os outros $r'$s, pode ser visto como uma mistura dos preditores global e local Yasui & Lele (1997). Especificamente,

$$\mathbb{E}(r_i|r_j, j \sim i) = (1-\rho)\mathbb{E}(r_i) + \rho \sum_{j\sim i} \frac{w_{ij}}{w_{i+}} r_j, \tag{2.23}$$

em que, $\rho$ é um parâmetro que representa o grau de dependência espacial e deve ser estimado dos dados. Usando a equação (2.23) podemos escrever que a esperança condicional de $Y_i$ dado todos os outros $Y$'s é,

$$\mathbb{E}(Y_i|Y_j, j \neq i) = \mu_i + \rho \sum_{j \sim i} w_{ij}(Y_j - \mu_j)/w_{i+} \tag{2.24}$$

com a esperança marginal $\mathbb{E}(Y_i) = \mu_i$ e matriz de correlação $\mathbf{R}(\rho) = (\mathbf{I} - \rho\mathbf{MW})^{-1}\mathbf{M}$ onde, $\mathbf{I}$ é a matriz identidade, $\mathbf{M}$ é uma matriz diagonal $n \times n$ com $m_{ii} = 1/w_{i+}$ e $\mathbf{W}$ é uma matriz simétrica $n \times n$ com $w_{ij} = 1$ se $j \sim i$ $w_{ij} = 0$ caso contrário Cressie (1992). No contexto de um **GLM** é mais natural escrevermos,

$$g(\mathbb{E}(Y_i|Y_j, j \neq i)) = g(\mu_i) + \rho \sum_{j \sim i} w_{ij}(g(Y_j) - g(\mu_j))/w_{i+}, \tag{2.25}$$

com $\mathbb{E}(g(Y_i)) \approx g(\mu_i)$. Como $\mathbf{R}^{-1}(\rho) = \mathbf{M}^{-1} - \rho\mathbf{W}$, teremos no contexto de **GEE** (veja, equação (2.21)) que $\widetilde{\mathbf{V}}^{-1}$ é facilmente obtida, facilitando o cálculo dos estimadores do modelo. Por isso, esta estrutura de correlação espacial será adota neste trabalho.

## 2.6 Algoritmo *Expectation-Solution* (ES)

Em um processo de estimação em mistura de modelos é comum utilizar o algoritmo **EM** (veja a Seção 2.1.1). No entanto, o algoritmo **EM** somente pode ser utilizado quando dispomos da função de verossimilhança para o modelo. Por exemplo, na estimação via misturas de quase-verossimilhanças o algoritmo **EM** não pode ser utilizado, pois não conhecemos o modelo probabilístico gerador dos dados. No contexto de medidas repetidas, para explicar a correlação entre as observações repetidas em um mesmo indivíduo, Rosen *et al.* (2000) inseriu as **GEE** no passo **M** do algoritmo **EM**, resultando em uma generalização dele, ao qual chamou de algoritmo *Expectation-Solution* (**ES**). Para ver a definição formal e prova de algumas propriedades assintóticas do algoritmo **ES**, consultar Rosen *et al.* (2000).

O algoritmo **ES**, também, é executado em dois passos. O primeiro é similar ao passo de esperança **E** do algoritmo **EM**. O segundo consiste na substituição do passo de maximização (**M**) no algoritmo **EM** por um passo que requer a solução (**S**) de um sistema de **GEE's** (2.14). No contexto de mistura de GLM's, Rosen et al.(2000) mostraram que se o algoritmo **ES** convergir, ele convergirá para um estimador não-viciado, consistente e assintoticamente Normal, sob suaves condições de regularidade. Sendo assim, utilizaremos esse algoritmo no processo de estimação dos parâmetros da regressão do modelo proposto.

## 2.7 Intervalos de Confiança Bootstrap

A inferência em modelos complexos, é cada vez mais difícil por não se conseguir expressões analiticamente tratáveis ou de fácil cálculo e interpretação na estimação dos parâmetros de interesse. Em nosso caso específico, não é possível garantir a convergência dos estimadores para uma distribuição normal, pois não assumimos qualquer distribuição. Sendo assim, optamos pela inferência via abordagem de intervalos de confiança *Bootstrap* (ver Gentle (2009a)).

**Intervalo de Confiança Bootstrap Percentil**

Dada uma amostra aleatória $(y_1, ..., y_n)$, cuja distribuição $P$ é desconhecida, queremos estimar intervalos de confiança para um parâmetro $\theta$ a partir do estimador pontual $T$. Para isso, podemos utilizar um estimador *Bootstrap* $T^*$ baseado na amostra *Bootstrap* $(y_1^*, ..., y_n^*)$. Se $G_{T^*}(t)$ é a função de distribuição de $T^*$, tal que $G_{T^*}(t^*_{(1-\alpha)}) = 1-\alpha$, em que $t^*_{(1-\alpha)}$ é o limite superior exato do intervalo de confiança $(1-\alpha)$ para o parâmetro $\theta$. Então, o intervalo de confiança *Bootstrap* Percentil (**ICBP**) é configurado da seguinte forma:

$$\left[t^*_{(\frac{\alpha}{2})}; t^*_{(1-\frac{\alpha}{2})}\right] \tag{2.26}$$

em que $t^*_{(\pi)}$ é a $(\pi m)^{\text{th}}$ estatística de ordem de uma amostra de tamanho $m$ de $T^*$. O **ICBP** é um *bootstrap* empírico e pode ser estimado através de simulação de Monte Carlo.

### 2.7.1 Intervalo de Confiança Bootstrap-t

O intervalo de confiança *Bootstrap*-t (**ICB-t**) é um dos intervalos aproximados bastante útil para a estimação intervalar de parâmetros, ele pode geralmente ser construído usando como referência o intervalo de confiança para a média de uma distribuição normal,

$$\left[\bar{Y} - t^*_{(\frac{1-\alpha}{2})}\frac{S}{\sqrt{n}}; \bar{Y} - t^*_{(\frac{\alpha}{2})}\frac{S}{\sqrt{n}}\right],$$

em que $t_{(\pi)}$ é o percentil da distribuição t-*Student*, $\bar{Y}$ é a média amostral e $S^2$ é a variância amostral. Então, um **ICB-t** para qualquer parâmetro construído neste padrão é da forma:

$$\left[T - \widehat{t}_{(\frac{1-\alpha}{2})}\sqrt{\widehat{V}(T)}; T - \widehat{t}_{(\frac{\alpha}{2})}\sqrt{\widehat{V}(T)}\right] \tag{2.27}$$

em que $\widehat{t_{(\pi)}}$ é o percentil estimado da estatística estudentizada:

$$\frac{T^* - T_0}{\sqrt{\widehat{V}(T^*)}}$$

em que $T_0$ é o valor de $T$ calculado a partir da amostra observada.

Para diversos estimadores $T$, não existe uma expressão simples para a variância $\hat{V}(T^*)$. Por isso, podemos estimar a variância utilizando um *bootstrap* e a equação:

$$\widehat{V}(T) = \widehat{V}(T^*) = \frac{1}{m-1}\sum_{j=1}^{m}(T_j^* - \bar{T}^*)^2$$

onde $T_j^*$ é a $j$-ésima observação *bootstrap* de $T$. A vantagem dos **ICB-t** é que eles são mais precisos do que os **ICBP** , porém a desvantagem é que eles são muito mais trabalhosos. Se a distribuição empírica é normal e $T$ é uma medida amostral, o **ICB-t** (2.27) é um intervalo de confiança exato $(1-\alpha)100\%$ de menor tamanho, caso contrário ele pode não ter boas propriedades.

# Capítulo 3

# Modelos de Quase-Verossimilhança Inflacionados de Zeros para Dados de Contagem Espacialmente Dependentes

Como vimos no capítulo anterior (Seção 2.2), modelos de quase verossimilhança são muito úteis na ausência da especificação correta da distribuição dos dados, sendo necessária apenas a definição da relação entre a média e variância dos dados, como em um modelo linear generalizado. Essa abordagem faz-se interessante, no caso de dados que possuam estruturas complexas, como por exemplo apresentando inflação de zeros, sobredispersão e dependência espacial, pois encontrar uma distribuição de probabilidade adequada, que modele conjuntamente essas três características, pode se tornar uma tarefa árdua e talvez até momentaneamente impossível.

Em processos de contagem já existem modelos como **ZIGP**, **ZIDP** e **ZINB**, que são flexíveis para incorporar a inflação de zeros, a sobredispersão e o ajuste por covariáveis. Entretando, tais modelos são limitados por não assumirem dependência ou existência de correlação espacial, fato este que é comum em problemas dessa natureza, pois dados coletados em áreas vizinhas tendem a ser mais similares ou correlacionados do que os obtidos em áreas mais distantes geograficamente. Um exemplo comum desta situação ocorre na área de saúde pública, onde epidemiologistas estudam a variação geográfica dos casos de doenças para gerar e refinar hipóteses testáveis sobre a sua etiologia ou visualizar variáveis que possam estar influenciando no aparecimento da epidemia (Imbiriba *et al.* (2009a); Imbiriba *et al.* (2009b))

Descreveremos, a seguir, o modelo proposto (**QIZ**) quando as observações da variável resposta forem independentes e, posteriormente, quando assumem dependência espacial.

## 3.1 Modelo QIZ para dados com Independência

Primeiramente, vamos considerar a descrição do modelo **ZI** (Capítulo 2, Seção 2.1), com representação estocástica

$$\mathbf{Y}|\mathbf{U} = (\mathbf{1} - \mathbf{U})\mathbf{Z},$$

em que $\mathbf{Y} = (Y_1, ..., Y_L)^T$ é o vetor resposta de observações de contagem independentes, $\mathbf{U} = (U_1, ..., U_L)^T$ é o componente da mistura não observável e $\mathbf{Z} = (Z_1, ..., Z_L)^T$ é o componente observável, com representação estocástica (2.3), em que $\mathbf{U}$ e $\mathbf{Z}$ são independentes. E ainda, que $Y_i$ segue uma distribuição inflacionada de zeros (2.1), ou seja,

$$Y_i \sim p_i \, \mathbb{I}_{\{y_i=0\}} + (1 - p_i)\mathbb{P}(\theta),$$

em que $\mathbb{I}$ é a função indicadora, $\mathbb{P}(\theta)$ é uma distribuição de contagem, $p_i$ é a probabilidade de $Y_i \sim 0$ e $(1 - p_i)$ é a probabilidade de $Y_i \sim \mathbb{P}(\theta)$. Assim, podemos escrever a probabilidade marginal de $Y_i$ (2.2) como sendo

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i)P_\theta(Y_i = y_i), & \text{se } y_i = 0, \\ (1 - p_i)P_\theta(Y_i = y_i), & \text{se } y_i > 0, \end{cases}$$

ou ainda,

$$\begin{aligned} P(Y_i = y_i) &= p_i \mathbb{I}_{\{y_i=0\}} + (1 - p_i)P_\theta(Y_i = y_i) \\ &= p_i \mathbb{I}_{\{y_i=0\}} + (1 - p_i)e^{\log P_\theta(Y_i=y_i)} \end{aligned}$$

cuja probabilidade $P(Y_i = y_i)$ depende de $p_i$ e da probabilidade $P_\theta(Y_i = y_i)$, que por sua vez vem da distribuição de contagem $\mathbb{P}(\theta)$.

Então, a verossimilhança do modelo **ZI** é escrita como

$$\begin{aligned} l(\theta) &= \prod_{i=1}^{L} P(Y_i = y_i) \\ &= \prod_{i=1}^{L} [p_i \mathbb{I}_{\{y_i=0\}} + (1 - p_i)e^{\log P_\theta(Y_i=y_i)}] \end{aligned} \tag{3.1}$$

Uma das principais ideias deste trabalho foi a de substituir o $\log P_\theta(Y_i = y_i)$, em (3.1), pela quase-verossimilhança estendida $Q^+$ (2.17), obtendo a aproximação

$$l(\theta) \approx \prod_{i=1}^{L} [p_i \mathbb{I}_{\{y_i=0\}} + (1 - p_i)e^{Q^+(y_i;\theta_i)}] = Q(\mathbf{p}, \theta),$$

em que $Q(\mathbf{p}, \theta)$ é a quase-verossimilhança.

A partir destas especificações, definiremos primeiramente o modelo **QIZ** sob a suposição de independência.

A construção do modelo **QIZ** foi baseada na log-verossimilhança completa do modelo **ZI** (2.6), dada no Capítulo anterior da seguinte forma:

$$\begin{aligned} l^c(p_i,\theta_i;y_i,u_i) &= [u_i \log p_i + (1-u_i)\log(1-p_i)] + (1-u_i)\log P_{\theta_i}(Y_i=y_i) \\ &= l^c(p_i;u_i) + l^c(\theta_i;y_i,u_i), \end{aligned}$$

Assim, o modelo **QIZ** é representado por uma quase verossimilhança inflacionada de zeros completa, considerando $\theta_i = (\mu_i, \phi_i)$, dada por:

$$Q^c(p_i,\mu_i,\phi_i;y_i,u_i) = u_i \log(p_i) + (1-u_i)\{\log(1-p_i) + Q^+(\mu_i,\phi_i;y_i)\}, \tag{3.2}$$

em que substituímos a $l^c(\theta_i,y_i,u_i)$ em (2.6) pela quase verossimilhança estendida $Q^+$ (2.17), a partir da qual poderemos modelar o parâmetro de dispersão como função de covariáveis, obtendo a quase verossimilhança inflacionada de zeros (3.2). Marginalmente, $Q^c$ (3.2) é a quase verossimilhança estendida para o modelo inflacionado de zeros. Através do modelo **ZI** em (2.4), definimos a esperança e variância do modelo **QIZ**, representadas como:

$$\mathbb{E}(Y_i) = (1-p_i)\mu_i \quad \text{e} \quad Var(Y_i) = (1-p_i)\phi_i V(\mu_i) + p_i(1-p_i)\mu_i^2. \tag{3.3}$$

em que $\mu_i$ é a esperança e $\phi_i V(\mu_i)$ é a variância, ambas do componente da mistura referente às contagens positivas, e $p_i$ é a probabilidade de $Y_i$ vir do excesso de zeros. E, para a diferenciação da $Q^c$ (3.2) com relação aos parâmetros de regressão que modelam a média $\beta$, a dispersão $\lambda$ e o excesso de zeros $\gamma$, são válidas as seguintes expressões:

(i) $\frac{\partial Q^c}{\partial \gamma} = \frac{u_i - p_i}{\tilde{V}(p_i)}\left(\frac{\partial p_i}{\partial \gamma}\right)$, reescrevendo na forma matricial:
$\frac{\partial Q^c}{\partial \gamma} = \sum_{i=1}^{L}\left\{\frac{\partial p_i}{\partial \gamma}\right\}^T [\mathbf{A}_i^{1/2}\mathbf{I}\mathbf{A}_i^{1/2}]^{-1}(u_i^{(k)} - p_i)$,
com $\mathbf{A}_i = diag\{p_1(1-p_1),...,p_L(1-p_L)\}$;

(ii) $\frac{\partial Q^c}{\partial \beta} = (1-u_i)\frac{y_i - \mu_i}{\phi_i V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \beta}\right)$, ou na forma matricial:
$\frac{\partial Q^c}{\partial \beta} = \sum_{i=1}^{L}\left\{\frac{\partial \mu_i}{\partial \beta}\right\}^T [\mathbf{B}_i^{1/2}\mathbf{I}\mathbf{B}_i^{1/2}]^{-1}\mathbf{U}^{(k)}(y_i - \mu_i)$,
com $\mathbf{B}_i = diag\{\phi_1 V(\mu_1),...,\phi_L V(\mu_L)\}$ e $\mathbf{U}^{(k)} = diag\{(1-u_1^{(k)}),...,(1-u_L^{(k)})\}$;

(iii) $\frac{\partial Q^c}{\partial \lambda} = (1-u_i)\frac{D(y_i,\mu_i) - \phi_i}{2\phi_i^2}\left(\frac{\partial \phi_i}{\partial \lambda}\right)$, e também na forma matricial:
$\frac{\partial Q^c}{\partial \lambda} = \sum_{i=1}^{L}\left\{\frac{\partial \phi_i}{\partial \lambda}\right\}^T [\mathbf{C}_i^{1/2}\mathbf{I}\mathbf{C}_i^{1/2}]^{-1}\mathbf{U}^{(k)}(D(y_i,\mu_i) - \phi_i)$,
com $\mathbf{C}_i = diag\{2\phi_1^2,...,2\phi_L^2\}$ e $\mathbf{U}^{(k)} = diag\{(1-u_1^{(k)}),...,(1-u_L^{(k)})\}$;

(iv) $\mathbb{E}_u\left(\frac{\partial Q^c}{\partial \gamma}\right) = \mathbb{E}_y\left(\frac{\partial Q^c}{\partial \beta}\right) = \mathbb{E}_y\left(\frac{\partial Q^c}{\partial \lambda}\right) = 0.$

As expressões (i), (ii) e (iii) são chamadas de funções quase escore, similar aquelas definidas no Capítulo 2 (equação 2.14) para seus respectivos parâmetros. A expressão (iv) é o valor esperado de cada função quase escore, que de forma similar às propriedades da quase verossimilhança definidas por Wedderburn (1974), são iguais a zero, pois tem a forma de uma **GEE**.

A função quase desvio é dada de forma similar a descrita na seção sobre quase verossimilhança (2.13), no Capítulo 2. Agora, vejamos exemplos de modelos, com suas respectivas funções quase desvio, quase verossimilhança e outras.

### 3.1.1 Modelo Quase-Poisson Inflacionado de Zeros

O primeiro modelo que iremos exemplificar é o de Quase-Poisson Inflacionado de Zeros (**QPIZ**). Nesse modelo, para calcular a função quase desvio (2.13), considera-se $V(t) = t$. A partir daí, podemos então calcular a função quase desvio da seguinte forma:

$$\begin{aligned} D(y_i;\mu_i) &= -2\int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)}dt = -2\int_{y_i}^{\mu_i} \frac{y_i - t}{t} = -2\Big[\int_{y_i}^{\mu_i} \frac{y_i}{t}dt - \int_{y_i}^{\mu_i} dt\Big] = -2\Big[y_i \ln t - t\Big]_{y_i}^{\mu_i} \\ &= -2\Big[y_i \ln\Big(\frac{\mu_i}{y_i}\Big) - (\mu_i - y_i)\Big] = 2\Big[y_i \ln\Big(\frac{y_i}{\mu_i}\Big) - (y_i - \mu_i)\Big]. \end{aligned} \tag{3.4}$$

Neste caso, a função de quase verossimilhança estendida é dada por:

$$\begin{aligned} Q^+(\mu_i,\phi_i;y_i) &= -\frac{1}{2}D(y_i;\mu_i) - \frac{1}{2}\log(\phi_i) = -\frac{1}{2}\Big\{2\Big[y_i \ln\Big(\frac{y_i}{\mu_i}\Big) - (y_i - \mu_i)\Big]\Big\} - \frac{1}{2}\log(\phi_i) \\ &= y_i\Big[1 - \ln\Big(\frac{y_i}{\mu_i}\Big)\Big] - \mu_i - \frac{1}{2}\log(\phi_i) \end{aligned} \tag{3.5}$$

Então, se $y_i > 0$, a função de quase verossimilhança completa será da forma:

$$\begin{aligned} Q^c(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i) + (1-u_i)\Big\{\log(1-p_i) + Q^+(\mu_i,\phi_i;y_i)\Big\} \\ &= u_i\log(p_i) + (1-u_i)\Big\{\log(1-p_i) + y_i\Big[1 - \ln\Big(\frac{y_i}{\mu_i}\Big)\Big] \\ &\quad -\mu_i - \frac{1}{2}\log(\phi_i)\Big\}, \end{aligned} \tag{3.6}$$

e se $y_i = 0$, em que assumimos $y_i * a = 0$, com $a$ assumindo qualquer valor, temos a

$$\begin{aligned} Q^c(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i) + (1-u_i)\Big\{\log(1-p_i) + Q^+(\mu_i,\phi_i;y_i)\Big\} \\ &= u_i\log(p_i) + (1-u_i)\Big\{\log(1-p_i) - \mu_i - \frac{1}{2}\log(\phi_i)\Big\}. \end{aligned} \tag{3.7}$$

A esperança condicional de $\mathbb{E}(u_i|y_i;p_i,\mu_i,\phi_i) \equiv \widetilde{u}_i$ será descrita como:

$$\begin{aligned}\widetilde{u}_i &= \Big(1+\exp\Big\{-logit(p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\}\Big)^{-1}\mathbb{I}_{\{y_i=0\}}\\ &= \Big(1+\exp\Big\{-logit(p_i)-\mu_i-\frac{1}{2}\log(\phi_i)\Big\}\Big)^{-1},\end{aligned}$$

### 3.1.2 Modelo Quase-Binomial Inflacionado de Zeros

Outro modelo, por exemplo, é o de Quase-Binomial Inflacionado de Zeros (**QBIZ**), no qual considera-se $V(t)=\frac{t(n_i-t)}{n_i}$, equivalente a $\mu_i=n_i\pi_i$ da distribuição Binomial $(n_i,\pi_i)$, a função quase-desvio é dada por:

$$\begin{aligned}D(y_i;\mu_i) &= -2\int_{y_i}^{\mu_i}\frac{y_i-t}{V(t)}dt=-2\int_{y_i}^{\mu_i}\frac{y_i-t}{\frac{t(n_i-t)}{n_i}}\\ &= -2n_i\Big[\int_{y_i}^{\mu_i}\frac{y_i}{t(n_i-t)}dt-\int_{y_i}^{\mu_i}\frac{t}{t(n_i-t)}dt\Big]\\ &= 2\Big[y_i\ln\Big(\frac{y_i}{\mu_i}\Big)+(n_i-y_i)\ln\Big(\frac{n_i-y_i}{n_i-\mu_i}\Big)\Big] \qquad (3.8)\end{aligned}$$

Logo, a função de quase verossimilhança estendida desse modelo, pode ser escrita como:

$$\begin{aligned}Q^{+}(p_i,\mu_i,\phi_i;y_i,u_i) &= -\frac{1}{2}D(y_i;\mu_i)-\frac{1}{2}\log(\phi_i)\\ &= -\frac{1}{2}\Big\{2\Big[y_i\ln\Big(\frac{y_i}{\mu_i}\Big)+(n_i-y_i)\ln\Big(\frac{n_i-y_i}{n_i-\mu_i}\Big)\Big]\Big\}-\frac{1}{2}\log(\phi_i)\\ &= (y_i-n_i)\ln\Big(\frac{n_i-y_i}{n_i-\mu_i}\Big)-y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-\frac{1}{2}\log(\phi_i) \qquad (3.9)\end{aligned}$$

Então, se $y_i>0$, a função de quase verossimilhança completa, para esse modelo, será da forma:

$$\begin{aligned}Q^{c}(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\}\\ &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+(y_i-n_i)\ln\Big(\frac{n_i-y_i}{n_i-\mu_i}\Big)\\ &\quad -y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}, \qquad (3.10)\end{aligned}$$

e se $y_i=0$, então a

$$\begin{aligned}Q^{c}(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\}\\ &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)-n_i\ln\Big(\frac{n_i}{n_i-\mu_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}, \qquad (3.11)\end{aligned}$$

A esperança condicional de $\mathbb{E}(u_i|y_i;p_i,\mu_i,\phi_i) \equiv \widetilde{u}_i$ pode ser expressa como:

$$\begin{aligned}\widetilde{u}_i &= \Big(1+\exp\Big\{-logit(p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\}\Big)^{-1}\mathbb{I}_{\{y_i=0\}} \\ &= \Big(1+\exp\Big\{-logit(p_i)-n_i\ln\Big(\frac{n_i}{n_i-\mu_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}\Big)^{-1},\end{aligned}$$

### 3.1.3 Modelo Quase-Binomial Negativo Inflacionado de Zeros

No modelo Quase-Binomial Negativo Inflacionado de Zeros (**QBNIZ**), consideramos $V(t)=\frac{t}{r_i}(t+r_i)$, de forma equivalente a $\mu_i=\frac{r_i}{\pi_i}$ da distribuição Binomial Negativa $(r_i,\pi_i)$, onde a função quase-desvio é dada por:

$$\begin{aligned}D(y_i;\mu_i) &= -2\int_{y_i}^{\mu_i}\frac{y_i-t}{V(t)}dt=-2\int_{y_i}^{\mu_i}\frac{y_i-t}{\frac{t}{r_i}(t+r_i)} \\ &= -2r_i\Big[\int_{y_i}^{\mu_i}\frac{y_i}{t+r_i}dt-\int_{y_i}^{\mu_i}\frac{t}{t+r_i}dt\Big] \\ &= 2\Big[y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-(y_i+r_i)\ln\Big(\frac{y_i+r_i}{\mu_i+r_i}\Big)\Big] \qquad (3.12)\end{aligned}$$

Então, a função de quase verossimilhança estendida para esse modelo, é escrita da forma:

$$\begin{aligned}Q^{+}(\mu_i,\phi_i;y_i) &= -\frac{1}{2}D(y_i;\mu_i)-\frac{1}{2}\log(\phi_i) \\ &= -\frac{1}{2}\Big\{2\Big[y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-(y_i+r_i)\ln\Big(\frac{y_i+r_i}{\mu_i+r_i}\Big)\Big]\Big\}-\frac{1}{2}\log(\phi_i) \\ &= (y_i+r_i)\ln\Big(\frac{y_i+r_i}{\mu_i+r_i}\Big)-y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-\frac{1}{2}\log(\phi_i) \qquad (3.13)\end{aligned}$$

Assim, se $y_i>0$, a função de quase verossimilhança completa, será descrita como:

$$\begin{aligned}Q^{c}(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\} \\ &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+(y_i+r_i)\ln\Big(\frac{y_i+r_i}{\mu_i+r_i}\Big) \\ &\quad -y_i\ln\Big(\frac{y_i}{\mu_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}, \qquad (3.14)\end{aligned}$$

e se $y_i=0$, então a

$$\begin{aligned}Q^{c}(p_i,\mu_i,\phi_i;y_i,u_i) &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\} \\ &= u_i\log(p_i)+(1-u_i)\Big\{\log(1-p_i)+r_i\ln\Big(\frac{r_i}{\mu_i+r_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}, \\ & \qquad (3.15)\end{aligned}$$

A esperança condicional de $\mathbb{E}(u_i|y_i;p_i,\mu_i,\phi_i) \equiv \widetilde{u}_i$ pode ser descrita como:

$$\begin{aligned}\widetilde{u}_i &= \Big(1+\exp\Big\{-logit(p_i)+Q^{+}(\mu_i,\phi_i;y_i)\Big\}\Big)^{-1}\mathbb{I}_{\{y_i=0\}} \\ &= \Big(1+\exp\Big\{-logit(p_i)+r_i\ln\Big(\frac{r_i}{\mu_i+r_i}\Big)-\frac{1}{2}\log(\phi_i)\Big\}\Big)^{-1},\end{aligned}$$

## 3.2 Modelo QIZ para dados com Dependência Espacial

Para incorporar a dependência espacial, utilizamos no processo de estimação, do modelo **QIZ** com independência, as equações de estimação generalizadas (**GEE**), propostas por Liang & Zeger (1986), que construíram funções de estimação para os parâmetros de interesse na ausência da verossimilhança totalmente especificada e presença de correlação (Zeger & Liang (1986)), que é exatamente o nosso caso. Definimos o modelo de quase verossimilhança inflacionado de zeros com dependência espacial (**QIZDE**), como segue.

**Definição 2.** *Suponha que* $\{Y_i \equiv Y(\mathbb{S}_i), \mathbb{S}_i \subset \mathbb{S}, i = 1,2,\ldots,L)\}$ *é a realização de processo espacial de contagens na área* $\mathbb{S}_i$ *com* $\mathbb{S} \subset \mathscr{R}^2$*. Mais especificamente, vamos assumir que* $\mathbb{S}$ *é um mapa particionado em L áreas* $\mathbb{S}_i$ *com* $y_i$ *representando uma contagem observada em* $\mathbb{S}_i$ *com excesso de zeros. Assuma que* $Y_i$ *pertence à classe de modelos com quase verossimilhança completa (dada na Seção 2.2), com vetor de parâmetros* $\vartheta_i = (\mu_i, \phi_i, p_i)$*, dada por:*

$$Q^c(\vartheta_i;y_i,u_i) = u_i\log(p_i)+(1-u_i)\{\log(1-p_i)+Q^{+}(\mu_i,\phi_i;y_i)\},$$

*então, para acomodar a dependência espacial, no processo de estimação substitui-se a matriz identidade* **I***, nas funções quase escore, por uma matriz simétrica de covariância espacial* $\mathbf{R}(\rho)$*, com* $\rho$ *representando o parâmetro que mede o grau de dependência espacial. A essa nova estrutura chamamos modelo de quase verossimilhança inflacionado de zeros com dependência espacial* ***QIZDE****.*

Sendo assim, partimos da definição do modelo **QIZDE**, em que substituímos a matriz identidade **I**, nas funções quase escore definidas em (i), (ii) e (iii), pela matriz de covariância espacial $\mathbf{R}(\rho)$, resultando nas seguintes funções quase escore para cada parâmetro de

regressão, respectivamente, dadas por:

$$\frac{\partial Q^c}{\partial \gamma} = \sum_{i=1}^{L} \left\{ \frac{\partial p_i}{\partial \gamma} \right\}^T [\mathbf{A}_i^{1/2} \mathbf{R}_\gamma(\rho) \mathbf{A}_i^{1/2}]^{-1} (u_i - p_i) = 0, \tag{3.16}$$

$$\frac{\partial Q^c}{\partial \beta} = \sum_{i=1}^{L} \left\{ \frac{\partial \mu_i}{\partial \beta} \right\}^T [\mathbf{B}_i^{1/2} \mathbf{R}_\beta(\rho) \mathbf{B}_i^{1/2}]^{-1} \mathbf{U} (y_i - \mu_i) = 0, \tag{3.17}$$

$$\frac{\partial Q^c}{\partial \lambda} = \sum_{i=1}^{L} \left\{ \frac{\partial \phi_i}{\partial \lambda} \right\}^T [\mathbf{C}_i^{1/2} \mathbf{R}_\lambda(\rho) \mathbf{C}_i^{1/2}]^{-1} \mathbf{U} (D(y_i, \mu_i) - \phi_i) = 0, \tag{3.18}$$

em que as matrizes de covariância espacial $\mathbf{R}_\gamma(\rho), \mathbf{R}_\beta(\rho)$ e $\mathbf{R}_\lambda(\rho)$ representam a estrutura de dependência espacial na estimação dos parâmetros de regressão $\gamma$, $\beta$ e $\lambda$, respectivamente, e $\mathbf{U} = diag\{(1-u_1), ..., (1-u_L)\}$ é uma matriz de valores não observados.

Como havíamos comentado anteriormente, essas funções quase escore têm uma estrutura de equação de estimação generalizada. Por exemplo, no caso **QIZDE** Poisson:

a) Para o parâmetro de regressão $\gamma$, cuja função de ligação é $logit(p_i) = G_i^T \gamma$, a função quase escore é dada por:

$$\frac{\partial Q^c}{\partial \gamma} = \sum_{i=1}^{L} \left\{ G_i^T p_i (1-p_i) \right\}^T [\mathbf{A}_i^{1/2} \mathbf{R}_\gamma(\rho) \mathbf{A}_i^{1/2}]^{-1} (u_i - p_i) = 0; \tag{3.19}$$

b) Já para o parâmetro de regressão $\beta$, cuja função de ligação é $\log(\mu_i) = X_i^T \beta$, a função quase escore será:

$$\frac{\partial Q^c}{\partial \beta} = \sum_{i=1}^{L} \left\{ X_i^T \mu_i \right\}^T [\mathbf{B}_i^{1/2} \mathbf{R}_\beta(\rho) \mathbf{B}_i^{1/2}]^{-1} \mathbf{U} (y_i - \mu_i) = 0; \tag{3.20}$$

c) Finalmente, para o parâmetro de regressão $\lambda$, cuja função de ligação é $\log(\phi_i) = Z_i^T \lambda$, a função quase escore será:

$$\frac{\partial Q^c}{\partial \lambda} = \sum_{i=1}^{L} \left\{ Z_i^T \phi_i \right\}^T [\mathbf{C}_i^{1/2} \mathbf{R}_\lambda(\rho) \mathbf{C}_i^{1/2}]^{-1} \mathbf{U} (D(y_i, \mu_i) - \phi_i) = 0. \tag{3.21}$$

### 3.2.1 Estimação dos parâmetros no modelo QIZDE

Por simplicidade na estimação e sem perda de generalidade, assumimos que as matrizes de dependência espacial que modelam a inflação de zeros e a sobredispersão são iguais a matriz identidade, ou seja, $\mathbf{R}_\gamma(\rho) = \mathbf{R}_\lambda(\rho) = \mathbf{I}$. E que a matriz de dependência espacial que modela a média das contagens positivas $\mathbf{R}_\beta(\rho)$ é uma representação de um processo condicional autorregressivo (**CAR**), veja a Seção 2.5, especificada da seguinte forma:

$$\mathbf{R}_\beta(\rho) = (I - \rho \mathbf{M}\mathbf{W})^{-1} \mathbf{M} \tag{3.22}$$

em que $\mathbf{W}$ é uma matriz $(L \times L)$ simétrica, que define a estrutura espacial de vizinhança, com elementos definidos por pesos adjacentes conhecidos, representados por $w_{ij}$ se $i \sim j$ (lê-se: $i$ e $j$ são áreas vizinhas) e 0 caso contrário, $\mathbf{M}$ é uma matriz diagonal $(L \times L)$ com $i$-ésimo elemento na diagonal igual a $w_{i+} = \sum_{j \sim i} w_{ij}$, $\rho$ é o parâmetro que mede o grau de dependência espacial e $\mathbf{R}_\beta^{-1}(\rho) = \mathbf{M}^{-1} - \rho\mathbf{W}$, se $\rho \in (\rho_{min}, \rho_{max})$, caso contrário $\mathbf{R}_\beta^{-1}(\rho)$ é singular, com $\rho_{min} = \imath_1^{-1}$ e $\rho_{max} = \imath_n^{-1}$, tal que $\imath_1 < 0 < \imath_n$, $\imath_1$ e $\imath_n$ são, respectivamente, o menor e maior autovalor da matriz $\mathbf{M}^{-1/2}\mathbf{M}\mathbf{W}\mathbf{M}^{1/2}$ (ver Lawson *et al.* (1999), pg. 66).

Nessa estrutura de covariância, representada através de um modelo **CAR**, podemos escrever a esperança condicional da variável resposta $Y$ de uma área $i$ dada outra $j$, sendo $i \neq j$, como:

$$\mathbb{E}[Y_i | Y_j, i \neq j] = \mu_i(\beta) + \rho \sum_{j \sim i} \frac{w_{ij}}{w_{i+}} (Y_j - \mu_j(\beta)) \tag{3.23}$$

em que o termo $\frac{w_{ij}}{w_{i+}}$ é resultado da matriz de correlação espacial $\mathbf{R}_\beta(\rho)$, descrita em (3.22).

Então, o valor esperado marginal de $Y_i$ é calculado a partir da esperança de (3.23), obtendo a expressão:

$$\mathbb{E}(Y_i) = \mu_i(\beta) \tag{3.24}$$

E, ainda, o valor esperado da função quase desvio é aproximado pelo valor do parâmetro de sobredispersão $\phi_i$, ou seja,

$$\mathbb{E}[D(y_i, \mu_i)] \approx \phi_i \tag{3.25}$$

Os modelos **CAR** podem assumir que a resposta é uma função de ambas as variáveis explicativas e os valores da resposta em locais vizinhos (equação 3.23) .

### Estimação de $\beta$

Suponha que a sobredispersão é a mesma em todas as áreas, ou seja, $\phi_l = \phi, l = 1, 2, ..., L$. Então, a estimação do vetor de parâmetros de regressão $\beta$ é realizada via algoritmo **ES**, em dois passos.

**Passo E**: defina $\mathbf{Q}^c(\vartheta; \mathbf{y}, \mathbf{u}) = \sum_{i=1}^{L} Q^c(\vartheta_i; y_i, u_i)$. Então, na $k$-ésima iteração, inicializamos o processo com $\vartheta^{(k)} = (\gamma^{(k)}, \beta^{(k)}, \rho^{(k)}, \phi^{(k)})$ e calculamos a $E_{\mathbf{u}|\vartheta^{(k)};\mathbf{y}}[\mathbf{Q}^c(\vartheta; \mathbf{y}, \mathbf{u})]$. Como a $\mathbf{Q}^c(\vartheta; \mathbf{y}, \mathbf{u})$ é linear em $\mathbf{u}$, esta quantidade é dada por $\mathbf{Q}^c(\vartheta; \mathbf{y}, \widetilde{\mathbf{u}}^{(k)})$, com i-ésimo elemento dado por:

$$\widetilde{u}_i^{(k)} = \left(1 + \exp\{-logit(p_i(\gamma^{(k)})) + Q^+(\mu_i(\beta^{(k)}), \phi_i(\lambda^{(k)}), y_i)\}\right)^{-1} \mathbb{I}_{\{y_i = 0\}},$$

em que $Q^+(\cdot)$ é a função de quase verossimilhança estendida, especificada de acordo com os modelos da Seção 3.1.

**Passo S**: definimos $\widetilde{V}^{-1}(\rho^{(k)},\beta^{(k)}) = [\mathbf{B}^{1/2}\mathbf{R}_\beta(\rho^{(k)})\mathbf{B}^{1/2}]^{-1}\mathbf{U}^{(k)}$, em que a matriz $\mathbf{B} = diag(V(\mu_1^{(k)}),...,V(\mu_L^{(k)}))$, e estimamos o vetor de parâmetros de regressão $\beta$ na $k+1$-ésima iteração solucionando iterativamente a **GEE** (3.17), ou seja, calculando a expressão abaixo, que é o estimador para $\beta$, dado por

$$\beta^{(k+1)} = \beta^{(k)} + \left(\mathbf{H}^T\tilde{V}^{-1}(\rho^{(k)},\beta^{(k)})\mathbf{H}\right)^{-1}\mathbf{H}^T\tilde{V}^{-1}(\rho^{(k)},\beta^{(k)})(\mathbf{y}-\mu(\beta^{(k)})), \tag{3.26}$$

em que $\mathbf{H}$ é a matriz de derivadas parciais $[\frac{\partial\mu}{\partial\beta}]^T$, sendo estas matrizes avaliadas em $\beta^{(k)}$.

**Estimação de $\gamma$**

A estimação de $\gamma$ é realizada solucionando (3.16) de maneira similar a um **GLM($p$)**, com $\mathbf{I} = \mathbf{R}_\gamma(\rho)$. Nesse caso teremos

$$\gamma^{(k+1)} = (\mathbf{G}^T\mathbf{A}^{-1(k)}\mathbf{G})^{-1}\mathbf{G}\mathbf{A}^{-1(k)}\mathbf{h}^{(k)}, \tag{3.27}$$

com $\mathbf{h}^{(k)} = \mathbf{G}\gamma^{(k)} + \Delta(\mathbf{u}^{(k)} - p^{(k)})$ e $\Delta = diag(g'(p_1),...,g'(p_L))$ avaliado em $p^{(k)} = p(\gamma^{(k)})$.

**Estimação de $\rho$**

Para estimar $\rho$ utilizamos um procedimento via mínimos quadrados condicionais (Klimko & Nelson (1978)) minimizando a soma de quadrados dos desvios em relação a função de ligação da esperança condicional dada em (2.25), somente com respeito aos $y_i$ pertencentes ao conjunto

$$\widetilde{y} = \{y_i : i = 1,2,\ldots,L\}\backslash\{y_i : y_i = 0, p_i > 0{,}05\}, \tag{3.28}$$

pois nesse caso específico a dependência espacial está sendo acomodada somente na distribuição de contagem e consideramos que os zeros pertencentes ao conjunto $\{y_i : y_i = 0, p_i > 0{,}05\}$ são considerados como estruturais. Na prática, os valores de $p_i$ são desconhecidos, sendo substituídos pelos seus valores estimados com os dados observados. De modo que, dado $\beta$ minimizamos, com relação a $\rho$, a seguinte expressão:

$$\sum_{y_i\in\widetilde{y}}\left\{g(y_i) - g(\mu_i(\beta)) + \rho\sum_{j\sim i}\frac{w_{ij}}{w_{i+}}[g(y_j) - g(\mu_j(\beta))]\right\}^2,$$

obtendo a expressão

$$\sum_{y_i \in \widetilde{y}} \left\{ g(y_i) - g(\mu_i(\beta)) + \rho \sum_{j \sim l} \frac{w_{ij}}{w_{i+}} [g(y_j) - g(\mu_j(\beta))] \right\} \times \left\{ \sum_{j \sim i} \frac{w_{ij}}{w_{i+}} [g(y_j) - g(\mu_j(\beta))] \right\}. \tag{3.29}$$

Então, na $(k+1)-$ésima iteração, dado $\beta^{(k+1)}$, o valor de $\rho$ é atualizado em $(k+1)$ através do estimador

$$\rho^{(k+1)} = \frac{\sum_{y \in \tilde{y}} \left\{ \sum_{j \sim i} \frac{w_{ij}}{w_{i+}} [g(y_i) - g(\mu_i(\beta^{(k+1)}))][g(y_j) - g(\mu_j(\beta^{(k+1)}))] \right\}}{\sum_{y \in \tilde{y}} \left\{ \sum_{j \sim i} \frac{w_{ij}}{w_{i+}} [g(y_j) - g(\mu_j(\beta^{(k+1)}))] \right\}^2}, \tag{3.30}$$

em que $\rho^{(k+1)} \in (\rho_{min}, \rho_{max})$, $\rho_{min}$ e $\rho_{min}$ são o inverso do menor e maior autovalores, respectivamente, da matriz $M^{1/2}MWM^{1/2}$ (Lawson *et al.* (1999), pg. 66).

Se não existe a inflação de zeros, $g$ é a função identidade ($g(a) = a$) e $w_{ij}$ é do tipo 0 ou 1. Não é difícil verificar que $\rho \in (-1,1)$ é um índice de Moran baseado nos resíduos do modelo (ver Druck *et al.* (2004), cap. 5), estimado por

$$\widehat{\rho} = \frac{\sum_{i=1}^{L} \frac{1}{w_{i+}} \left\{ \sum_{j \sim i} (y_i - \mu_i(\widehat{\beta}))(y_j - \mu_j(\widehat{\beta})) \right\}}{\sum_{i=1}^{L} \frac{1}{w_{i+}^2} \left\{ \sum_{j \sim i} (y_j - \mu_j(\widehat{\beta})) \right\}^2} \tag{3.31}$$

**Estimação de $\phi$**

Na estimação de $\phi$, solucionamos iterativamente a **GEE** (3.18) condicionada ao valor de $\beta^{(k+1)}$, com $\phi_l = \phi, l = 1,2,...,L$ e obtemos na $(k+1)$-ésima iteração,

$$\phi^{(k+1)} = \frac{\sum_{l=1}^{L} (1 - u_l^{(k)}) D(y_l, \mu_l(\beta^{(k+1)}))}{\sum_{l=1}^{L} (1 - u_l^{(k)})}, \tag{3.32}$$

em que $D$ é a função quase desvio avaliada em $\beta^{(k+1)}$.

As estimativas dos parâmetros de regressão $\beta$ e $\gamma$, do parâmetro que mede a sobredispersão $\phi$ e do parâmetro de dependência espacial $\rho$, são obtidas solucionando iterativamente as quatro equações (3.26), (3.27), (3.30) e (3.32), até obter-se a convergência. Em cada iteração, $\widehat{\beta}$ é calculado dado $\widehat{\phi}$ e $\widehat{\rho}$, acima na expressão (3.31), em seguida $\widehat{\phi}$ e $\widehat{\rho}$ são obtidos a partir do $\hat{\beta}$ calculado e $\gamma$ é estimado a partir de $p$.

### 3.2.2 Estimação Geral dos parâmetros no modelo QIZDE

A partir das funções quase escore (3.16), (3.17) e (3.18), considerando que cada matriz de correlação espacial é similar a definida como no modelo **CAR**, na expressão (3.22), obtemos os estimadores para os parâmetros de regressão do modelo proposto, de modo geral, no formato matricial. Assim, no passo **S** do algoritmo **ES**, teremos os seguintes estimadores:

(i) para $\gamma$:

$$\gamma^{(k+1)} = \gamma^{(k)} + \left(\mathbf{H}_\gamma^T \tilde{V_\gamma}^{-1}(\rho^{(k)}, \gamma^{(k)})\mathbf{H}_\gamma\right)^{-1} \mathbf{H}_\gamma^T \tilde{V_\gamma}^{-1}(\rho^{(k)}, \gamma^{(k)})(\mathbf{u} - \mathbf{p}(\gamma^{(k)})); \quad (3.33)$$

(ii) para $\beta$:

$$\beta^{(k+1)} = \beta^{(k)} + \left(\mathbf{H}_\beta^T \tilde{V_\beta}^{-1}(\rho^{(k)}, \beta^{(k)})\mathbf{H}_\beta\right)^{-1} \mathbf{H}_\beta^T \tilde{V_\beta}^{-1}(\rho^{(k)}, \beta^{(k)})(\mathbf{y} - \mu(\beta^{(k)})); \quad (3.34)$$

(iii) para $\lambda$:

$$\lambda^{(k+1)} = \lambda^{(k)} + \left(\mathbf{H}_\lambda^T \tilde{V_\lambda}^{-1}(\rho^{(k)}, \lambda^{(k)})\mathbf{H}_\lambda\right)^{-1} \mathbf{H}_\lambda^T \tilde{V_\lambda}^{-1}(\rho^{(k)}, \lambda^{(k)})[D(\mathbf{y}, \mu) - \phi(\lambda^{(k)})]; \quad (3.35)$$

(iv) para $\rho$:

$$\rho^{(k+1)} = \mathbf{M}^{-1}(\mathbf{y} - \mu(\beta^{(k+1)}))\mathbf{W}(\mathbf{y} - \mu(\beta^{(k+1)}))[\mathbf{M}^{-1}\mathbf{W}(\mathbf{y} - \mu(\beta^{(k+1)}))^T(\mathbf{y} - \mu(\beta^{(k+1)}))\mathbf{M}^{-1}]^{-1}. \quad (3.36)$$

em que $\mathbf{H}_\gamma$, $\mathbf{H}_\beta$ e $\mathbf{H}_\lambda$ são as respectivas matrizes de derivadas parciais $[\frac{\partial \mathbf{p}}{\partial \gamma}]^T$, $[\frac{\partial \mu}{\partial \beta}]^T$ e $[\frac{\partial \phi}{\partial \lambda}]^T$, avaliadas em $\gamma^{(k)}$, $\beta^{(k)}$ e $\lambda^{(k)}$, e as matrizes $\widetilde{V_\gamma}^{-1}(\rho^{(k)}, \gamma^{(k)}) = [\mathbf{A}^{1/2}\mathbf{R}_\gamma(\rho^{(k)})\mathbf{A}^{1/2}]^{-1}$, com $\mathbf{A} = diag(p_1^{(k)}(1-p_1^{(k)}), ..., p_L^{(k)}(1-p_L^{(k)}))$, $\widetilde{V_\beta}^{-1}(\rho^{(k)}, \beta^{(k)}) = [\mathbf{B}^{1/2}\mathbf{R}_\beta(\rho^{(k)})\mathbf{B}^{1/2}]^{-1}\mathbf{U}^{(k)}$, com $\mathbf{B} = diag(V(\mu_1^{(k)}), ..., V(\mu_L^{(k)}))$ e $\widetilde{V_\lambda}^{-1}(\rho^{(k)}, \lambda^{(k)}) = [\mathbf{C}^{1/2}\mathbf{R}_\lambda(\rho^{(k)})\mathbf{C}^{1/2}]^{-1}\mathbf{U}^{(k)}$, com $\mathbf{C} = diag(2\phi_1^{2(k)}, ..., 2\phi_L^{2(k)})$.

## 3.3 Distribuição dos Estimadores via Bootstrap

Liang & Zeger (1986) mostraram que os estimadores **GEE** são consistentes e assintoticamente normais, para qualquer escolha da matriz de correlação, desde que o modelo de regressão para a média esteja corretamente especificado. No entanto, o **GEE** padrão não envolve qualquer variável latente. Em nosso modelo a variável latente (U) existe e o algoritmo **ES** substitui essa variável pela média condicional dada a variável resposta e as estimativas dos parâmetros. Mas, ignorando a variação dessa substituição, de cada variável latente pela

média condicional, teremos para os estimadores variâncias estimadas menores do que as verdadeiras (Kong *et al.* (2015)). Por isso, optamos por uma abordagem *Bootstrap* em dois estágios para a realização de inferências sob os parâmetros do modelo.

Para descrever o processo de reamostragem a partir dos estimadores dos parâmetros, considere no primeiro estágio o vetor $\mathbf{Z}^* = (Z_1^*, ..., Z_L^*)$ como sendo uma amostra de uma distribuição de contagem proposta $\mathbb{P}(\mu_i(\widehat{\beta}))$, como por exemplo a Poisson, com média $\mu_i(\widehat{\beta})$, que é o estimador da média obtido através do processo de estimação, descrito na seção anterior, obtido através do processo de estimação com os dados reais, e

$$\mu_i^*(\mathbf{Z}^*) = \mu_i(\widehat{\beta}) + \widehat{\rho} \sum_{j \sim l} \frac{w_{lj}}{w_{l+}} [Z_j^* - \mu_j(\widehat{\beta})]. \tag{3.37}$$

No segundo estágio, para reamostrar valores do Modelo **QIZDE** usamos os seguintes passos,

1. gere $s_i \sim \mathbb{U}(0,1)$;
2. se $s_i \leq p_i(\widehat{\gamma})$, $Y_i^* = 0$. Senão, gere $Y_i^* \sim \widehat{\phi} \times F_i$, com $F_i \sim Poisson\big(\frac{\mu_i^*(\mathbf{Z}^*)}{\widehat{\phi}}\big)$;
3. Na $n$-ésima reamostra, repetindo os passos 1. e 2. para $i = 1, 2, ..., L$. Teremos o vetor reamostrado $\mathbf{Y}_n^* = (Y_1^*, ..., Y_L^*)_n$;
4. Repita o passo 3. para $n = 1, 2, ..., N$ e obtenha as amostras Bootstrap $\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_N^*$.

Nota-se que a esperança condicional de cada componente do vetor de reamostras $\mathbf{Y}^*$ proveniente do processo de contagem, é da forma $E(Y_i^*|\mathbf{Z}^*) = \mu_i^*(\mathbf{Z}^*)$, sua esperança marginal é igual a $E(Y_i^*) = E[E(Y_i^*|\mathbf{Z}^*)] = \mu_i(\widehat{\beta})$. E, portanto, $E(Y_i^* - \mu_i(\widehat{\beta})) = 0$, $E(D(Y_i^*; \mu_i(\widehat{\beta}))) \cong \widehat{\phi}$ e ainda, a estrutura de dependência dada em (3.23) é preservada. Desta forma, sejam os estimadores $(\widehat{\mu}_i, \widehat{p}_i) \equiv (\mu_i(\widehat{\beta}) p_i(\widehat{\gamma}))$, teremos que

$$\sum_{i=1}^{L} \left\{ \frac{\partial \widehat{p}_i}{\partial \gamma^*} \right\}^T [\mathbf{A}_i^{1/2} \mathbf{R}_{\gamma^*}(\rho) \mathbf{A}_i^{1/2}]^{-1} (u_i^* - \widehat{p}_i) = 0, \tag{3.38}$$

$$\sum_{i=1}^{L} \left\{ \frac{\partial \widehat{\mu}_i}{\partial \beta^*} \right\}^T [\mathbf{B}_i^{1/2} \mathbf{R}_{\beta^*}(\rho) \mathbf{B}_i^{1/2}]^{-1} \mathbf{U}^* (y_i^* - \widehat{\mu}_i) = 0, \tag{3.39}$$

$$\sum_{i=1}^{L} \left\{ \frac{\partial \widehat{\phi}_i}{\partial \lambda^*} \right\}^T [\mathbf{C}_i^{1/2} \mathbf{R}_{\lambda^*}(\rho) \mathbf{C}_i^{1/2}]^{-1} \mathbf{U}^* (D(y_i^*, \widehat{\mu}_i) - \widehat{\phi}_i) = 0, \tag{3.40}$$

são **GEE** empíricas para obter os estimadores *Bootstrap* $(\widehat{\beta}^*, \widehat{\gamma}^*, \widehat{\phi}^*)$ de $(\beta, \gamma, \phi)$. Sendo assim, testes de hipótese e intervalos de confiança *Bootstrap* podem ser realizados no modelo proposto através da sequência de estimadores *Bootstrap* $(\widehat{\beta}^*, \widehat{\gamma}^*, \widehat{\phi}^*)_n, n = 1, 2, ..., N$.

# Capítulo 4

# Estudo de Simulação

## 4.1 Descrição do Estudo

Neste capítulo mostraremos diversos estudos de simulação realizados a fim de verificar o comportamento dos estimadores propostos pelo novo modelo (**QIZDE**). Consideramos como cenário o mapa do estado do Amazonas (ver Figura 4.1), com seus 62 municípios, e executamos esses estudos simulados para algumas variações dos parâmetros do modelo, que serão descritos posteriormente.

Seja a variável resposta $Y_i$ pertencente à classe de modelos com quase verossimilhança completa, com $Y_i \sim \mathbf{QIZDE}(p_i, \mu_i, \phi_i, \rho)$, $i = 1, 2..., 62$, em que $p_i$ é a probabilidade de ocorrer um zero estrutural, $\mu_i$ é a média de ocorrências, $\phi_i$ é a sobredispersão e $\rho$ é o parâmetro que mede a dependência espacial entre as áreas. Em cada estudo, consideramos as covariáveis $X_i \sim U(0,1)$ para modelar as contagens positivas e $G_i \sim U(0, 0.5)$ para modelar o zero estrutural (essas mesmas covariáveis foram utilizadas no estudo de simulação em Nieto-Barajas & Bandyopadhyay (2013)), em que $i$ é o índice que referencia a i-ésima área. Assim, os modelos de regressão são descritos por:

$$X_i^T \beta = \beta_0 + \beta_1 X_i \qquad \text{e} \qquad G_i^T \gamma = \gamma_0 + \gamma_1 G_i. \tag{4.1}$$

Para o componente do modelo de mistura, referente a modelagem da média da distribuição de contagem, a função de ligação adotada foi a logarítmica ($\eta_i = \log(\mu_i) = X_i^T \beta$). E, para o componente referente ao zero estrutural, a função de ligação utilizada foi o logit($p_i$). Em todos os cenários, utilizamos uma matriz de vizinhança espacial $\mathbf{W}$, considerando que $w_{ij} = 1$ se $i \sim j$, ou seja, se a área i é vizinha da área j assume-se peso 1, e 0 caso contrário. Sob essas especificações, a variável resposta $Y_i$, que é a contagem observada na i-ésima área, foi gerada da seguinte forma:

(i) se uma variável aleatória $s_i \sim U(0,1) \leq p_i(\gamma)$, então $y_i = 0$;

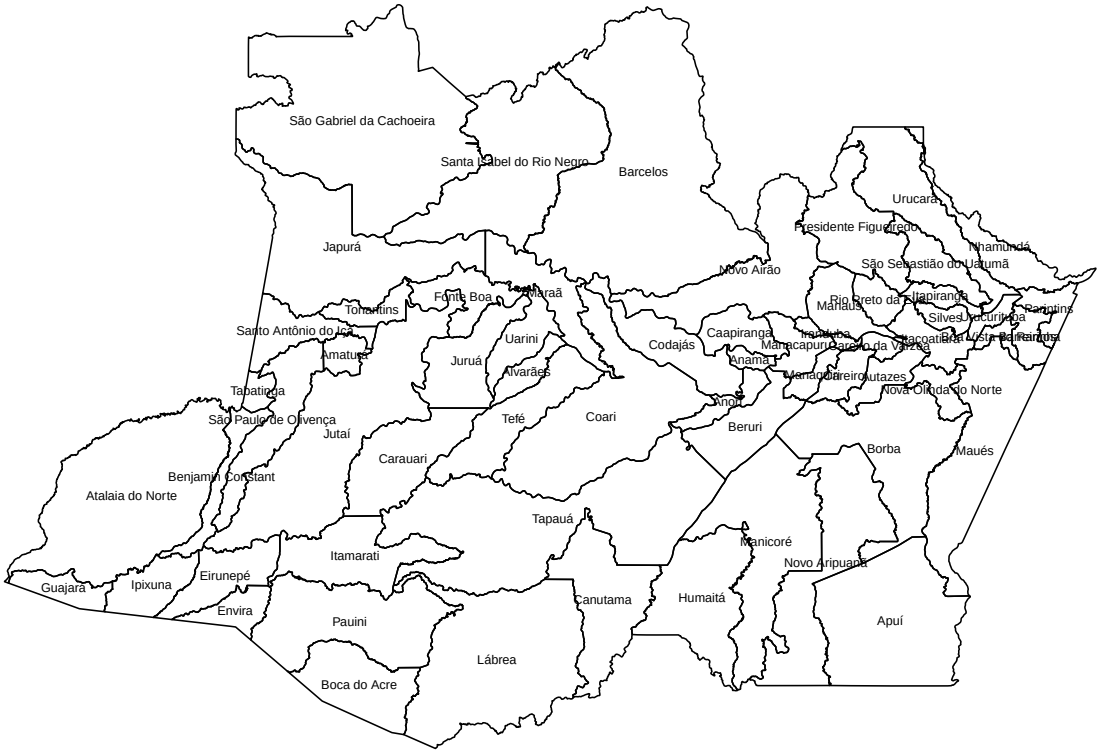


Figura 4.1: Mapa do Estado do Amazonas, com seus 62 municípios.

(ii) caso contrário, $y_i \sim \phi * T_i$, com $T_i \sim Poisson(\mu_i^*/\phi)$, em que:

$$\mu_i^* = \mu_i(\beta) + \rho \sum_{j \in \partial_i} \frac{w_{ij}}{w_{i+}} (Y_j - \mu_j(\beta))) \tag{4.2}$$

é escrita de acordo com o modelo **CAR** (ver expressão 3.24), com a parametrização descrita pela expressão (4.2).

Com a finalidade de garantir a comparabilidade entre os estudos de simulação para os cenários propostos, foi necessário desenvolver uma fórmula geral que nos permitisse manipular a proporção de zeros vindas do excesso e da distribuição de contagem. Sendo assim, consideramos a quase verossimilhança completa para o modelo proposto (**QIZDE**), definida em (3.7), em que $Y_i \sim \phi * T_i$ é uma quase-Poisson inflacionada de zeros, com $T_i \sim Poisson(\mu_i^*/\phi)$, e $y = 0$, obtendo a relação da quase verosimilhança estendida, que compõe a estrutura de (3.7), com o logaritmo da probabilidade de ocorrência do zero

($\log[P(Y_i = 0)]$), dada da seguinte forma:

$$Q^{+}(\mu_i, \phi, y_i = 0) = -\frac{\mu_i}{\phi} - \frac{1}{2}\log(\phi), \tag{4.3}$$

Aplicando a função exponencial a expressão (4.3), obtemos:

$$\begin{aligned} P(Y_i = 0|T_i) &\approx \exp\Big\{-\frac{\mu_i}{\phi} - \frac{1}{2}\log(\phi)\Big\} \\ &\approx e^{-\frac{\mu_i}{\phi}} \exp\{\log(\phi^{-\frac{1}{2}})\} \\ &\approx \phi^{-\frac{1}{2}} e^{-\frac{\mu_i^*}{\phi}}. \end{aligned} \tag{4.4}$$

Como a proporção de zeros, gerada para o modelo **QIZDE**, pode ser calculada como a probabilidade da variável resposta ser igual a zero, quando ele vier do zero estrutural ou das contagens positivas, resultando em uma soma de probabilidades, dada pela seguinte expressão:

$$P(Y_i = 0) = P(Y_i = 0|T_i \sim 0) + P(Y_i = 0|T_i \sim \phi Poi(\mu_i^*/\phi)) = p_i + \phi^{-\frac{1}{2}} e^{-\frac{\mu_i}{\phi}}, \tag{4.5}$$

em que $p_i$ é a probabilidade de ocorrer um zero estrutural.

Assim, fixamos os valores de $\gamma_1$ e $\beta_1$, e obtemos uma fórmula geral para obtenção de $\gamma_0$ e $\beta_0$, sob cada variação de $\phi$, onde assumimos que a proporção de zeros $\widetilde{p_i} = \phi^{-\frac{1}{2}} e^{-\frac{\mu_i}{\phi}}$, se $Y_i \sim \phi * T_i$, que representa a proporção de zeros proveniente da distribuição de contagem, e $\overline{p_i} = p_i$, se $Y_i \sim 0$, que representa a proporção de zeros vinda do zero estrutural. Então, fixando $\beta_1$, obtivemos que:

$$E_X[\log(\mu_i)] = E[\beta_0 + \beta_1 X_i] = \beta_0 + \beta_1 \frac{1}{2}, \tag{4.6}$$

logo o $\log(\mu_i) \cong \beta_0 + \beta_1 \frac{1}{2}$. Fixando $\gamma_1$, tivemos que:

$$E_G[logit(p_i)] = E[\gamma_0 + \gamma_1 G_i] = \gamma_0 + \gamma_1 \frac{1}{4}, \tag{4.7}$$

logo o $logit(p_i) \cong \gamma_0 + \gamma_1 \frac{1}{4}$.

Fixando $\phi = 2$ e utilizando as equações (4.6) e (4.7), desenvolvemos as seguintes

expressões, das quais para $\beta_0$ obtivemos que:

$$\begin{aligned}
\log(\widetilde{p}_i) &\cong \log\left(\phi^{-\frac{1}{2}} e^{-\frac{\mu_i}{\phi}}\right) = -\frac{1}{2}\log(2) - \frac{\mu_i}{2} \\
2\log(\widetilde{p}_i) &\cong -\mu_i - \log(2) \\
\mu_i &\cong -\log(2) - 2\log(\widetilde{p}_i) \\
\exp\left[\beta_0 + \beta_1 \frac{1}{2}\right] &\cong -\log(2) - 2\log(\widetilde{p}_i) \\
\beta_0 &\cong \log\Big\{ -\log(2) - 2\log(\widetilde{p}_i)\Big\} - \beta_1 \frac{1}{2}, \qquad (4.8)
\end{aligned}$$

e para $\gamma_0$ obtivemos:

$$\begin{aligned}
logit(p_i) &= \gamma_0 + \gamma_1 Z_i \\
\gamma_0 &\cong logit(p_i) - \gamma_1 \frac{1}{4} \qquad (4.9)
\end{aligned}$$

Com esse mesmo procedimento, encontramos outras expressões para alguns valores fixados de $\phi$, que são apresentados na Tabela (4.1).

Tabela 4.1: Fórmula Geral para $\beta_0$ e $\gamma_0$ de acordo com o $\phi$.

| $\phi$ | $\beta_1$ (fixo) ($\widetilde{p}_i$, % de zeros da Poisson) | $\gamma_1$ (fixo) ($p_i$, % de zeros da Bernoulli) |
|---|---|---|
| 1 | $\beta_0 \cong log(-log(\widetilde{p}_i)) - \frac{\beta_1}{2}$ | $\gamma_0 \cong logit(p_i) - \frac{\gamma_1}{4}$ |
| 2 | $\beta_0 \cong log[-2log(\widetilde{p}_i) - log(2)] - \frac{\beta_1}{2}$ | |
| 4 | $\beta_0 \cong log\{-4[log(\widetilde{p}_i) + log(2)]\} - \frac{\beta_1}{2}$ | |

Todos os estudos de simulação foram realizados através do software estatístico **R**, versão 3.0.2. E, ainda, para cada estudo, utilizamos os resultados da Tabela 4.1, na qual fixamos $\gamma_1 = 2$, $\beta_1 = 1$ e $\rho = 0.5$, para obter os respectivos valores de $\beta_0$ e $\gamma_0$. O valor fixado de $\rho = 0.5$ é bastante comum na literatura (Yasui & Lele (1997)). Cada uma dessas simulações foram realizadas com Bootstrap duplo, dos quais temos um Bootstrap externo de tamanho 1.000, para calcular as probabilidades de cobertura, e um interno de tamanho 300, para gerar os intervalos de confiança, ou seja, foram gerados um total de 300.000 mapas distintos nesse processo de simulação. Fizemos alguns testes iniciais e percebemos que os resultados obtidos com um *bootstrap* interno de 300 é equivalente ao obtido com tamanhos maiores, como por exemplo o de tamanho 500. Os valores de $\beta_0$ e $\gamma_0$ foram calculados, também, para cada valor fixado do parâmetro $\phi$, mas escolhemos apenas $\phi = 2,4$ que caracterizam a sobredispersão. Desse modo, obtivemos os valores de $\beta_0$ e $\gamma_0$ para o estudo de simulação (Tabela 4.2), no qual a proporção de zeros total é igual a 45% .

Os Intervalos de Confiança *Bootstrap*-t (**ICBt**) (ver Seção 2.7) para a média das estimativas dos parâmetros, foram criados ao nível de 95% de confiança, cujas médias foram calculadas das 300 reamostras *Bootstrap* (*Bootstrap* interno), que por sua vez foram geradas

a partir do valor estimado em cada repetição (*Boostrap* externo), de forma similar ao procedimento de estimação descrito na Seção 4.3. As probabilidades de cobertura *Bootstrap* nada mais são que a proporção de intervalos **ICBt** que contiveram o valor estimado, em cada repetição, para seus respectivos parâmetros.

No processo de reamostragem *Bootstrap*, um fato importante é que se a matriz $\widetilde{\mathbf{V}}(\rho^{(k)}, \beta^{(k)})$ inserida no estimador de $\beta$ (ver equação 3.26) for computacionalmente singular ($det(\widetilde{\mathbf{V}}(\rho^{(k)}, \beta^{(k)})) \to 0$), a amostra é descartada gerando-se uma nova, pois no processo de estimação é necessário obter a inversa dessa matriz.

Tabela 4.2: Estudo de Simulação, para 45% de Zeros.

| $p_i$ (% de zeros da Inflação) | $\widetilde{p}_i$ (% de zeros da Poisson) | $\gamma_0$ | $\phi$ | $\beta_0$ |
|---|---|---|---|---|
| 22,5 | 22,5 | -1.7368 | 1 | -0.1001 |
| | | | 2 | 0.3286 |
| | | | 4 | 0.6613 |
| 30 | 15 | -1.3473 | 1 | 0.1403 |
| | | | 2 | 0.6317 |
| | | | 4 | 1.0719 |
| 15 | 30 | -2.2346 | 1 | -0.3144 |
| | | | 2 | 0.0393 |
| | | | 4 | 0.2146 |

A partir dessas especificações iniciais, realizamos estudos de simulação sob três perspectivas a serem comparadas, nas quais a proporção de zeros estruturais é maior, menor ou igual a proporção de zeros vinda da distribuição de contagem. Em cada estudo, analisamos as qualidades dos estimadores dos parâmetros de regressão do modelo proposto, através de suas estimativas médias, vícios médios, erros padrão e probabilidades de cobertura baseadas em intervalos de confiança Bootstrap-t.

## 4.2 Resultados

O cenário deste estudo foi o mapa do estado do Amazonas-AM, com suas 62 áreas (municípios), ou seja, $L = 62$. Todos os 1000 intervalos **ICBt** dos parâmetros foram gerados a partir das reamostras *bootstrap*, para cada valor fixado do parâmetro de sobredispersão $\phi$, que no caso foram 2 e 4. Os resultados deste estudo estão dispostos nas tabelas 4.3, 4.4 e 4.5.

No que tange a estimativa do vetor $\beta$, sob a perspectiva de que a proporção de zeros vinda do excesso é maior do que a da distribuição de contagem (Tabela 4.3), os estimadores $(\widehat{\beta}_0, \widehat{\beta}_1)$ obtiveram em média estimativas razoáveis e vícios pequenos, evidenciando o não viés desses estimadores, com erros padrão pequenos. As probabilidades de cobertura estão um pouco abaixo do esperado (95%), no entanto são maiores e mais próximas do esperado quando $\phi = 2$, sofrendo uma redução com o aumento da sobredispersão ($\phi = 4$), sendo mais

Tabela 4.3: Mapa do AM, para 45% de Zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

| $\phi$ fixo | Parâmetro | Valor Verdadeiro | Estimativa Média | Vício Médio | Erro Padrão | Prob. de Cobertura Bootstrap-t |
|---|---|---|---|---|---|---|
| 2 | $\beta_0$ | 0.6317 | 0.7428 | 0.1000 | 0.3095 | 0.8620 |
| | $\beta_1$ | 1.0000 | 0.9344 | -0.0788 | 0.4863 | 0.8970 |
| | $\gamma_0$ | -1.3473 | -1.2613 | 0.0654 | 0.4880 | 0.9690 |
| | $\gamma_1$ | 2.0000 | 2.0370 | 0.0189 | 1.7186 | 0.9720 |
| 4 | $\beta_0$ | 1.0719 | 1.4462 | 0.1333 | 0.2340 | 0.6130 |
| | $\beta_1$ | 1.0000 | 0.7430 | -0.1109 | 0.3725 | 0.8010 |
| | $\gamma_0$ | -1.3473 | -0.9996 | 0.1055 | 0.5227 | 0.9500 |
| | $\gamma_1$ | 2.0000 | 1.7622 | -0.0761 | 1.8034 | 0.9820 |

precisos no primeiro caso. Este fato foi mais perceptível em $\widehat{\beta_0}$, que teve maior redução, em torno de $24,9\%$. Com isso, a sobredispersão aparenta exercer certa influência sobre as estimativas dos betas. Essa probabilidade de cobertura abaixo do esperado, também, pode ser explicada, pelo fato de que o **ICBt** possui um erro de aproximação da ordem $\frac{1}{\sqrt{L}} \cong 0.127$, o qual é um valor razoavelmente grande.

Para os parâmetros que modelam o excesso de zeros $(\gamma_0, \gamma_1)$, sob a mesma perspectiva (Tabela 4.3), obtiveram em média estimativas bem próximas ao seu verdadeiro valor e vícios pequenos, em ambos valores fixados de $\phi$, nos dando indícios de estimadores não viesados. O estimador $\widehat{\gamma_0}$ apresentou erros padrão pequenos, ao contrário de $\widehat{\gamma_1}$. Em contrapartida, ambos tiveram as maiores probabilidades de cobertura, acima do esperado (95%).

Tabela 4.4: Mapa do AM, para 45% de Zeros, sendo 22,5% da $\sim$ Ber e 22,5% da $\sim$ Poi.

| $\phi$ fixo | Parâmetro | Valor Verdadeiro | Estimativa Média | Vício Médio | Erro Padrão | Prob. de Cobertura Bootstrap-t |
|---|---|---|---|---|---|---|
| 2 | $\beta_0$ | 0.3286 | 0.5280 | 0.1736 | 0.3072 | 0.9090 |
| | $\beta_1$ | 1.0000 | 0.9410 | -0.0929 | 0.5306 | 0.9180 |
| | $\gamma_0$ | -1.7368 | -1.6127 | 0.1506 | 0.3887 | 0.9760 |
| | $\gamma_1$ | 2.0000 | 1.8925 | -0.1089 | 1.3756 | 0.9590 |
| 4 | $\beta_0$ | 0.6613 | 1.0880 | 0.2399 | 0.3646 | 0.7030 |
| | $\beta_1$ | 1.0000 | 0.9177 | -0.1593 | 0.5829 | 0.7840 |
| | $\gamma_0$ | -1.7368 | -1.3918 | 0.1982 | 0.4602 | 0.9800 |
| | $\gamma_1$ | 2.0000 | 1.9730 | -0.0501 | 1.5256 | 0.9880 |

A Tabela 4.4, apresenta os resultados sob outra perspectiva, de que a proporção de zeros estruturais é igual a da distribuição de contagem. Na qual, os estimadores $\widehat{\beta_0}$ e $\widehat{\beta_1}$, também, mostraram-se não tendenciosos, pois suas estimativas médias foram próximas aos valores verdadeiros dos parâmetros e seus vícios foram pequenos. Aqui, também, tiveram erros padrão pequenos, evidenciando a consistência desses estimadores. Suas probabilidades de cobertura ainda continuam pouco abaixo do esperado (95%), no entanto são maiores aqui, do que no resultado anterior. Essa melhora pode ter ocorrido devido à redução na proporção

de zeros estruturais. Essas probabilidades são mais próximas do esperado em $\phi = 2$, sofrendo uma redução quando $\phi = 4$, sendo mais precisos no primeiro caso. Note que, assim como no resultado anterior, essa redução foi mais perceptível em $\widehat{\beta}_0$, que obteve a maior redução ($\cong 21,5\%$). Esse fato corrobora à suposição de que a sobredispersão realmente afete, de alguma forma, as estimativas desses parâmetros.

Avaliando os estimadores de $\gamma_0$ e $\gamma_1$, sob o mesmo ponto de vista (Tabela 4.4), observamos que em média obtiveram estimativas bastante próximas do verdadeiro valor, com vícios pequenos. Da mesma forma que o resultado anterior, o estimador $\widehat{\gamma}_0$ apresentou erros padrão pequenos, e $\widehat{\gamma}_1$ erros padrão maiores. No entanto, ambos mostraram-se precisos, com probabilidades de cobertura bem acima do esperado (95%), comportamento similar ao estudo anterior.

Tabela 4.5: Mapa do AM, para 45% de Zeros, com 15% da $\sim$ Ber e 30% da $\sim$ Poi.

| $\phi$ fixo | Parâmetro | Valor Verdadeiro | Estimativa Média | Vício Médio | Erro Padrão | Prob. de Cobertura Bootstrap-t |
|---|---|---|---|---|---|---|
| 2 | $\beta_0$ | 0.0393 | 0.2318 | 0.1701 | 0.2928 | 0.8980 |
| | $\beta_1$ | 1.0000 | 1.0009 | -0.0603 | 0.5000 | 0.9240 |
| | $\gamma_0$ | -2.2346 | -2.3963 | 0.0208 | 0.3499 | 0.9650 |
| | $\gamma_1$ | 2.0000 | 3.1588 | 0.5009 | 1.3428 | 0.9380 |
| 4 | $\beta_0$ | 0.2146 | 0.2427 | 0.2997 | 0.4348 | 0.8220 |
| | $\beta_1$ | 1.0000 | 1.6017 | -0.1178 | 0.5739 | 0.7060 |
| | $\gamma_0$ | -2.2346 | -1.7771 | 0.4098 | 0.3311 | 0.9610 |
| | $\gamma_1$ | 2.0000 | 1.9225 | -0.2830 | 1.2081 | 0.9560 |

Considerando agora que a proporção de zeros estruturais é menor do que a da distribuição de contagem (ver Tabela 4.5), os estimadores dos betas ($\widehat{\beta}_0, \widehat{\beta}_1$), assim como nos estudos anteriores, obtiveram em média estimativas próximas aos verdadeiros valores dos parâmetros e vícios pequenos, ratificando a suposição de não viés, e erros padrão pequenos. E, suas probabilidades de cobertura permaneceram abaixo do esperado (95%), porém são mais próximas de 95% quando $\phi = 2$, sofrendo uma redução com o aumento da sobredispersão, ou seja, são mais acurados com menor sobredispersão. Diferente dos demais estudos, pôde-se perceber mais claramente essa redução na probabilidade de cobertura para o estimador $\widehat{\beta}_1$, que apresentou maior redução ($\cong 21,8\%$) aqui neste estudo. Novamente, é notável que a sobredispersão interfere nas estimativas desses parâmetros.

Sob a mesma perspectiva (Tabela 4.5), os estimadores $\widehat{\gamma}_0$ e $\widehat{\gamma}_1$ mostraram-se não viesados, pois obtiveram em média estimativas próximas ao verdadeiro valor, o que também foi observado nos estudos anteriores. O estimador $\widehat{\gamma}_0$ apresentou erros padrão pequenos, e $\widehat{\gamma}_1$ obteve erros padrão maiores, mas ambos obtiveram altas probabilidades de cobertura, quase todas foram acima do esperado (95%), exceto para $\widehat{\gamma}_1$, com $\phi = 2$, que teve sua probabilidade pouco abaixo, mas bastante próxima de 95%, não interferindo na sua precisão.

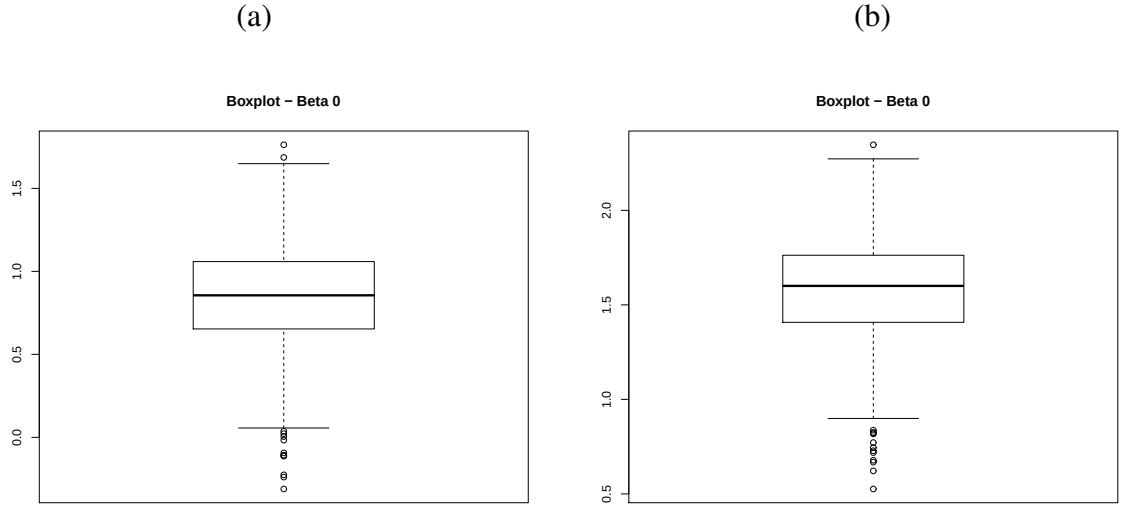


Figura 4.2: Boxplot das estimativas de $\beta_0$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

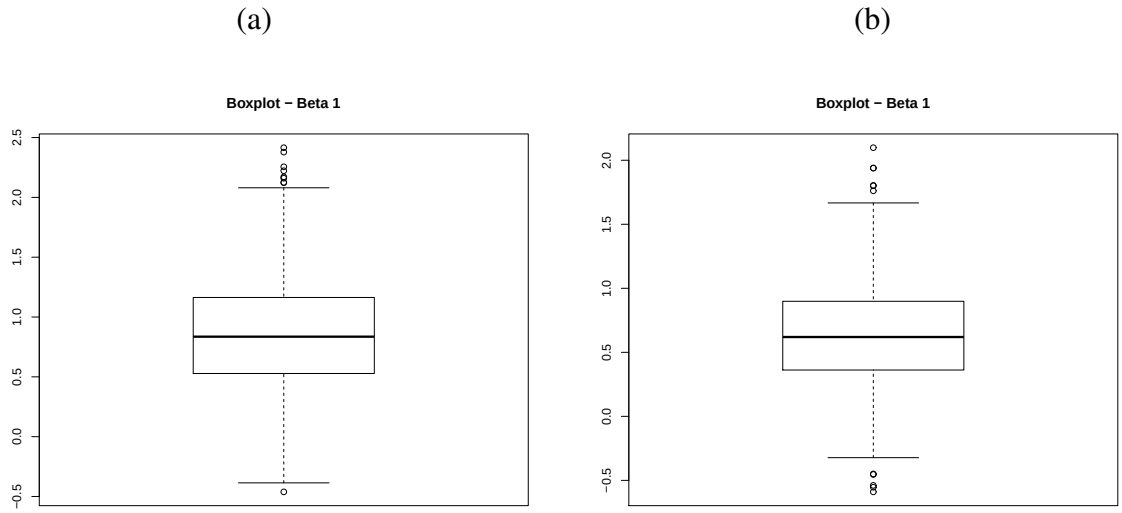


Figura 4.3: Boxplot das estimativas de $\beta_1$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

Outro comportamento que podemos visualizar para o estimador $\widehat{\beta_0}$ (ver Figura 4.2) é o de simetria no lado (a) e no lado (b), evidenciando uma distribuição assintoticamente Normal, com a presença de *outliers* em ambos os lados da figura. O estimador $\widehat{\beta_1}$ (ver Figura 4.3) mostrou-se simétrico em ambos os lados (a) e (b), nos dando indícios de que possuem distribuição Normal assintótica, com a presença de *outliers*.

O estimador $\widehat{\gamma_0}$ apresentou comportamento levemente assimétrico à direita (ver Figura 4.4), afastando-se da suposição de normalidade assintótica, e amplitudes equivalentes nos lados (a) e (b). E o estimador $\widehat{\gamma_1}$ (ver Figura 4.5) apresentou simetria em (a) e (b),

(a) (b)

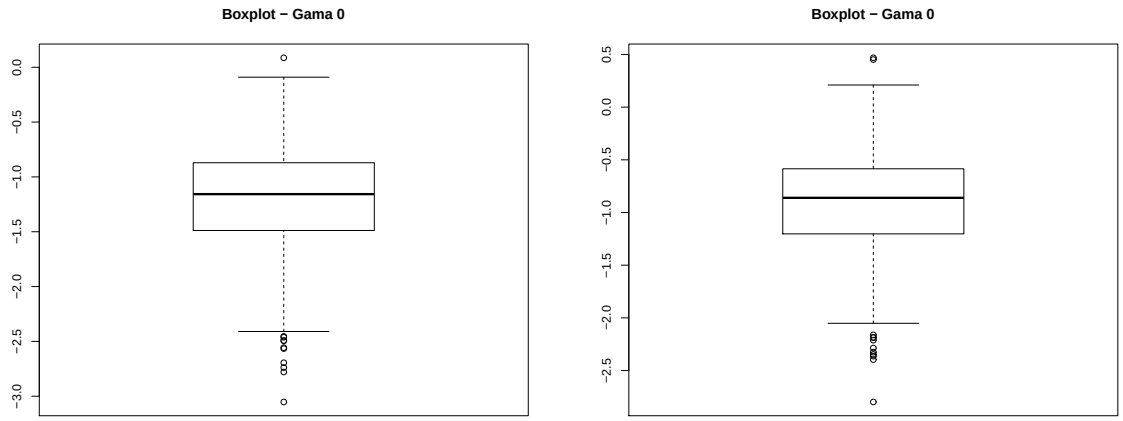


Figura 4.4: Boxplot das estimativas de $\gamma_0$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

(a) (b)

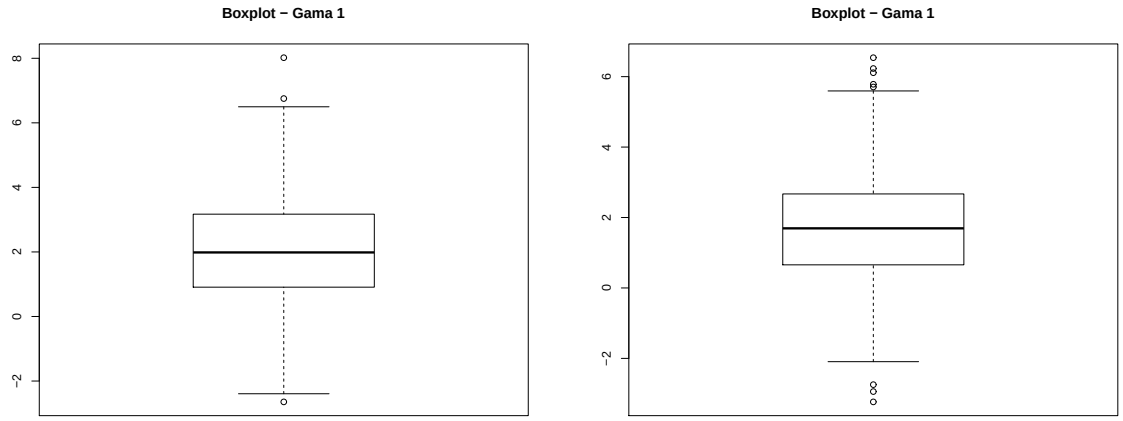


Figura 4.5: Boxplot das estimativas de $\gamma_1$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

aproximando-se da suposição de normalidade assintótica.

Como os parâmetros $\rho$ e $\phi$ são limitados, respectivamente, por $[\rho_{min}, 1]$ e $(0, \infty)$, podemos avaliar também a mediana e a simetria deles, em todos os estudos realizados. Sendo assim, sob a perspectiva da Figura 4.6, $\widehat{\rho}$ mostrou-se simétrico tanto no lado (a), com mediana $\cong 0.22$, quanto no lado (b), com mediana $\cong 0.14$, aparentando comportamento assintoticamente Normal. Nota-se ainda a presença de *outliers* em ambos, onde apesar do lado (b) apresentar a menor amplitude, suas estimativas foram abaixo do valor verdadeiro (0.5), e a mediana no lado (a) foi bem mais próxima dele.

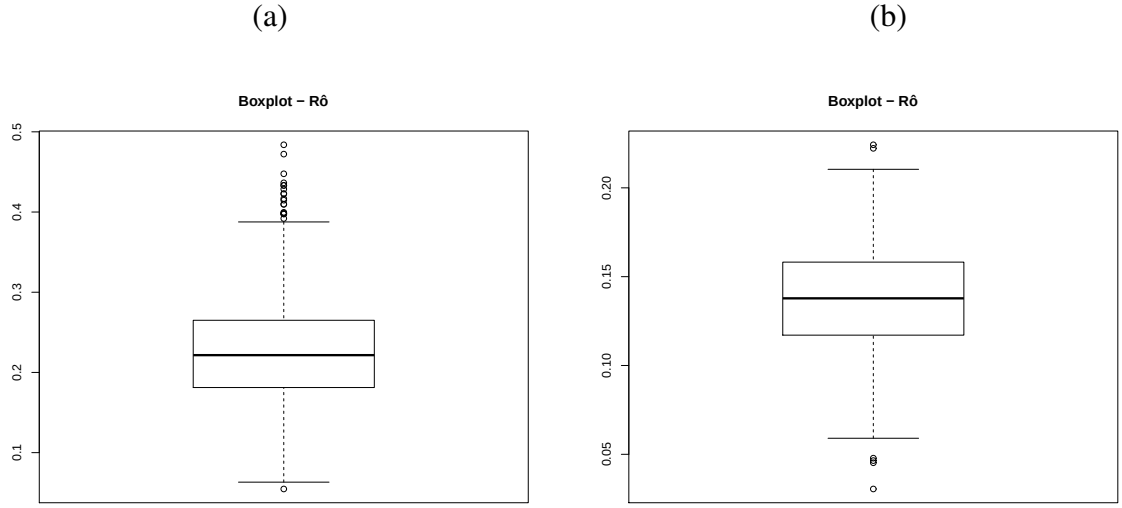


Figura 4.6: Boxplot das estimativas de $\rho$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

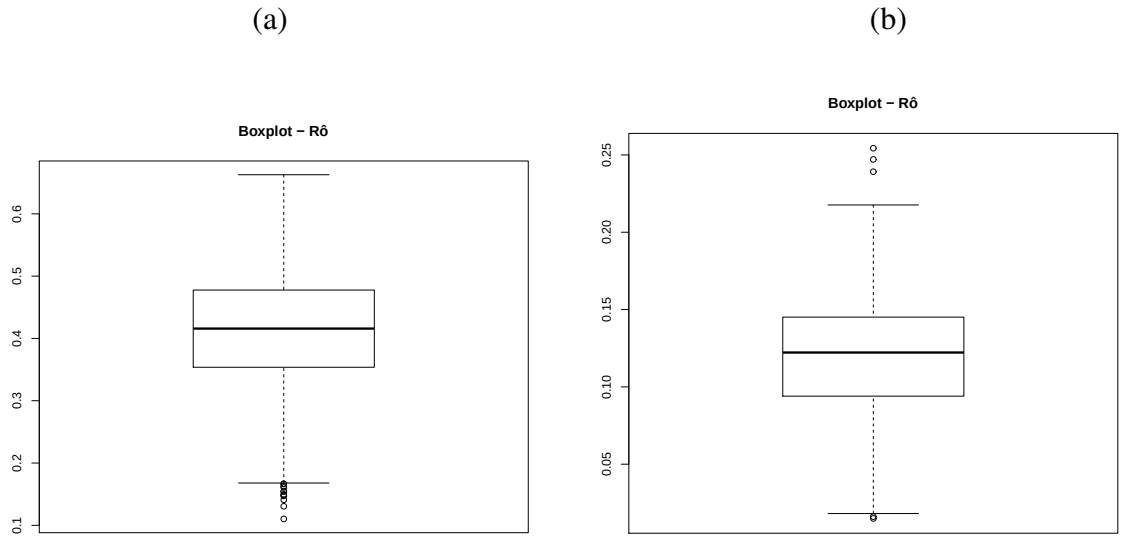


Figura 4.7: Boxplot das estimativas de $\rho$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 22,5% da $\sim$ Ber e 22,5% da $\sim$ Poi.

Na Figura 4.7, o desempenho de $\widehat{\rho}$ melhorou, mas somente em (a), continuando com comportamento simétrico, ratificando a hipótese de normalidade assintótica, e com mediana ($\cong 0.41$), bem mais próxima do verdadeiro valor. Ao contrário do resultado anterior, apresentou assimetria à direita no lado (b), afastando-se da suposição de normalidade, com mediana $\cong 0.13$ distante do verdadeiro valor.

O melhor desempenho de $\widehat{\rho}$ foi observado na Figura (4.8), que continuou apresentando simetria em (a), confirmando a distribuição assintoticamente Normal, com mediana de

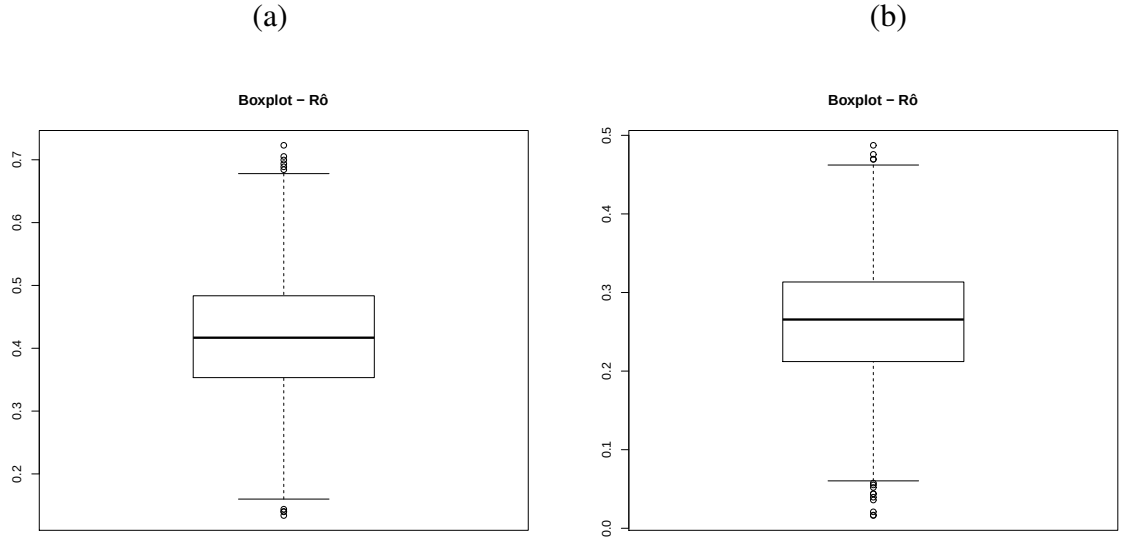


Figura 4.8: Boxplot das estimativas de $\rho$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 15% da $\sim$ Ber e 30% da $\sim$ Poi.

$\cong 0.41$, e assimetria à direita em (b), afastando-se da suposição de normalidade, mas com mediana $\cong 0.27$, bem mais próxima de 0.5 do que nos estudos anteriores. O comportamento em (b) neste cenário foi bem melhor, comparado aos anteriores. Com isso, é possível notar que o estimador $\widehat{\rho}$ tem melhor desempenho quando a proporção de zeros vinda da distribuição de contagem é maior do que a da inflação de zeros e a sobredispersão apresenta-se menor.

Como o parâmetro de dependência espacial é inserido apenas na distribuição de contagem, já era de se esperar que quando a proporção de zeros estrutural fosse menor do que a da distribuição de contagem (Figura 4.8), a performance do seu estimador seria melhor, ou seja, o excesso de zeros pode estar mascarando o valor estimado da dependência espacial, subestimando-o nos outros casos, sendo mais visível quando a sobredispersão é maior ($\phi = 4$).

O estimador do parâmetro de sobredispersão $\widehat{\phi}$ foi assimétrico à direita em (a) (ver Figura 4.9), com cauda superior mais pesada e mediana ($\cong 1.7$) próxima do verdadeiro valor (2), e simétrico em torno de 2.4 em (b), com maior amplitude, no entanto distante do valor esperado (4). Nota-se que o estimador obteve melhor desempenho quando o valor verdadeiro da sobredispersão é menor.

No contexto apresentado pela Figura 4.7, $\widehat{\phi}$ obteve um bom desempenho, mas somente em (a), com mediana ($\cong 2$) equivalente ao verdadeiro valor, e em (b) mostrou-se assimétrico à esquerda, com mediana ($\cong 2.4$), resultados contrários aos anteriores, porém mais próximos do esperado.

O melhor desempenho de $\widehat{\phi}$ foi observado na Figura (4.8), que apesar de assimé-

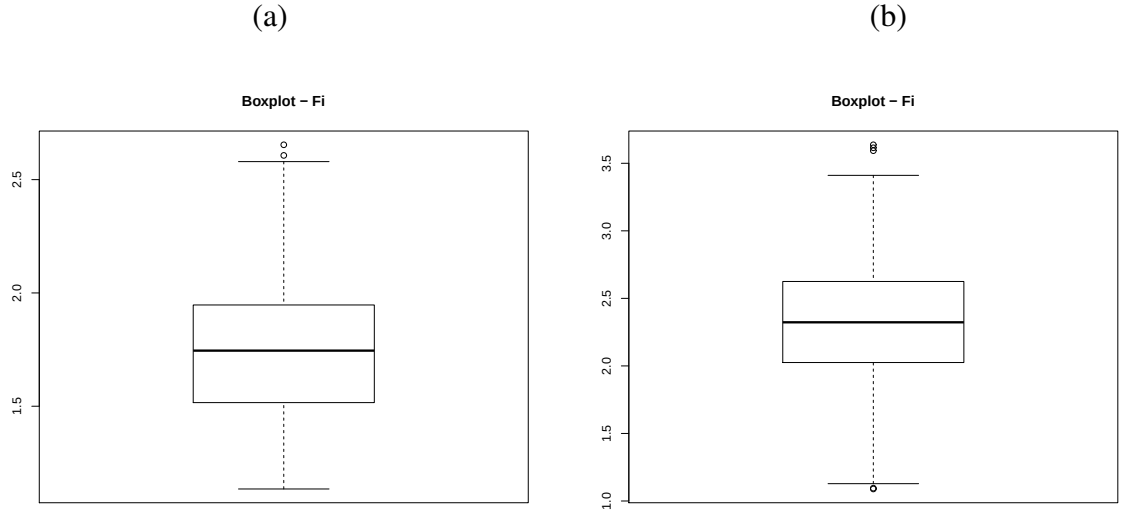


Figura 4.9: Boxplot das estimativas de $\phi$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 30% da $\sim$ Ber e 15% da $\sim$ Poi.

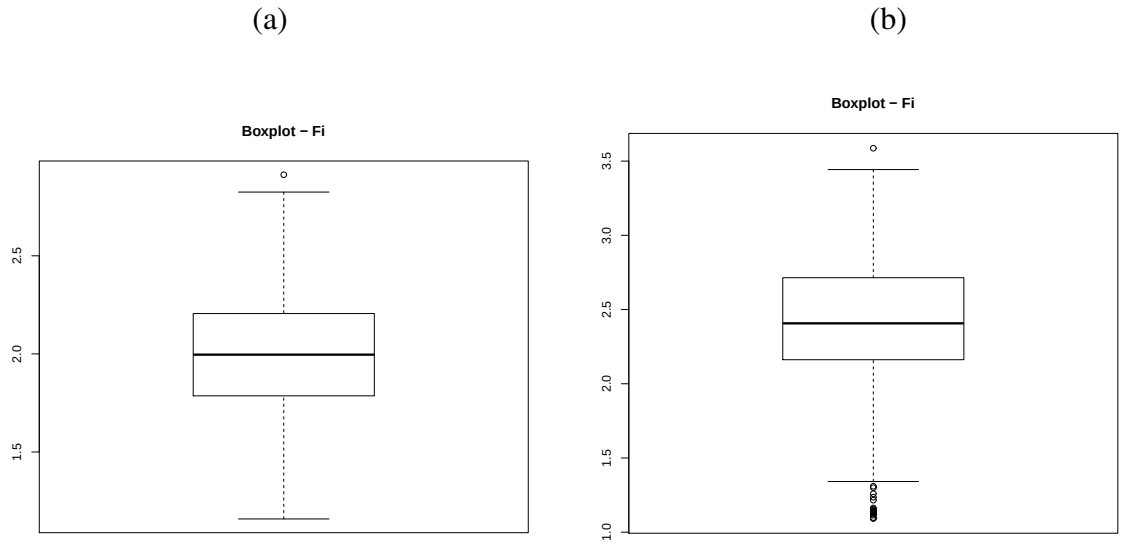


Figura 4.10: Boxplot das estimativas de $\phi$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 22,5% da $\sim$ Ber e 22,5% da $\sim$ Poi.

trico à direita em (a) e em (b), fugindo da suposição de normalidade assintótica, continuou apresentando mediana de $\cong 2$ em (a), equivalente ao esperado, e em (b) obteve mediana de $\cong 3$, muito mais próxima do esperado (4), obtendo assim melhor desempenho do que nos resultados anteriores.

Uma avaliação geral dos resultados obtidos (Tabelas: 4.3, 4.4 e 4.5) é que os estimadores dos betas mostraram-se não viesados em todos os estudos realizados, mas foram mais precisos quando a sobredispersão apresentou-se menor ($\phi = 2$) e obteve melhor desempe-

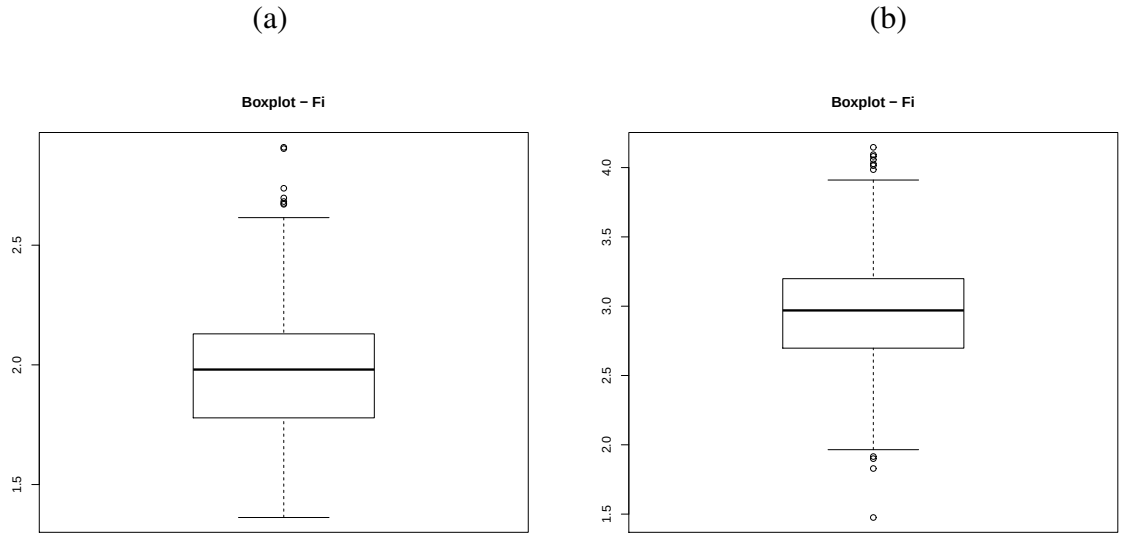


Figura 4.11: Boxplot das estimativas de $\phi$, com $\phi = 2$ (a) e $\phi = 4$ (b), mapa do AM, para 45% de zeros, sendo 15% da $\sim$ Ber e 30% da $\sim$ Poi.

nho quando a proporção de zeros estruturais era menor ou igual do que a da distribuição de contagem. E, graficamente, apresentaram uma comportamento assintoticamente Normal.

Os gamas mostraram-se não viesados e precisos em todos os estudos e com melhor desempenho quando a proporção de zeros estruturais era maior ou igual do que a da distribuição de contagem. No entanto, avaliando os gráficos, $\widehat{\gamma_0}$ não apresentou comportamento assintoticamente Normal, enquanto que $\widehat{\gamma_1}$ mostrou seguir tal comportamento.

No que concerne ao estimador $\widehat{\rho}$, em geral obteve comportamento assintoticamente Normal, apresentou melhor desempenho sob a perspectiva de que a proporção de zeros estruturais é menor do que a da distribuição de contagem, principalmente quando a sobredispersão foi fixada em 2. O que era esperado, pois a dependência espacial foi inserida na distribuição de contagem, então quando a quantidade de zeros vindos da inflação é maior do que a da distribuição de contagem, percebemos que o excesso de zeros acaba mascarando a dependência espacial.

Por fim, o estimador $\widehat{\phi}$ apresentou comportamento assimétrico, fugindo da suposição de Normalidade. E, assim como observado no comportamento de $\widehat{\rho}$, obteve melhor desempenho sob o aspecto de que a proporção de zeros estruturais é menor do que a da distribuição de contagem, com medianas muito mais próximas do verdadeiro valor.

# Capítulo 5

# Aplicação do Modelo em Dados Reais

## 5.1 Descrição dos Dados

Neste capítulo, realizamos uma aplicação do modelo proposto (**QIZDE**) em dados de contagem de novos casos de hanseníase em menores de 15 anos, notificados dentre os anos de 2009 a 2012, no estado do Amazonas, região norte do Brasil, para cada um dos seus 62 municípios. Para uma visualização da distribuição espacial desses novos casos de hanseníase veja a Figura 5.1.

A hanseníase, também conhecida como lepra, é uma doença infecto contagiosa causada por uma bactéria chamada *Mycobacterium leprae*, que foi descoberta em 1873 por um cientista chamado Hansen. É uma doença curável, porém se não for devidamente tratada pode ser preocupante, devido à sua magnitude e seu alto poder incapacitante. Por isso, o acompanhamento epidemiológico e variação geográfica da hanseníase são de grande importância à saúde pública, para seu controle e monitoramento.

O surgimento e desenvolvimento da hanseníase estão relacionados com as condições de vida da população, e seu acompanhamento epidemiológico é realizado por meio do coeficiente de detecção de novos casos, que é obtido dividindo-se o total de novos casos pela população em risco e multiplicando o resultado por 10 mil habitantes. Os países com maior incidência são os menos desenvolvidos ou com condições precárias de higiene e superpopulação.

O número de casos em menores de 15 anos é um indicador que reflete a gravidade do nível endêmico da Hanseníase e a exposição precoce à doença, pois a detecção de casos nessa faixa etária tem relação com a doença detectada recente e focos de transmissão ativos. Por isso, o grupo com menores de 15 anos (grupo de risco) foi considerado o público alvo desta aplicação. Em geral, as taxas elevadas refletem baixos níveis de condições de vida, desenvolvimento socioeconômico e atenção à saúde.

Nos anos de 2009 a 2012 foram notificados 242 novos casos de hanseníase, do total de 4.649.905 menores de 15 anos residentes no estado do AM, com uma taxa de 0,5204 casos

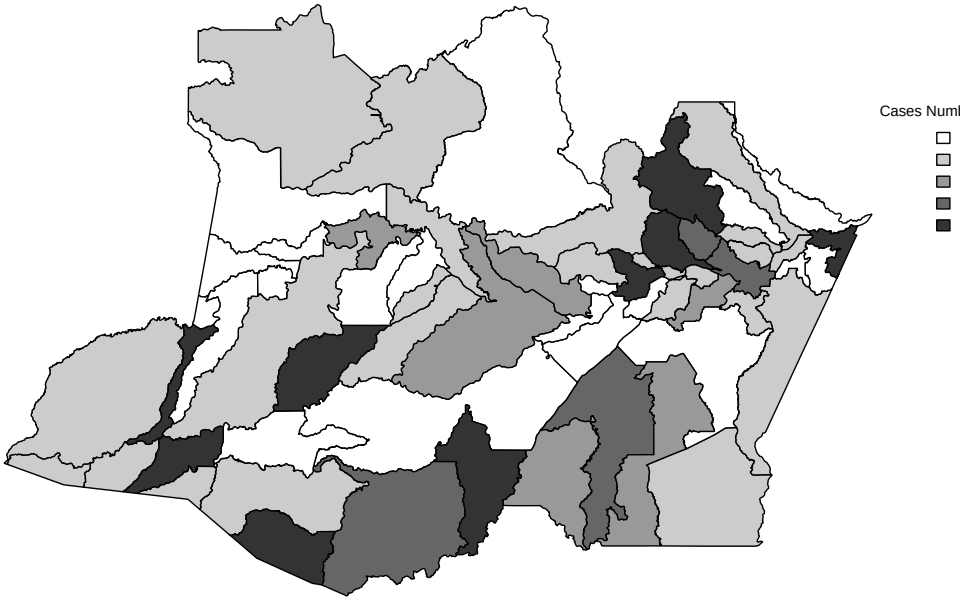


Figura 5.1: Distribuição espacial dos novos casos de Hanseníase, notificados no Estado do Amazonas-Brasil, no período de 2009 a 2012.

para cada 10 mil habitantes, sendo classificada como hiperendêmica, segundo o Ministério da Saúde do Brasil.

Cerca de $32,26\%$ dos municípios não registraram ou não notificaram novos casos, evidenciando a existência de excesso de zeros nos dados. Na Figura 5.1, podemos notar a existência de municípios sem notificações de casos durante 4 anos, que são as áreas em branco. Este fato nos sugeriu a hipótese de que estes zeros podem ter ocorrido estruturalmente devido a subnotificação de casos causada, possivelmente, pelas condições de vida social e econômica destas populações. Por isso, inicialmente foi realizada uma análise preliminar dos dados utilizando um modelo Poisson inflacionado de zeros (**ZIP**) com covariáveis que podem influenciar a ocorrência e o registro de novos casos de hanseníase. A descrição do modelo e resultados dessa análise serão apresentados na próxima seção.

## 5.2 Modelo ZIP para os novos casos notificados de hanseníase no Amazonas

O banco de dados é composto por registros de novos casos de hanseníase, em menores de 15 anos, notificados no período de 2009 a 2012 nos 62 municípios do Estado do Amazonas (fonte: *http://dtr2004.saude.gov.br/ sinanweb/index.php*) e por outras variáveis: taxa padronizada de domicílios com saneamento inadequado de 2010, média de unidades básicas de saúde de 2009 a 2012, índice de desenvolvimento humano de 2010, quantidade de recursos humanos na área de saúde em 2010, distância padronizada dos municípios para a capital (Manaus) e taxa de analfabetismo de 2010.

Inicialmente analisamos esses dados com um modelo para dados inflacionados de zeros, considerando que $Y_i \sim \mathbf{ZIP}(\mu_i, p_i), i = 1, 2..., 62$, representa o número de novos casos de hanseníase registrados ou notificados no $i$-ésimo município do estado do Amazonas. A representação do modelo é dada por:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \log(n_i), \tag{5.1}$$

em que $\log(\mu_i)$ é o preditor linear para a média das contagens positivas e,

$$logit(p_i) = \gamma_0 + \gamma_1 G_{i1} + \gamma_2 G_{i2} + \gamma_3 G_{i3} + \gamma_4 G_{i4} \tag{5.2}$$

é o preditor linear para a probabilidade de ocorrer o zero estrutural, possivelmente devido a subnotificações. Essas covariáveis são descritas como:

$X_{i1}$: é a taxa padronizada de domicílios que possuem saneamento inadequado, da i-ésima área, dada por:

$$X_{i1} = \frac{X_{i1} - \min\{X_{i1} : i = 1, ..., 62\}}{\max\{X_{i1} : i = 1, ..., 62\} - \min\{X_{i1} : i = 1, ..., 62\}},$$

A diferença entre a capital do Estado do Amazonas, Manaus, e os demais municípios no que concerne a taxa de domicílioas com saneamento inadequado é muito grande, por isso adotamos os valores padronizados.

$X_{i2} = G_{i3}$: é a quantidade média de unidades básicas de saúde, no período de 2009 a 2012, da $i$-ésima área;

$X_{i3}$: é o índice de desenvolvimento humano (IDH), em 2010, da $i$-ésima área;

$G_{i1}$: é a quantidade de recursos humanos na área de saúde, em 2010, da $i$-ésima área;

$G_{i2}$: é a distância padronizada da $i$-ésima área para Manaus, dada por:

$$G_{i2} = \frac{d_{im} - \min\{d_{im} : i = 1, ..., 62\}}{\max\{d_{im} : i = 1, ..., 62\} - \min\{d_{im} : i = 1, ..., 62\}},$$

em que $d_{im}$ é a distância da $i$-ésima área para Manaus-AM.

$G_{i4}$: é a taxa de analfabetismo, em 2010, da $i$-ésima área;

$\log(n_i)$: é a variável (*offset*) utilizada para incorporar a heterogeneidade da população, em que $n_i$ corresponde a população média projetada de menores de 15 anos de idade, residentes no estado do Amazonas, no período de 2009 a 2012, na $i$-ésima área (município).

As variáveis $G_{i1}$, $G_{i3}$ e $G_{i4}$ são consideradas covariáveis de registro ou notificações no sentido de que inibem a ocorrência de zeros. A covariável $G_{i2}$ é considerada de subnotificação, pois a capital Manaus concentra mais de 50% de toda população, apresenta condições de recursos humanos e estrutura física (hospitais, clínicas, etc...) que favorecem o diagnóstico da doença. No entanto, devido a extensão territorial do estado, Manaus é muito distante de algumas cidades, das quais o deslocamento só pode ser realizado através de embarcações. Tal fato, pode dificultar a busca pelo diagnóstico da doença gerando o não registro e possibilitando o surgimento de um zero estrutural.

Para análise do modelo, utilizamos no **R** a função "*zeroinfl*"do pacote "*pscl*"(*Package* 'pscl') e os resultados obtidos para modelo $\mathbf{ZIP}(\mu_i, p_i)$ são apresentados na Tabela 5.1. Como pode ser observado na Tabela 5.1, de acordo com essa modelagem, há apenas duas covariáveis significativas para a média do modelo de contagem, que são $X_{i2}$ (média de unidades básicas de saúde) e $X_{i3}$ (índice de desenvolvimento humano), e para a inflação de zeros nenhuma das covariáveis avaliadas foi significativas.

Note que há inconsistência de informação, por exemplo, a estimativa para $X_{i2}$ foi negativa, implicando na seguinte relação, de que quanto mais unidades básicas de saúde existirem em determinada área, espera-se que menor seja a média de novos casos notificados de hanseníase, resultado este que é o oposto do esperado. Observa-se também, que nenhuma das covariáveis preditoras para o excesso de zeros, foram significativas. No entanto, o índice de inflação de zeros calculado com esse modelo ajustado, definido no Capítulo 2, foi estimado em $zi = 0.189$, o qual é maior que zero, implicando na presença da inflação de zeros nos dados. Podemos notar a existência de municípios sem notificação, o que não faz sentido, pois o esperado é que ao menos uma dessas covariáveis implicasse na inflação de zeros. Baseados na razão entre a soma de quadrados dos resíduos e os graus de liberdade do modelo **ZIP** utilizado, obtivemos uma estimativa para sobredispersão igual a 8.67. O que nos leva a crer, que essa modelagem não consegue capturar adequadamente a inflação de zeros e a sobredispersão, possivelmente porque não levam em consideração a existência da relação de dependência entre os municípios vizinhos.

Para verificar uma possível estrutura de dependência espacial no dados, utilizou-se no **R** o correlograma dos resíduos do modelo escolhido, gerado pela função "*correlog*"do pacote "ncf", que utiliza o Índice de Moran, para o caso univariado, e a estatística de Mantel centrada, para o caso multivariado (Epperson (1993); Bjørnstad *et al.* (1999); Bjørnstad & Falck (2001)). O gráfico desses resíduos apresenta em sua estrutura o coeficiente de

Tabela 5.1: Resultados para o modelo **ZIP**$(\mu_i, p_i)$, gerados pela função "*zeroinfl*", para os novos casos de hanseníase, notificados no Estado do Amazonas-Brasil-2009/2012.

| Count model coefficients (poisson with log link): | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
| (Intercept) | -9.834055 | 0.802098 | -12.260 | < 2e-16 | *** |
| $X_{i1}$ | -0.092210 | 0.536577 | -0.172 | 0.86356 | |
| $X_{i2}$ | -0.004258 | 0.001407 | -3.026 | 0.00248 | ** |
| $X_{i3}$ | 2.945614 | 1.253587 | 2.350 | 0.01879 | * |
| Zero-inflation model coefficients (binomial with logit link): | | | | | |
| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
| (Intercept) | 303.08 | 289.47 | 1.047 | 0.295 | |
| $G_{i1}$ | -10.75 | 10.27 | -1.046 | 0.295 | |
| $G_{i2}$ | 344.99 | 331.88 | 1.040 | 0.299 | |
| $G_{i3}$ | -19.77 | 19.11 | -1.035 | 0.301 | |
| $G_{i4}$ | -12.55 | 12.02 | -1.044 | 0.297 | |

dependência espacial, no eixo das ordenadas, e a distância média de classe, no eixo das abscissas, onde o primeiro ponto do gráfico é referente a média de distância das cidades mais próximas, que por sua vez foi a referência utilizada para as análises. Um decaimento no valor do coeficiente em função da distância, sugere uma dependência espacial nos dados que deve ser incorporada pelo modelo.

(a) (b)

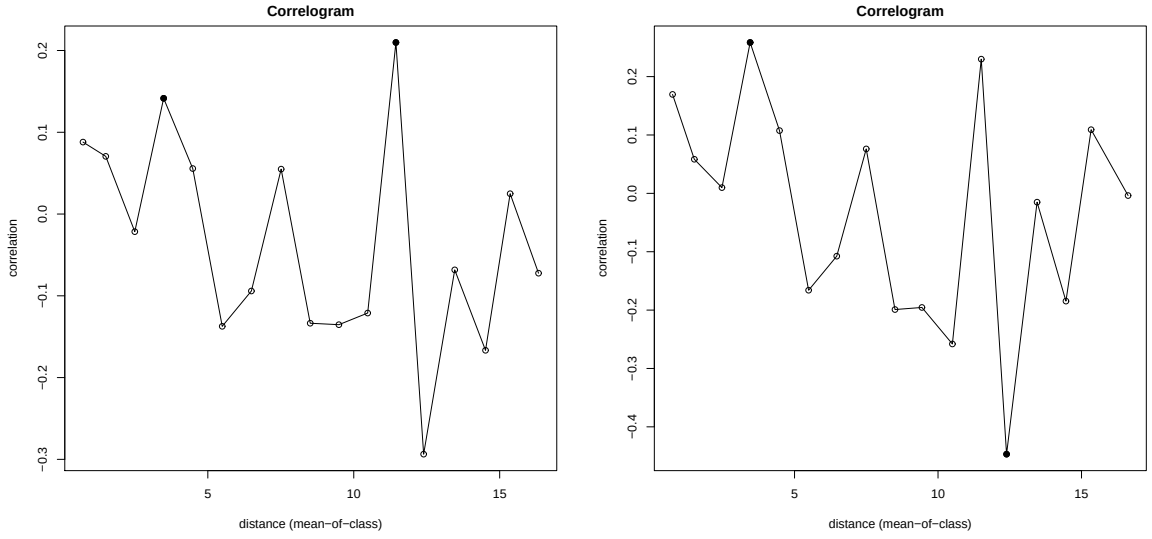


Figura 5.2: Correlograma espacial dos resíduos do modelo ajustado para novos casos de Hanseníase no Estado do Amazonas-Brasil, nos anos de 2009 a 2012.

A Figura 5.2 apresenta o correlograma dos resíduos do modelo, referente aos novos casos de hanseníase, notificados no período de 2009 a 2012, onde o lado (a) apresenta o correlograma dos resíduos com todas as observações e o lado (b) sem os resíduos referentes

às cidades que não apresentaram nenhum caso. Note, que a inflação de zeros mascara um pouco dependência espacial em (a), mas os dados de contagem positiva possuem sim tal dependência, como vemos claramente em (b), cuja correlação é interpretada como positiva entre áreas vizinhas. E ainda, o segundo correlograma (b) mostra que os vizinhos mais próximos possuem correlação espacial positiva de $\cong 0,2$. Esse comportamento gráfico do tipo zig-zag para os resíduos, pode ser um indicador de que existam pequenos aglomerados de municípios com excesso de zeros ou pequenos aglomerados de contagens positivas.

Portanto, como as três características abordadas: inflação de zeros, sobredispersão e dependência espacial, foram identificadas nesses dados de hanseníase, os modelos tradicionais existentes, bem como os modelos para dados inflacionados de zeros, ambos no contexto frequentista, não ajustariam bem esses dados por não conseguirem modelar ao mesmo tempo as três características. Então, na próxima seção, aplicaremos o modelo (**QIZDE**) o qual propõe-se a modelar essas três características conjuntamente, com abordagem frequentista, e discutiremos os resultados obtidos.

## 5.3 Descrição do Modelo Proposto

Considere a variável de interesse sendo $Y_i, i = 1, ..., 62$, que corresponde ao número de novos casos de hanseníase notificados no $i$-ésimo município do estado do Amazonas, no período de 2009 a 2012, cujo mapa do estado é composto de 62 municípios. Como observado na seção anterior, os dados referentes a esses novos casos de hanseníase notificados apresentam inflação de zeros, sobredispersão e as áreas vizinhas possuem dependência espacial. Vamos assumir, ainda, que $Y_i$ pertença à classe de modelos com quase verossimilhança completa. Sendo assim, pela definição 2 dada na Seção 3.2, podemos representar $Y_i$ por um modelo quase inflacionado de zeros com dependência espacial **QIZDE**. Então, $Y_i \sim \mathbf{QIZDE}(p_i, \mu_i, \phi_i, \rho)$, $i = 1, 2..., 62$, em que $p_i$ é a probabilidade de ocorrer um zero estrutural, subnotificado, $\mu_i$ é a média de novos casos de hanseníase notificados, $\phi_i$ é a sobredispersão e $\rho$ é o parâmetro que mede o grau de dependência espacial entre municípios vizinhos.

Para a modelagem dos dados, vamos adotar o modelo **QIZDE** Poisson, exemplificado na Seção 3.1.1, com função de quase-verossimilhança completa dada pela expressão (3.6), se houverem casos notificados no i-ésimo município, e pela expressão (3.7), se não houver nenhum caso notificado. A partir dessas definições, para encontrarmos os estimadores é necessário definir cada componente das funções quase-escore do modelo proposto (ver expressões (3.16), (3.17) e (3.18)), de acordo com os dados utilizados neste estudo.

Adotamos o modelo **QIZDE** Poisson, com as equações quase-escore descritas na Seção 3.2. Para a modelagem da inflação de zeros, utilizamos a função de ligação logit, em que $logit(p_i) = G_i^T \gamma$, com matriz de covariáveis $G_i$ e vetor de parâmetros de regressão $\gamma$,

resultando na equação quase-escore (3.19). Assumimos a relação de dependência espacial apenas na média de notificação de novos casos de hanseníase, para modelar a inflação de zeros, consideramos a matriz identidade $I$ no lugar da matriz de correlação espacial $\mathbf{R}_\gamma(\rho)$, resultando na expressão definida no modelo para independência (ver Seção 3.1). Assim temos que:

$$G_i^T \gamma = \gamma_0 + \gamma_1 G_{i1} + \gamma_2 G_{i2} + \gamma_3 G_{i3} + \gamma_4 G_{i4}, \tag{5.3}$$

cujas covariáveis $G_{i\cdot}$ são exatamente as mesmas definidas anteriormente, na aplicação com o modelo **ZIP**, Seção 5.2.

Para a modelagem da média de novos casos de hanseníase notificados, consideramos a função logarítmica como sendo a de ligação, onde $\log(\mu_i) = X_i^T \beta$, com matriz de covariáveis $X_i$ e vetor de parâmetros de regressão $\beta$, cuja função quase-escore é dada pela expressão (3.20). Assim temos que:

$$X_i^T \beta = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \log(n_i), \tag{5.4}$$

em que as covariáveis $X_{i\cdot}$, também, foram definidas na aplicação com o modelo **ZIP** (Seção 5.2).

Em sua composição, a função quase escore (3.20) conta com a matriz de correlação espacial $\mathbf{R}_\beta(\rho)$, definida em (3.22), em que a matriz $\mathbf{W}$, que define a estrutura espacial de vizinhança, será composta por elementos $w_{ij} = 1$, se os municípios $i$ e $j$ forem vizinhos, e $w_{ij} = 0$, caso contrário. E a matriz de pesos $\mathbf{M}$, referente à essa estrutura de vizinhança, terá como i-ésimo elemento da diagonal o $w_{i+}$, que corresponde à soma de quantos vizinhos o i-ésimo município possui, ou seja, equivale à soma desses zeros e uns em cada linha da matriz $\mathbf{W}$. Com isso, obtemos a esperança condicional da ocorrência de casos no i-ésimo município dada a sua vizinhança (ver expressão 3.23). Desta forma, substituindo adequadamente as matrizes de covariáveis ($\mathbf{X}$ e $\mathbf{G}$) e de correlação espacial ($\mathbf{R}$), nas expressões (3.17), (3.16) e (3.18), os estimadores de $(\beta, \gamma, \phi)$ são obtidos, respectivamente, a partir das seguintes **GEE**'s, em sua forma matricial:

$$\begin{aligned}
\mathbf{X}^T(\mathbf{M}^{-1} - \rho\mathbf{W})\mathbf{U}(\mathbf{Y} - \mu) &= 0, & (5.5)\\
\mathbf{G}^T(\mathbf{U} - \mathbf{p}) &= 0, & (5.6)\\
diag((\mathbf{1} - \mathbf{u})(D(\mathbf{y} - \mu) - \phi)) &= 0, & (5.7)
\end{aligned}$$

Para os resíduos obtidos com os $\widetilde{y_i}$ pertencente ao conjunto definido em (3.28), defina o vetor de resíduos como:

$$\widetilde{\mathbf{r}} = [g(\widetilde{y_i}) - g(\mu_i), ..., g(\widetilde{y_L}) - g(\mu_L)]. \tag{5.8}$$

Então, usando a expressão (3.29) a **GEE** para $\rho$ pode ser escrita matricialmente da seguinte forma:

$$\widetilde{\mathbf{r}}^T \mathbf{MW}(\mathbf{I} - \rho \mathbf{MW})\widetilde{\mathbf{r}}. \tag{5.9}$$

Com as funções quase-escore definidas, a estimação dos parâmetros de regressão é realizada de forma iterativa via algoritmo **ES** (ver seção 2.6), com dois passos de esperança (passo **E**) e solução (passo **S**) das funções quase-escore, definidas acima, que nada mais são que **GEE**'s. Todos os estimadores foram definidos na seção 3.2.1, onde os estimadores de $\beta$, $\gamma$, $\phi$ e $\rho$ são obtidos, respectivamente, pelas expressões (3.26), (3.27), (3.32) e (3.30).

A construção dos intervalos de confiança *Bootstrap*-t (**ICBt**) dos parâmetros, conforme descrito na seção 2.7.1, foi realizada para verificar a significância dos estimadores no modelo. Geramos 1000 reamostras *Bootstrap* do modelo **QIZDE**, a partir das estimativas de cada parâmetro do modelo proposto, com $\mathbb{P}(\mu_i(\hat{\beta}))$ sendo uma Poisson$(\mu_i(\hat{\beta}))$, afim de avaliar a distribuição empírica desses estimadores (ver seção 3.3). As equações quase-escore empíricas *Bootstrap* para $\beta$, $\gamma$, $\lambda$ e $\rho$ são, respectivamente, dadas por:

$$\begin{aligned}
\mathbf{X}^T(\mathbf{M}^{-1} - \widehat{\rho}\mathbf{W})\mathbf{U}^*(\mathbf{Y}^* - \widehat{\mu}) &= 0, && (5.10)\\
\mathbf{G}^T(\mathbf{U}^* - \widehat{\mathbf{p}}) &= 0, && (5.11)\\
diag((\mathbf{1} - \mathbf{u}^*)(\hat{D}(\mathbf{y}^* - \widehat{\mu}) - \widehat{\phi})) &= 0, && (5.12)\\
\tilde{\mathbf{r}}^T \mathbf{MW}(\mathbf{I} - \widehat{\rho}\mathbf{MW})\tilde{\mathbf{r}} &= 0. && (5.13)
\end{aligned}$$

Usando o modelo proposto, os resultados desta aplicação estão descritos na próxima seção.

## 5.4 Resultados

Aplicamos o modelo proposto aos dados de hanseníase, conforme descrito na seção anterior, para estimar os parâmetros de regressão. Foram consideradas 1000 reamostras *bootstrap*, afim de construir os intervalos de confiança *Bootstrap*-t para cada parâmetro, cujos resultados estão apresentados na Tabela (5.2), e avaliar a qualidade desses estimadores.

Vale ressaltar que o algoritmo proposto foi bastante sensível a valores iniciais, no que se refere a $\beta$, pois foi por método de tentativa a escolha do valor inicial na execução do processo de estimação, que no caso foi $\beta^{(0)} = (0.01, 0.01, 0.01, 0.01)$. Tentamos utilizar o valor estimado a partir da função "*zeroinfl*", mas não obtivemos êxito. Os gamas foram inicializados com as estimativas geradas pela função "*zeroinfl*", $\gamma^{(0)} = (303.08, -10.75, 344.99, -19.77, -12.55)$. O $\phi$ inicial ($\phi^{(0)} = 8.6681$) foi o *deviance* calculado do modelo ajustado pela mesma função e o $\rho$ inicial ($\rho^{(0)} = 0.5$) é o mesmo utili-

Tabela 5.2: Resultados para o modelo proposto, para os novos casos notificados de hanseníase de 2009 a 2012, com valor estimado, erro padrão e intervalo de confiança Bootstrap-t.

| Parâmetro | Estimativa | Erro Padrão | ICBt 2.5% | ICBt 97.5% |
|---|---|---|---|---|
| $\beta_0$ | -1.7386 | 1.7984 | -4.5785 | 2.2804 |
| $\beta_1$ | -1.4049 | 1.0706 | -3.5171 | 0.6529 |
| $\beta_2$ | 0.0078 | 0.0026 | 0.0025 | 0.0126 |
| $\beta_3$ | 5.9801 | 2.7618 | 0.0350 | 10.4869 |
| $\gamma_0$ | 734.7286 | 24.1065 | 698.0866 | 760.6774 |
| $\gamma_1$ | -26.1965 | 0.8624 | -27.1742 | -24.9212 |
| $\gamma_2$ | 844.7811 | 28.3961 | 804.2731 | 876.9653 |
| $\gamma_3$ | -48.4173 | 1.6536 | -50.2068 | -46.0436 |
| $\gamma_4$ | -30.6969 | 1.0213 | -31.8470 | -29.2166 |
| $\rho$ | 0.5149 | 0.3477 | 0.1982 | 1.0000 |
| $\phi$ | 2.8006 | 0.4266 | 1.7409 | 3.4241 |

zado nos estudos de simulação (Capítulo 4), valor muito comum na literatura (Yasui & Lele (1997)).

Analisando os resultados da Tabela 5.2, no que se refere ao vetor *beta*, observamos que apenas dois dos intervalos **ICBt**, para $\widehat{\beta}_2$ e $\widehat{\beta}_3$, não contiveram o zero, ou seja, duas das covariáveis analisadas foram significativas para a média, que são $X_{i2}$ e $X_{i3}$. Sendo assim, podemos dizer que a quantidade média de unidades básicas de saúde ($X_{i2}$) e o índice de desenvolvimento humano ($X_{i3}$) influenciam positivamente na média de novos casos notificados, ou seja, quanto maior for a quantidade dessas unidades de saúde e maior o valor desse índice, espera-se que maior seja a média de novos casos notificados. Tal resultado parece ser bastante razoável, considerando que a unidade básica é um dos estabelecimentos de saúde no qual se faz a notificação da doença e um valor alto do IDH indica maior desenvolvimento sócio-econômico do município, o que implica em uma maior capacidade de notificação por ter mais estabelecimentos de saúde, acarretando no aumento da média de notificações.

Um fato interessante é que o parâmetro $\beta_2$, referente à covariável $X_{i2}$, na análise realizada com o modelo **ZIP** obteve estimativa negativa, com sentido contraditório ao estimado pelo modelo proposto **QIZDE**, que, por sua vez, o estimou positivo. Neste caso, a interpretação mais razoável dessa covariável seria a do modelo proposto **QIZDE**. Isso pode ser em decorrência do modelo **QIZDE** levar em consideração a relação de dependência espacial entre os municípios, na distribuição de contagem, conseguindo obter melhor interpretação dessa covariável.

Avaliando os resultados dos estimadores para a inflação de zeros, notamos que todas as covariáveis escolhidas nessa modelagem foram significativas, pois seus respectivos intervalos (**ICB-t**) não contiveram o zero. Então, com relação ao excesso de zeros, podemos dizer que a quantidade de recursos humanos na área de saúde ($G_{i1}$), a quantidade média de

unidades básicas de saúde ($G_{i3}$) e a taxa de analfabetismo ($G_{i4}$) influenciam negativamente, ou seja, quanto maior o quadro de recursos humanos na saúde, mais unidades básicas de saúde existirem e maior a taxa de analfabetismo, menor será o número de municípios sem notificação de novos casos de hanseníase, culminando na redução da inflação de zeros. Pois, se há mais recursos humanos e unidades básicas de saúde para realizar essas notificações, espera-se que aumente a quantidade de notificações, reduzindo o número de zeros estruturais. Tendo em vista que a taxa de analfabetismo é uma variável que interfere na ocorrência de novos casos, então se um município tiver maior taxa de analfabetismo, significa que ele possui maior vulnerabilidade à doença, com isso espera-se que ele tenha maior ocorrência de novos casos de hanseníase, reduzindo a inflação de zeros.

Outra covariável significativa foi a distância padronizada para Manaus ($G_{i2}$), que influencia positivamente na inflação de zeros, ou seja, quanto mais distante de Manaus for o município maior será a probabilidade de ocorrência do zero estrutural. O que é razoável, tendo em vista que os municípios mais distantes da capital, Manaus, são menos desenvolvidos, com maior dificuldade de acesso e deslocamento. Assim, quanto mais distante da capital for esse município, espera-se que não ocorra a notificação dos casos, não pelo fato de que realmente não exista algum novo caso da doença, mas pela dificuldade de acesso dos agentes de saúde à esses lugares ou pela dificuldade de deslocamento do portador da doença à unidade de saúde notificadora, pois grande parte dos descolamentos no estado do Amazonas são realizados via fluvial.

Note que todas as covariáveis, consideradas na modelagem do excesso de zeros, obtiveram estimativas com o mesmo sinal em ambos modelos **ZIP** e **QIZDE**, culminando em interpretações similares, pois ambos referem-se à inflação de zeros. No entanto, essas covariáveis foram significativas apenas para o modelo proposto **QIZDE**, cujos erros padrão de $\widehat{\gamma}$ foram bem menores do que os do modelo **ZIP**.

O parâmetro que mede a dependência espacial ($\rho$) foi estimado em 0.5149, implicando em correlação espacial positiva entre as áreas vizinhas, ou seja, é esperado que municípios próximos a outros, que apresentem grande quantidade de casos, tenham mais casos também.

A sobredispersão ($\phi$) foi estimada em 2.8006, apontando a existência de sobredispersão nos dados. A sobredispersão em dados referentes a processos espaciais de contagem é muito comum, quando existe uma grande heterogeneidade populacional, o que de fato ocorre.

Portanto, o modelo de regressão estimado, com as covariáveis significativas, para os gamas correspondentes à inflação de zeros (5.3), é dado por:

$$G_i^T \widehat{\gamma} = 734.7286 - 26.1965 G_{i1} + 844.7811 G_{i2} - 48.4173 G_{i3} - 30.6969 G_{i4},$$

e para os betas correspondentes à média de novos casos notificados de hanseníase, dado por:

$$X_i^T \widehat{\beta} = 0.0078X_{i2} + 5.9801X_{i3} + \log(n_i),$$

cuja dependência espacial foi estimada em $\widehat{\rho} = 0.5149$ e a sobredispersão foi estimada em $\widehat{\phi} = 2.8006$.

O valor esperado e a variância estimados, de novos casos notificados para o i-ésimo município, definidos pela expressão (2.4), são descritos respectivamente por:

$$\widehat{E(Y_i)} \approx (1 - \widehat{p}_i)\widehat{\mu}_i(\widehat{\beta})$$

e

$$\widehat{Var(Y_i)} \approx (1 - \widehat{p}_i)\widehat{\mu}_i(\widehat{\beta}) + \widehat{p}_i(1 - \widehat{p}_i)\widehat{\mu}_i(\widehat{\beta})^2,$$

em que $\widehat{p}_i = \frac{\exp(G_i^T \widehat{\gamma})}{1+\exp(G_i^T \widehat{\gamma})}$ e $\widehat{\mu}_i(\widehat{\beta}) = \exp(X_i^T \widehat{\beta})$.

Podemos avaliar ainda a distribuição desses estimadores, pois a teoria sugere que eles tenham comportamento assintoticamente Normal. Contudo, essa suposição nem sempre mantém-se válida, tendo em vista que os dados possuem características bem diferentes da distribuição Normal.

(a) (b)

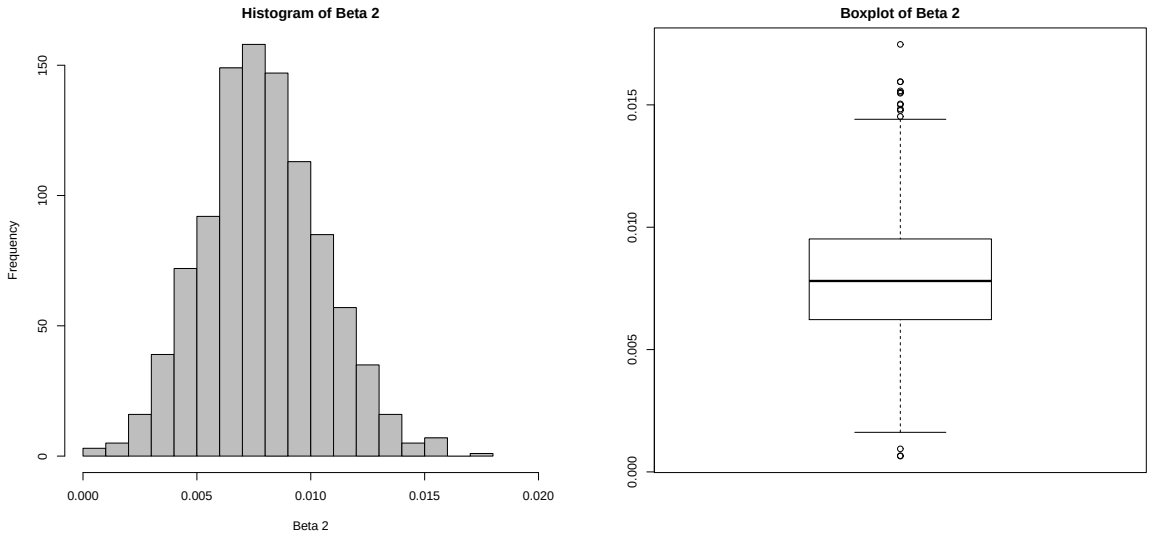


Figura 5.3: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\beta}_2$.

O comportamento dos estimadores pode ser visualizado através de gráficos, como por exemplo o de $\widehat{\beta}_2$ a partir da Figura 5.3, o qual apresenta uma forma simétrica em torno de 0.0075 no lado (a), corroborando à suposição de normalidade assintótica, com a presença de *outliers* em ambos os lados (b). O surgimento desses *outliers* pode ser ocasionado por haver mais unidades básicas de saúde em municípios desenvolvidos do que nos subdesenvolvidos.

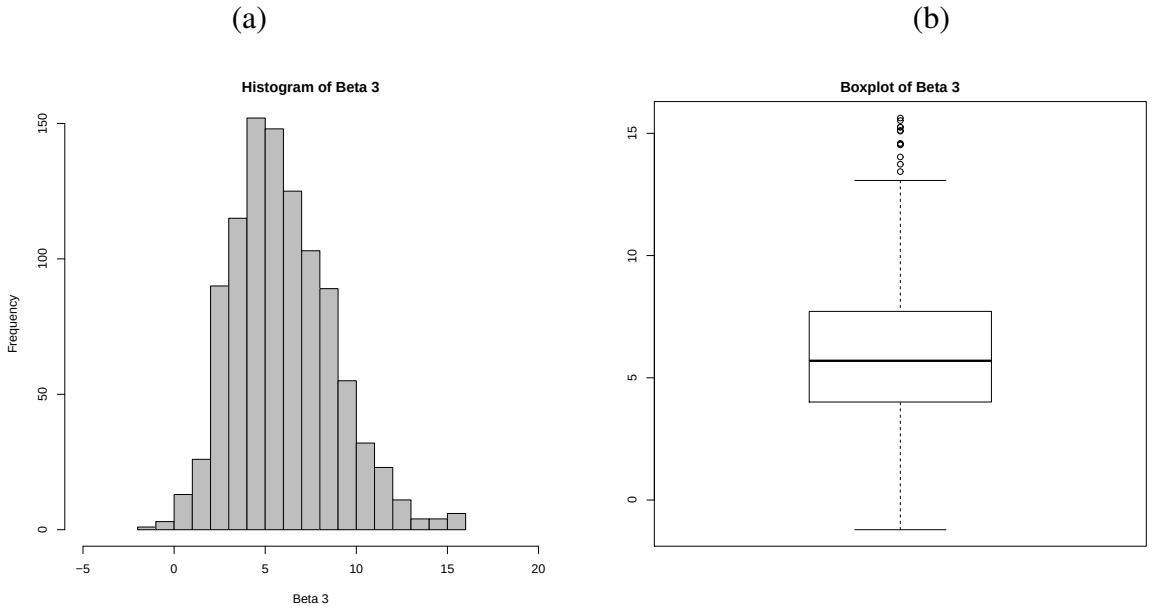


Figura 5.4: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\beta}_3$.

O estimador $\widehat{\beta}_3$ (ver Figura 5.4) apresentou leve assimetria à esquerda, com mediana em torno de 6 em (a), afastando-se um pouco da suposição de normalidade, e ainda há presença de *outliers* à direita em (b). Podemos interpretar essa assimetria e os pontos discrepantes, sendo decorrentes da existência de alguns municípios com índice de desenvolvimento humano muito superior ao da maioria, ocorrência bastante provável no estado do Amazonas, pois uma pequena parte de seus municípios são desenvolvidos.

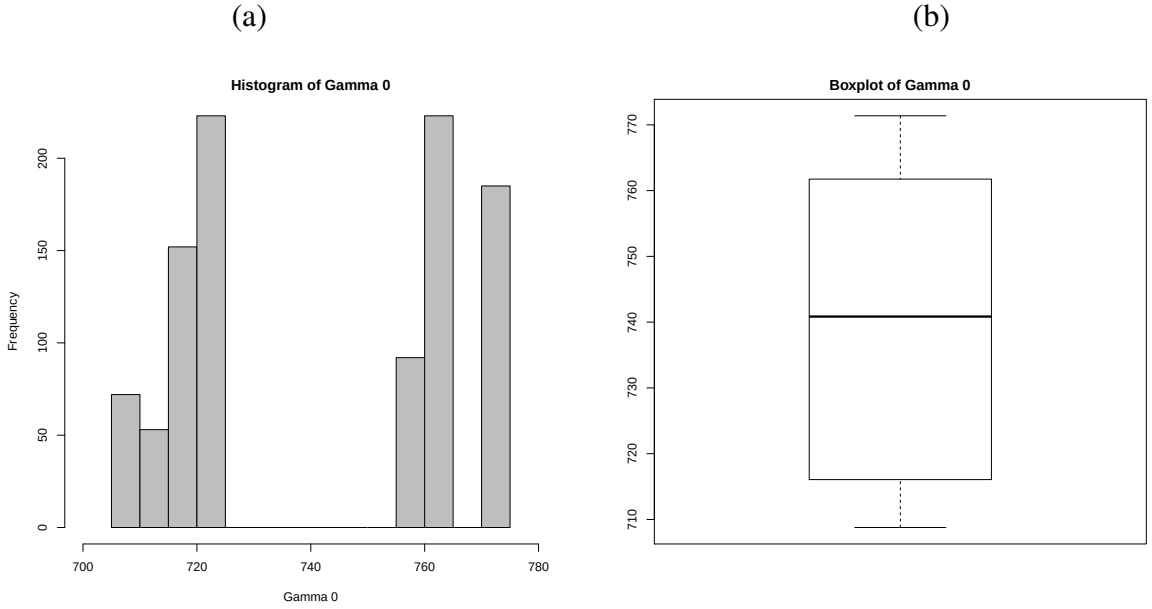


Figura 5.5: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\gamma}_0$.

Todos estimadores dos gamas, apresentaram claramente um comportamento bimodal, afastando-se completamente da suposição de normalidade assintótica. Esse comportamento

(a) (b)

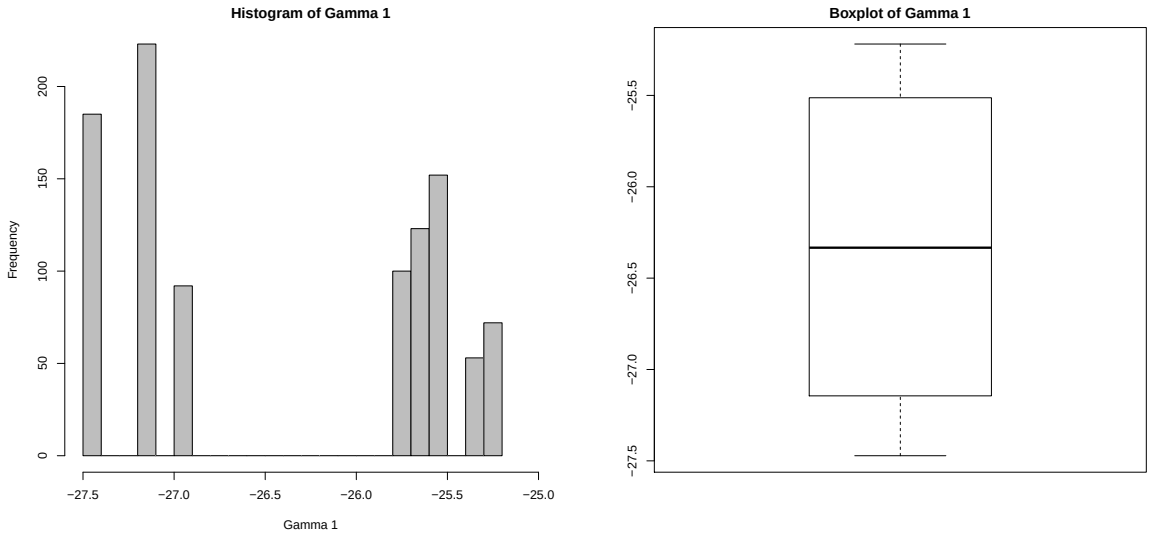


Figura 5.6: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\gamma_1}$.

(a) (b)

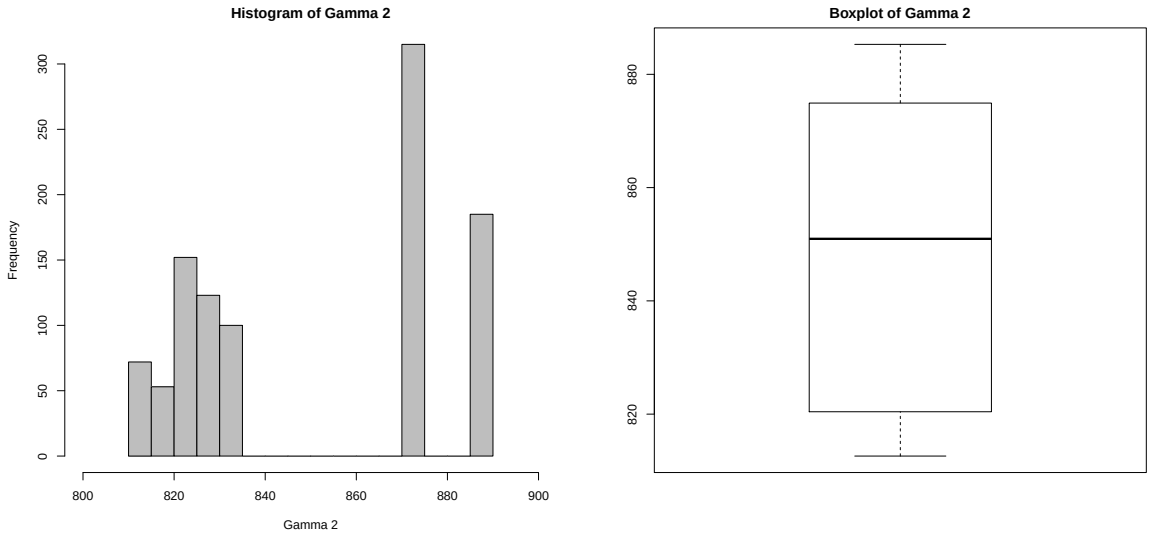


Figura 5.7: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\gamma_2}$.

pode ser decorrente do fato de não conseguirmos identificar a natureza dos zeros, se do excesso ou da distribuição de contagem. Os estimadores $\widehat{\gamma_0}$ (ver Figura 5.5) e $\widehat{\gamma_2}$ (ver Figura 5.7), apresentaram assimetria à direita. No caso de $\widehat{\gamma_2}$, isso pode ocorrer devido a grande distância da maioria dos municípios do estado do Amazonas para a capital. Os estimadores $\widehat{\gamma_3}$ (Figura 5.8) e $\widehat{\gamma_4}$ (Figura 5.9) apresentaram assimetria à esquerda, que pode ser ocasionada por mais da metade dos municípios possuírem poucas ou nenhuma unidade básica de saúde e taxas de analfabetismo acima da média.

Na Figura 5.10, visualizamos o comportamento suavemente assimétrico do estimador $\widehat{\rho}$.

(a) (b)

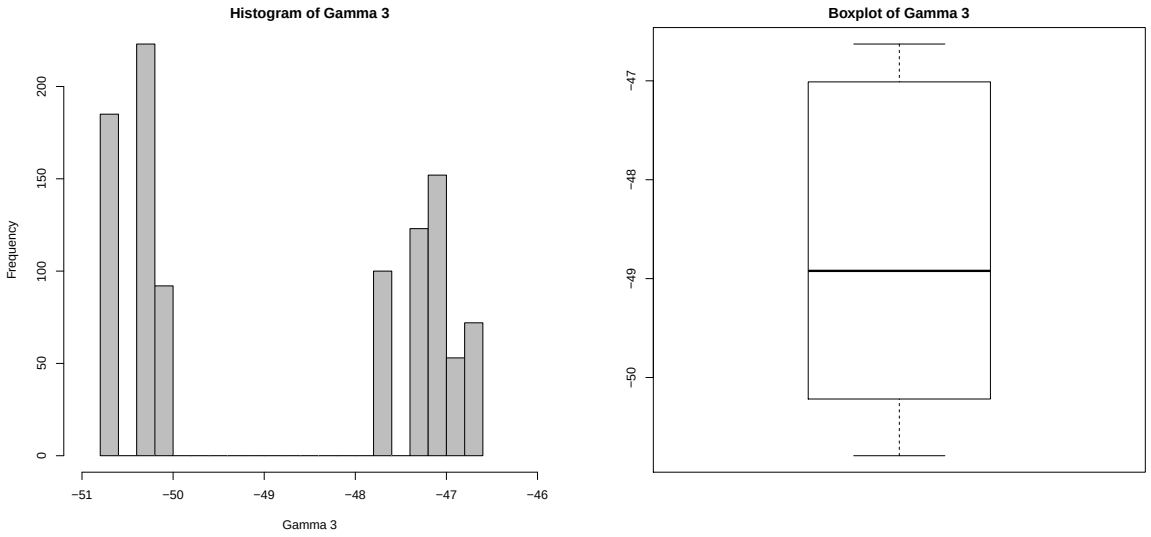


Figura 5.8: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\gamma_3}$.

(a) (b)

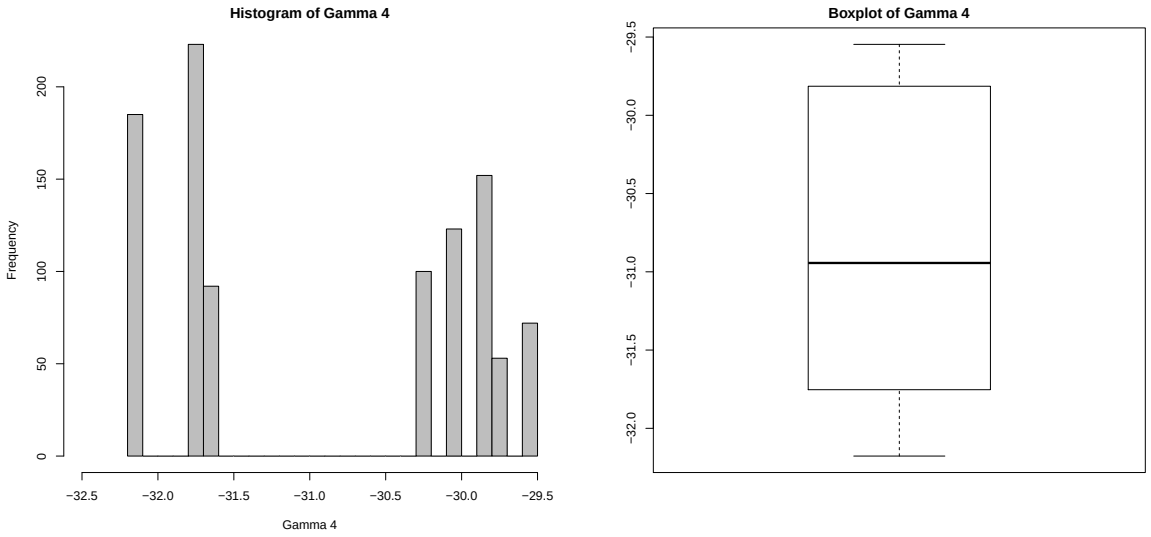


Figura 5.9: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\gamma_4}$.

E, por fim, o estimador $\widehat{\phi}$ (ver Figura 5.11) apresentou simetria e formato de sino em sua distribuição (a), ratificando a suposição de normalidade assintótica de sua distribuição.

(a) (b)

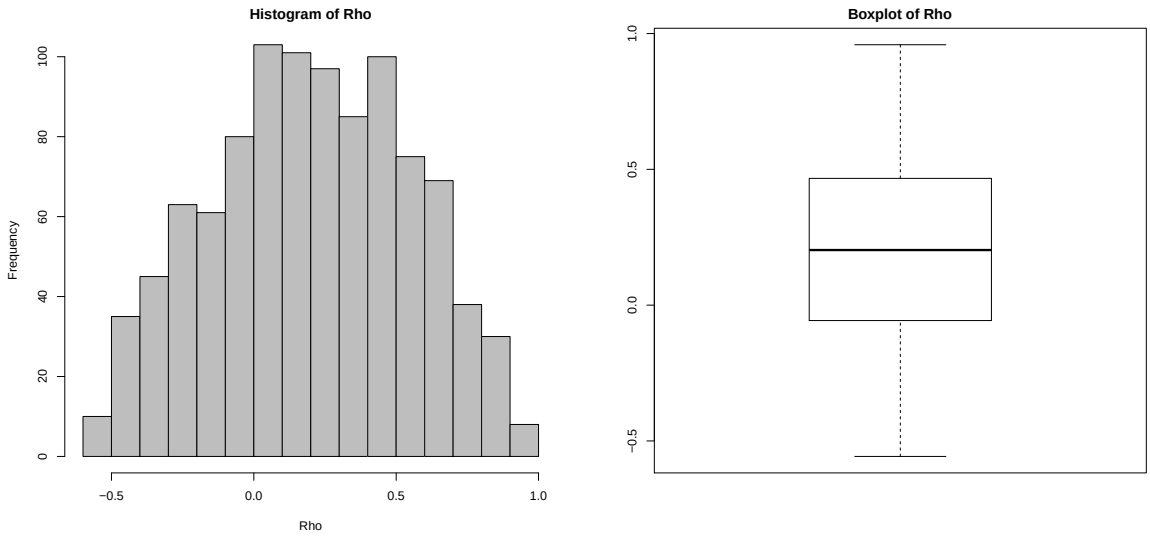


Figura 5.10: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\rho}$.

(a) (b)

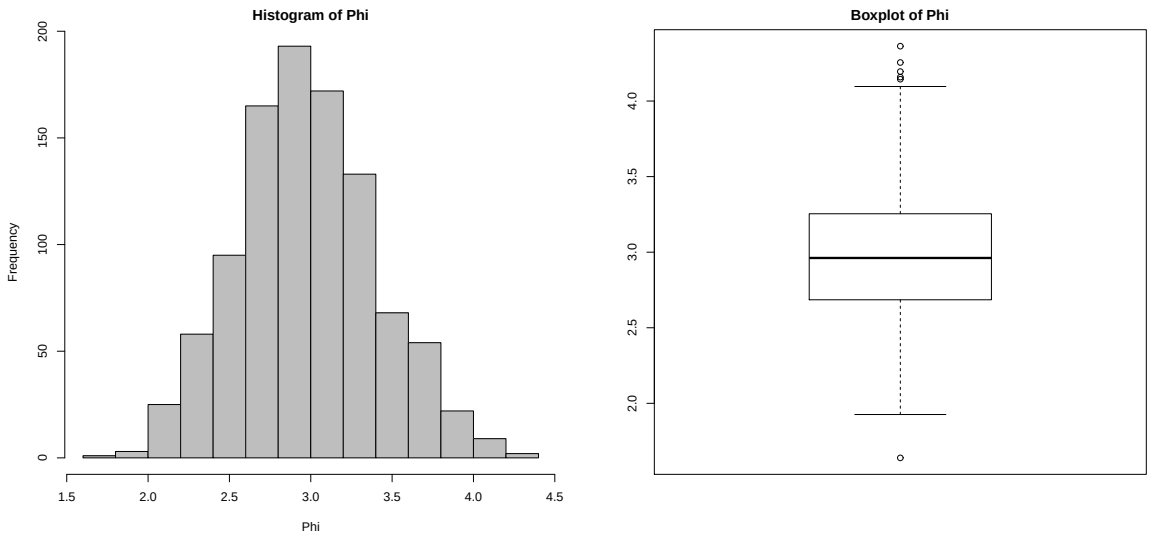


Figura 5.11: Histograma (a) e Boxplot (b) das estimativas de $\widehat{\phi}$.

# Capítulo 6

# Considerações Finais

## 6.1 Principais Conclusões

A abordagem proposta nesta dissertação é interessante e inovadora, do ponto de vista clássico, por oferecer a possibilidade da modelagem conjunta de três características muito comuns: inflação de zeros, sobredispersão e dependência espacial. Sendo de grande relevância, no que se refere a aplicabilidade em dados espaciais de contagem reais. Principalmente na atualidade, com o surgimento de diversos dados com essa natureza devido ao avanço tecnológico.

Desenvolvemos a modelagem para dados de contagem que possuem inflação de zeros, sobredispersão e dependência espacial, através de uma quase verossimilhança inflacionada de zeros capaz de acomodar a dependência espacial, e aplicamos essa nova metodologia em dados reais, realizamos estudos de simulação *Bootstrap* para avaliar a qualidade dos estimadores propostos e discutimos todos os resultados obtidos.

No que diz respeito à aplicação do modelo proposto **QIZDE** em dados reais (Capítulo 5), ele conseguiu evidenciar todas as características abordadas como pertinentes aos dados de novos casos de hanseníase, notificados entre 2009 e 2012, mostrou-se mais adequado na modelagem desses dados do que o modelo para dados inflacionados de zeros (**ZIP**), pois este último não conseguiu evidenciar tais características.

Um dos problemas levantados na aplicação, consistiu em identificar se realmente o problema de subnotificação estaria ocasionando a inflação de zeros, pois as covariáveis utilizadas na modelagem estão associadas à tal problema. Nenhuma dessas covariáveis foram avaliadas como significativas pelo modelo **ZIP**, porém o modelo proposto **QIZDE** detectou que todas eram significativas na modelagem do excesso de zeros, concluindo-se que a subnotificação está causando o surgimento desses zeros excedentes. Além disso, o modelo proposto (**QIZDE**) ratificou a presença da sobredispersão e da dependência espacial positiva entre os municípios vizinhos, o que foi evidenciado na análise preliminar dos dados.

Na modelagem da média de novos casos notificados de hanseníase, o modelo proposto

(**QIZDE**) foi mais coerente que o modelo **ZIP**, no que refere-se à influência da média de unidades básicas de saúde na notificação, seus resultados mostraram que com o aumento da média de unidades básicas a média de notificações também aumentaria, conclusão inversa apresentada pelo modelo **ZIP**.

À luz da aplicação em dados reais (Capítulo 5) e do estudo de simulação (Capítulo 4), no que diz respeito à qualidade dos estimadores dos parâmetros do modelo proposto (**QIZDE**), obtivemos excelentes resultados no que tange a não tendenciosidade e precisão deles. Percebemos, também, que a estimação dos parâmetros, principalmente a dos betas, são muito sensíveis a valores iniciais.

Um comentário importante a ser feito é a respeito dos estimadores de $\beta$, $\phi$ e $\rho$ que, no estudo simulado, apresentaram melhor desempenho na ocasião em que a proporção de zeros estruturais era menor do que a vinda da distribuição de contagem, podendo ser decorrente do fato de que a inflação de zeros esteja mascarando de certa forma o verdadeiro valor desses parâmetros. E os gamas apresentaram sempre maior precisão em quaisquer aspectos. Portanto, o modelo **QIZDE** aparenta melhor desempenho quando os dados possuem proporção de zeros estruturais menor do que a da distribuição de contagem.

Um detalhe importante do modelo **QIZDE** é que a dependência espacial foi inserida no processo de estimação, através de uma **GEE** e pela configuração de um modelo **CAR** na matriz de correlação espacial $\mathbf{R}_\beta(\rho)$, simplificando o processo de estimação. Outro ponto interessante foi a utilização do algoritmo **ES** no processo de estimação, que resulta na solução das **GEE**'s.

Uma das mais importantes vantagens dessa abordagem é a possibilidade de modelar conjuntamente essas três características: dependência espacial, excesso de zeros e sobredispersão, tendo em vista que não há trabalhos similares na literatura, no contexto clássico. Porém, há um trabalho recente com abordagem similar, mas sob o ponto de vista bayesiano (Monod (2012)). Outra, é a facilidade no tratamento dos dados por causa da quase verossimilhança, que não exige a definição de uma função de probabilidade, bastando apenas definir os dois primeiros momentos como em um **GLM**, em que as funções quase escore nada mais são que **GEE**'s, definidas no processo de estimação. E, ainda, a modelagem da matriz de correlação espacial como um modelo **CAR**, evitando difíceis inversas de matrizes, facilitando o cálculo dos estimadores e o processo de estimação, com uma forma simples de matriz inversa.

Algumas desvantagens, também, foram percebidas no desenvolvimento deste trabalho. Como por exemplo, não termos a garantia de que o algoritmo **ES** convirja para um estimador não-viesado, consistente e assintoticamente Normal, sendo apenas uma possibilidade. A dificuldade na escolha do valor para inicializar o processo de estimação, pois dependendo dessa escolha o algoritmo tende a não convergir. E, o fato de não conseguirmos estimar bem os gamas, tendo em vista que as estimativas são bastante dispersas e assimétricas, com a

presença de *outliers*, nos remetendo a ideia de que esse estimador não seja robusto.

## 6.2 Trabalhos Futuros

Diante dessas discussões, temos algumas propostas para trabalhos futuros, consideradas interessantes na continuação deste, que são:

1. A substituição do algoritmo **ES** pelo *Robust Expectation-Solution* (**RES**) (ver Hall & Shen (2010)), aprimorando o processo de estimação, utilizando estimadores robustos. Inicialmente, essa substituição é sugerida apenas para os gamas, que apresentaram estimativas assimétricas.

2. Podemos realizar estudos de simulação comparativos com outros cenários, que possuam mais áreas, a fim de avaliar o desempenho dos estimadores quanto à eficiência. E comparar os **ICB-t** com outros intervalos de confiança empíricos, como por exemplo o obtido através do método **BCa** (*bias-corrected and acelerated*), para correção de viés (ver Gentle (2009b)).

3. Utilizar um critério de seleção de modelos, para auxiliar na escolha de covariáveis e até mesmo da matriz de correlação mais adequada. Uma sugestão seria o **QIC** (ver Pan (2001)), que é um critério de seleção de modelos que substitui a log-verossimilhança, usada na definição do critério **AIC**, pela quase verossimilhança.

4. E, ainda, estender o modelo proposto (**QIZDE**) para a modelagem de dados espaço-temporais, com as características tratadas neste trabalho, porém avaliando a correlação espaço-temporal.

# Referências Bibliográficas


Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Bjørnstad, O. N. & Falck, W. (2001). Nonparametric spatial covariance functions: estimation and testing. *Environmental and Ecological Statistics*, **8**(1), 53–70.

Bjørnstad, O. N., Ims, R. A. & Lambin, X. (1999). Spatial population dynamics: analyzing patterns and processes of population synchrony. *Trends in Ecology & Evolution*, **14**(11), 427–432.

Cancado, A., da Silva, C. & da Silva, M. (2011). A zero-inflated poisson-based spatial scan statistic. *Emerging Health Threats Journal*, **4**.

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine*, **21**(10), 1461–1469.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, **4**(5), 613–617.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Druck, S., Carvalho, M. S., Câmara, G., Monteiro, A. M. V. et al. (2004). *Spatial analysis of geographic data.*. Embrapa Cerrados.

Elashoff, M. & Ryan, L. (2004). An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, **13**(1), 48–65.

Epperson, B. K. (1993). Recent advances in correlation studies of spatial patterns of genetic variation. In *Evolutionary biology*, pages 95–155. Springer.

Fahrmeir, L. & Tutz, G. (2001). Random effect model.

Gentle, J. E. (2009a). *Computational statistics*, volume 308. Springer.

Gentle, J. E. (2009b). Monte carlo methods for statistical inference. In *Computational Statistics*, pages 417–433. Springer.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, **56**(4), 1030–1039.

Hall, D. B. & Shen, J. (2010). Robust estimation for zero-inflated poisson regression. *Scandinavian Journal of Statistics*, **37**(2), 237–252.

Hall, D. B. & Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, **4**(3), 161–180.

Imbiriba, E. B., Basta, P. C., Pereira, E. d. S., Levino, A. & Garnelo, L. (2009a). Hanseníase em populações indígenas do amazonas, brasil: um estudo epidemiológico nos municípios de autazes, eirunepé e são gabriel da cachoeira (2000 a 2005). *Cad Saude Publica*, **25**(5), 972–84.

Imbiriba, E. N. B., Silva Neto, A. L. d., Souza, W. V. d., Pedrosa, V., Cunha, M. d. G. & Garnelo, L. (2009b). Social inequality, urban growth and leprosy in manaus: a spatial approach. *Revista de Saúde Pública*, **43**(4), 656–665.

Johnson, N. L. & Kotz, S. (1969). *Distribution in statistics: Discrete distribution*. Wiley.

Klimko, L. A. & Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *The Annals of Statistics*, pages 629–642.

Kong, M., Xu, S., Levy, S. M. & Datta, S. (2015). Gee type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational Statistics & Data Analysis*, **85**(0), 54–66.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., Bertollini, R. et al. (1999). *Disease mapping and risk assessment for public health.*. John Wiley & Sons.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K. & McLachlan, G. J. (2006). Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, **15**(1), 47–61.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lima, M. S. d. & Duczmal, L. H. (2014). Abordagem bayesiana adaptativa para vigilância online de cluster espaciais (pp. 6-10). *Revista da Estatística da Universidade Federal de Ouro Preto*, **3**(3).

Lima, M. S. d., Duczmal, L. H. & Pinto, L. P. (2013). Spatial scan statistics for models with excess zeros and overdispersion. *Online Journal of Public Health Informatics*, **5**(1).

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, pages 59–67.

McCullagh, P. & Nelder, J. (1983). Generalised linear modelling. *Chapman and Hall, London. Negro, JJ & Hiraldo, F.(1992) Sex ratios in broods of the lesser kestrel Falco naumanni. Ibis*, **134**, 190–191.

McCullagh, P. & Nelder, J. (1989). Quasi-likelihood functions. In *Generalized Linear Models*, pages 323–356. Springer.

Monod, A. (2007). *An analysis on count panel data using a zero-inflated poisson model*. Ph.D. thesis, Université de Neuchâtel-Faculté des sciences économiques-Institut de statistique.

Monod, A. (2012). *A Quasi-Likelihood Approach to Zero-Inflated Spatial Count Data*. Ph.D. thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

Nelder, J. & Wedderburn, R. (1972). General linearized models. *J. Roy. Stat. Soc. Ser. A*, **135**, 370–384.

Nelder, J. A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**(2), 221–232.

Nie, L., Wu, G., Brockman, F. J. & Zhang, W. (2006). Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, **22**(13), 1641–1647.

Nieto-Barajas, L. & Bandyopadhyay, D. (2013). A zero-inflated spatial gamma process model with applications to disease mapping. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**(2), 137–158.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**(1), 120–125.

Perumean-Chaney, S. E., Morgan, C., McDowall, D. & Aban, I. (2013). Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation*, **83**(9), 1671–1683.

Rosen, O., Jiang, W. & Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika*, **87**(2), 391–404.

Van den Broek, J. (1995). A score test for zero inflation in a poisson distribution. *Biometrics*, pages 738–743.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, **16**(3), 275–289.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussânewton method. *Biometrika*, **61**(3), 439–447.

Yang, M. (2012). Statistical models for count time series with excess zeros.

Yasui, Y. & Lele, S. (1997). A regression method for spatial disease rates: an estimating function approach. *Journal of the American Statistical Association*, **92**(437), 21–32.

Yau, K. K., Lee, A. H. & Carrivick, P. J. (2004). Modeling zero-inflated count series with application to occupational health. *Computer methods and programs in biomedicine*, **74**(1), 47–52.

Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

Zhang, T., Zhang, Z. & Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in medicine*, **31**(8), 762–774.