ALISSON DE LIMA BRITO

DADOS DE ÁREA NA FAMÍLIA GAMLSS EM ESTUDOS EPIDEMIOLÓGICOS

LAVRAS – MG

2021

ALISSON DE LIMA BRITO

DADOS DE ÁREA NA FAMÍLIA GAMLSS EM ESTUDOS EPIDEMIOLÓGICOS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Dr[a]. Izabela Regina Cardoso de Oliveira
Orientadora

LAVRAS – MG

2021

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo próprio autor.

Brito, Alisson de Lima
Dados de área na família GAMLSS em estudos epidemiológicos / Alisson de Lima Brito. – Lavras : UFLA, 2021.
91 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2021.
Orientadora: Dr[a]. Izabela Regina Cardoso de Oliveira.
Bibliografia.

1. Estatística Espacial. 2. Modelos de Regressão. 3. Saúde Pública. I. Cardoso de Oliveira, Izabela Regina. II. Título.

ALISSON DE LIMA BRITO

DADOS DE ÁREA NA FAMÍLIA GAMLSS EM ESTUDOS EPIDEMIOLÓGICOS
AREA DATA IN THE GAMLSS FAMILY IN EPIDEMIOLOGICAL STUDIES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 09 de Março de 2021.

Dr. João Domingos Scalon UFLA
Dr[a]. Fernanda De Bastiani UFPE

Dr[a]. Izabela Regina Cardoso de Oliveira
Orientadora

LAVRAS – MG
2021

À minha família.
Com amor, dedico.

AGRADECIMENTOS

À minha orientadora, Profª. Dª. Izabela Regina Cardoso de Oliveira, que desde minha entrada ao mestrado me deu o suporte do qual precisei, me incentivou e me apoiou diante de problemas para que continuasse firme até aqui. Aos ensinamentos, e a dedicação que teve durante o desenvolvimento deste trabalho e por sempre acreditar em meu potencial.

Aos amigos, Tiago Almeida de Oliveira e Ana Patrícia Bastos Peixoto, que sempre me ajudaram desde a minha graduação na Universidade Estadual da Paraíba, sempre me apoiaram e me incentivaram em relação aos meus estudos, contribuindo para meu crescimento pessoal e profissional.

Aos amigos, Arthur Oliveira Costa e Leandro Valter Gomes, que estiveram comigo desde a graduação em Estatística e que trilharam comigo mais uma jornada no mestrado. Sempre estivemos nos ajudando e contribuindo para o crescimento individual de cada um.

À todos os professores e funcionários do Departamento de Estatística da Universidade Federal de Lavras, que se mostraram bastante acolhedores e prestativos ao corpo discente do Programa de Pós-graduação em Estatística e Experimentação Agropecuária. Sempre empenhados em ajudar na solução de problemas relacionados ao curso e muito eficientes.

Aos colegas que conheci durante minha estadia na UFLA-Lavras, que apesar do pouco tempo que permaneci no programa, foram muitas experiências e conhecimentos compartilhados.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro desde o início até a conclusão deste trabalho.

A todos deixo meu sincero obrigado!

A genialidade é 1% inspiração e 99% transpiração.
(Thomas Edison)

RESUMO

O avanço no campo das análises estatísticas tem sido cada vez mais importante nos últimos anos. Em particular, os modelos de regressão compõem um conjunto de ferramentas estatísticas que tem recebido grandes contribuições em um curto intervalo de tempo. Estes modelos são uma das ferramentas mais utilizadas no meio científico para descrever diversos fenômenos, nas mais variadas áreas do conhecimento. Desde a sua introdução, o modelo linear normal tem sido utilizado em muitas pesquisas científicas para modelagem de dados. Porém, devido às limitações encontradas neste modelo, foram então introduzidos os Modelos Lineares Generalizados, que englobam um maior número de distribuições probabilísticas para a variável resposta. Posteriormente, foram propostos os Modelos Aditivos Generalizados que flexibilizaram a relação entre as variáveis explicativas e resposta. Em seguida foram introduzidos os modelos GAMLSS, que podem ser vistos como uma generalização dos demais modelos citados anteriormente. Essa nova classe de modelos oferece não só um maior número de distribuições probabilísticas para a modelagem da variável resposta como também maior flexibilização para a relação entre as variáveis, assim como a modelagem de outros parâmetros da distribuição, além do de locação. Recentemente, uma adaptação foi feita nos modelos GAMLSS para incorporar o efeito de uma estrutura de dependência espacial quando o pressuposto de independência das observações da variável resposta é quebrado e estas apresentam uma correlação no espaço. Neste estudo, objetivou-se estudar os modelos GAMLSS no contexto de análise espacial e aplicá-los a dados reais. Para isso, foram utilizados dados de ocorrência de Tuberculose bovina no estado de Minas Gerais e de dengue no estado da Paraíba. Foram obtidos ajustes satisfatórios para ambas as bases dados, mesmo quando estas apresentaram problemas de estrutura como forte assimetria e curtose e problema de superdispersão. Para os dados de dengue, foi introduzida uma componente espacial no modelo através de um modelo intrínseco autoregressivo, já que os dados apresentaram autocorrelação espacial significativa.

Palavras-chave: Dengue. Estatística espacial. Modelos de regressão flexíveis. Saúde pública. Tuberculose bovina.

ABSTRACT

The progress in the field of statistical analysis has been increasingly significant in recent years. In particular, regression models comprise a set of statistical tools that have received major contributions in a short period of time. These models are one of the most used tools in the scientific world to describe several phenomena in the most varied areas of knowledge. Since the beginning, the normal linear model has been used in many scientific researches for data modeling. However, due to the limitations found in this model, the Generalized Linear Models were introduced, which encompass more probabilistic distributions for the response variable. Subsequently, Generalized Additive Models were proposed, which made the relationship between variables more flexible. Then the GAMLSS models were introduced, which can be seen as a generalization of the other models mentioned above. This new class of models allows not only a greater number of probabilistic distributions for modeling the response variable, but also greater flexibility for the relationship between the variables, as well as the modeling of other distribution parameters, in addition to the location. Recently, an adaptation was made in the GAMLSS models to incorporate the effect of a spatial dependence structure when the assumption of independence of the observations of the response variable is not met and there is correlation in space. In this study, we aimed to study the GAMLSS models in the context of spatial analysis and apply them to real data. For this, data on the occurrence of bovine tuberculosis in the state of Minas Gerais and dengue in the state of Paraíba were used. Satisfactory adjustments were obtained for both databases, even when they presented structural problems such as strong asymmetry and kurtosis and a problem of overdispersion. For dengue data, a spatial component was introduced in the model through an intrinsic autoregressive model, since the data showed significant spatial autocorrelation.

Keywords: Dengue. Bovine tuberculosis. Flexible regression models. Public health. Spatial statistics.

LISTA DE FIGURAS

Figura 3.1 – *Worm plot* para um modelo bem ajustado. 28
Figura 3.2 – Representação gráfica do diagrama de dispersão de Moran. 34
Figura 3.3 – Grafo não direcionado e matriz de proximidade correspondente. 38
Figura 4.1 – Distribuição global da Tuberculose bovina (bTB) entre 2017 e 2018. . 46
Figura 4.2 – Distribuição da bTB no Brasil entre os anos de 1999 e 2019 (número de casos registrados). 48
Figura 4.3 – Localização geográfica do Estado de Minas Gerais, divisões e seus respectivos municípios. 51
Figura 4.4 – Visualização gráfica para o número de animais com diagnóstico positivo para a bTB (CFTA) no estado de Minas Gerais. 55
Figura 4.5 – Gráfico de dispersão para a relação entre as variáveis PB, ER e VO com a variável CFTA. 56
Figura 4.6 – Resultados positivos e negativos para os testes de Tuberculose bovina nos municípios do estado de Minas Gerais entre os anos 2014 e 2018. . 57
Figura 4.7 – Possíveis casos de subnotificação do número de animais positivos para a bTB nos municípios do estado de Minas Gerais entre os anos 2014 e 2018 a partir da relação AIA (número de animais abatidos com lesões sugestivas) e CFTA (número de animais com diagnóstico positivo). . . 58
Figura 4.8 – Visualização geográfica para o número de animais com diagnóstico positivo para a bTB entre os anos 2014-2018 Figura 4.8(a) e a diferença entre o número de animais abatidos e o número de animais positivos Figura 4.8(b). 59
Figura 4.9 – *Worm plot* para os resíduos dos modelos ZALG, ZANBI e ZASICHEL para a modelagem do número de bovinos com diagnóstico positivo para a bTB no estado de Minas Gerais. 60
Figura 4.10 –Termos de suavização para as variáveis explanatórias em relação ao preditor de μ. 62
Figura 4.11 –Análise dos resíduos para o modelo ZANBI. 63
Figura 5.1 – Localização geográfica do estado da Paraíba, mesorregiões e seus respectivos municípios. 70

Figura 5.2 – Municípios com ausência e presença de casos notificados por dengue no estado da Paraíba, segundo ano de notificação. 73
Figura 5.3 – Visualização gráfica da distribuição da taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018. 74
Figura 5.4 – Mapa de intervalos iguais (Figura 5.4(a)) e mapa de quartis (Figura 5.4(b)) para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018. 75
Figura 5.5 – Índice I_i de Moran local (Figura 5.5(a)) e LISA *Map* (Figura 5.5(b)) para os municípios do estado da Paíba em relação à taxa de incidência de dengue por 100 mil habitantes. 76
Figura 5.6 – Moran *Scatterplot* e mapa de *clusters* espaciais para a taxa de incidência de dengue por 100 mil habitantes no estado da Paraíba (2008-2018). 77
Figura 5.7 – *Worm plot* para os modelos Gama Generalizado e Inversa Gaussiana Generalizado. 78
Figura 5.8 – Relação entre as variáveis explanatórias e o preditor do parâmetro μ. . 80
Figura 5.9 – Relação entre as variáveis explanatórias e o preditor do parâmetro σ. . 81
Figura 5.10 – Relação entre a variável precipitação e o preditor do parâmetro ν. . . . 81
Figura 5.11 – *Worm plot* do modelo Gama Generalizado após a inclusão da componente espacial. 82
Figura 5.12 – Análise de diagnóstico para o modelo GG. 82

LISTA DE TABELAS

Tabela 3.1 – Ligações canônicas para algumas distribuições importantes. 20
Tabela 3.2 – Diferentes formas da curva ajustada no *worm plot*, irregularidades nos resíduos e no ajuste da distribuição. 28
Tabela 4.1 – Estatísticas descritivas para as variáveis número de animais com diagnóstico positivo para a bTB (CFTA), número de propriedades com bovinos (PB), efetivo de rebanho bovino (ER), número de vacas ordenhadas (VO) e número de animais abatidos com lesões sugestivas (AIA). 54
Tabela 4.2 – Teste para o τ de Kendall para a relação entre as variáveis em estudo. 56
Tabela 4.3 – Possíveis modelos e valores do Critério de Informação de Akaike Generalizado para o número de animais com diagnóstico positivo para a bTB no estado de Minas Gerais. 60
Tabela 4.4 – Estimativas dos parâmetros para o modelo ZANBI para o número de animais com diagnóstico positivo para a bTB. 63
Tabela 5.1 – Estatísticas descritivas para os casos notificados por dengue nos municípios do estado da Paraíba entre os anos 2008-2018. 73
Tabela 5.2 – Estatísticas descritivas para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018. 74
Tabela 5.3 – Testes de significância para a autocorrelação espacial para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018. 76
Tabela 5.4 – Possíveis modelos e valores do GAIC para a taxa de incidência de dengue no estado da Paraíba. 78

SUMÁRIO

1 INTRODUÇÃO GERAL 12
2 OBJETIVOS 16
2.1 Objetivo Geral 16
2.2 Objetivos Específicos 16
3 FUNDAMENTAÇÃO TEÓRICA 17
3.1 O Modelo Linear Normal 17
3.2 Modelos Lineares Generalizados - GLM 18
3.3 Modelos Aditivos Generalizados - GAM 20
3.4 Modelos GAMLSS 21
3.4.1 Estimação dos Parâmetros 24
3.4.2 Seleção dos Modelos 25
3.5 Estatística Espacial 29
3.5.1 Análise de Dados de Área 30
3.5.2 Matriz de Proximidade Espacial 30
3.5.3 Autocorrelação Espacial 31
3.5.4 Modelos de Regressão Espacial para Dados de Área 34
3.6 Campos Aleatórios Markovianos Gaussianos 36
3.6.1 Independência Condicional 36
3.6.2 Grafo não Direcionado 37
3.6.3 Modelos Autoregressivos Condicional (CAR) e Intrínseco (IAR) 37
3.7 O Modelo Autoregressivo Intrínseco na família GAMLSS 38
4 APLICAÇÃO 1: ESTUDO DE CASOS DE TUBERCULOSE BOVINA NO ESTADO DE MINAS GERAIS ENTRE OS ANOS 2014-2018 42
4.1 Introdução 42
4.2 Revisão Bibliográfica 44
4.3 Material e Métodos 50
4.4 Resultados 53
4.5 Discussão 64
4.6 Conclusões 65

5 APLICAÇÃO 2: ESTUDO DOS CASOS NOTIFICADOS DE DENGUE NO ESTADO DA PARAÍBA ENTRE OS ANOS 2008-2018 67
5.1 Introdução 67
5.2 Material e Métodos 69
5.3 Resultados 72
5.4 Discussão 82
5.5 Conclusões 84
6 CONSIDERAÇÕES FINAIS 85
REFERÊNCIAS 86

1 INTRODUÇÃO GERAL

A análise de regressão é uma área da Estatística que tem recebido grandes contribuições ao longo dos anos, especialmente com o avanço constante da tecnologia computacional. Os modelos de regressão, são em particular, uma das ferramentas estatísticas mais utilizadas por pesquisadores, uma vez que é possível descrever de forma simples um determinado fenômeno por meio de uma equação que estabelece uma relação entre uma variável resposta Y_i e uma ou mais variáveis explicativas $\mathbf{x}_i$. Nestes modelos, essa relação pode ser explicada por diversos fatores associados ao fenômeno de interesse.

A forma mais simples de um modelo de regressão é o modelo linear normal com apenas duas variáveis (uma independente e outra dependente), o qual exige normalidade para fins de inferência estatística. No entanto, na prática nem sempre isso ocorre e o uso do modelo linear normal pode não ser apropriado nessa situações. Uma maneira de lidar com a falta de normalidade é fazer uma transformação na variável resposta de maneira a obter uma aproximação para a normalidade. Todavia, isso nem sempre resolve o problema.

Contornando esse problema, Nelder e Wedderburn (1972) introduziram os Modelos Lineares Generalizados (*Generalized Linear Models* - GLM) conhecidos por serem uma extensão do modelo linear normal. Nos GLM a variável resposta pode ser representada por um conjunto de distribuições de probabilidade, inclusive a distribuição normal. No entanto, para isso, a função densidade de probabilidade da distribuição pretendida deve pertencer à família exponencial de distribuições. Esses modelos podem ser considerados quando a suposição de normalidade da variável resposta não é atendida.

Com a inclusão de uma função de ligação que relaciona a componente que corresponde a variável resposta com a componente correspondente às variáveis explanatórias, nos Modelos Lineares Generalizados, passa a ser possível a modelagem de variáveis de interesse que assumem a forma de contagem, contínuas simétricas e assimétricas, binárias e categóricas. Apesar de apresentar uma certa flexibilização para a relação entre Y_i e as variáveis explanatórias $\mathbf{x}_i$, devido à função de ligação, nos GLM, essa relação ocorre linearmente.

Para solucionar esse problema Hastie e Tibshirani (1990) propuseram uma adaptação dos GLM, em que a estrutura de relação entre a resposta e uma variável preditora pode ser feita por meio de técnicas de regressão não paramétrica, o que suaviza a rela-

ção entre os preditores e a função da média que será modelada, mas ainda assim assume que a distribuição dos dados segue uma distribuição de probabilidade com densidade pertencente a família exponencial. Esses modelos foram denominados de Modelos Aditivos Generalizados (*Generalized Additive Models* - GAM), nos quais, apenas a modelagem da média da distribuição é possível, sendo os demais parâmetros considerados constantes.

Com o objetivo de resolver as limitações encontradas também nos modelos GAM, Rigby e Stasinopoulos (2005) apresentaram os Modelos Aditivos Generalizados para Locação, Escala e Forma (*Generalized Additive Models for Location, Scale and Shape* - GAMLSS). Nessa nova classe, os modelos pertencem a uma família de distribuições mais flexíveis, chamada de família GAMLSS. Nos modelos GAMLSS, a variável resposta pode agora ser representada por uma variedade de distribuições de probabilidade pertencentes a família GAMLSS. Outra vantagem dos modelos GAMLSS em relação aos modelos GAM é a possibilidade de modelar os parâmetros de escala e de forma da distribuição. Essa diversidade de modelos permite também o tratamento de dados em situações adversas de forma eficaz, como por exemplo, excesso de zeros na variável resposta e/ou alta variabilidade, assimetria positiva, assimetria negativa, curtose, etc. Com a introdução da classe de modelos GAMLSS tornou-se possível a modelagem de muitos problemas antes intratáveis, devido a forma complexa da distribuição assumida pela variável resposta.

Existem ainda, modelos estatísticos que consideram a localização da variável resposta, ou seja, o espaço onde ela ocorre, durante o processo de análise. Esses modelos, são de grande valia para se avaliar problemas em que a pressuposição de independência da variável resposta não é atendida, uma vez que esta é autocorrelacionada no espaço. Nessas situações, espera-se uma melhoria na modelagem dessa variável quando consideramos a estrutura espacial presente no fenômeno em estudo. A área da Estatística que engloba as ferramentas para esse tipo de análise é chamada de Estatística Espacial.

Cressie (1993), classifica os dados espaciais em três principais categorias, são elas: Dados Geoestatísticos, Dados de Área e Processos Pontuais. De acordo com Plant (2018), a Geoestatística é definida quando a variável resposta varia continuamente no espaço, como por exemplo: precipitação, temperatura, focos de calor. Para Banerjee, Carlin e Gelfand (2014), na metodologia de Dados de Área a variável resposta é medida em partições (áreas) regulares ou irregulares da região de estudo. E segundo Bivand, Pebesma

e Gomez-Rubio (2013), um Processo Pontual é um processo estocástico em que observamos a localização de alguns eventos de interesse em uma região de estudo.

De acordo com De Bastiani et al. (2018), variação espacial discreta, em que as variáveis estão definidas em um domínio discreto, como em dados de área por exemplo, pode ser modelada por Campos Aleatórios de Markov. Mais especificamente por modelos IAR (*Gaussian Intrinsic Autoregressive Models*), que por sua vez, são um caso particular dos GMRF (*Gaussian Markov Random Fields*).

Rue e Held (2005) apresentam uma base teórica e aplicações práticas de GMRF. Segundo os autores, essa metodologia tem alta aplicabilidade na Estatística, bem como em diversas outras áreas do conhecimento. É bastante comum vermos aplicações de GMRF em áreas como Análise de Séries Temporais, Análise de Sobrevivência, Estatística Espacial, Análise de Imagem, Regressão Semi-paramétrica, entre outras.

Na literatura, alguns modelos de regressão para dados espacialmente dependentes, localizados em áreas bem definidas, foram propostos e são amplamente utilizados em diversas situações. Contudo, grande parte desses modelos ainda supõe normalidade para os resíduos, o que os torna bastante limitados do ponto de vista inferencial e de modelagem. Aplicações desses modelos são muito comuns em econometria.

Wood (2006) apresenta modelos IAR com abordagem nos modelos generalizados aditivos, como uma forma de flexibilizar a distribuição da variável resposta e considerar a estrutura espacial na análise. Mais recentemente, De Bastiani et al. (2018) apresentam uma base teórica e aplicação de modelos da família GAMLSS considerando a estrutura espacial através de um modelo IAR, dando maior flexibilização à distribuição da variável resposta e trazendo a possibilidade de modelar os demais parâmetros da distribuição com o incremento espacial incorporado ao modelo.

Atualmente tem-se levantado uma discussão interessante por pesquisadores de diversas áreas sobre a forma de abordagem das enfermidades de caráter zoonótico (doenças transmitidas do animal para o homem). Trata-se do conceito *One Health*, o qual defende que os cuidados relativos à saúde do homem, dos animais e do meio ambiente deve ser tratado de forma unificada, buscando a interação entre a medicina veterinária e humana. Neste contexto, as zoonoses representam um fator que reforça a iniciativa em discussão, visto que essas doenças apresentam impacto na saúde de ambos. Para o melhor entendimento do comportamento dessas doenças, estudos epidemiológicos devem ser conduzidos.

A estatística espacial tem grande importância na aplicação em estudos epidemiológicos, uma vez que é possível descrever por meio de suas ferramentas, o comportamento espacial de uma determinada doença na região de estudo e verificar se esta ocorre de maneira dependente na região, facilitando a tomada de decisão para um direcionamento de medidas mitigadoras em localidades específicas. Também o uso dos modelos de regressão se fazem importantes nessa perspectiva, uma vez que, por intermédio destes, pode-se avaliar a influência de outros fatores que podem estar ligados ao aumento ou diminuição da doença.

2 OBJETIVOS

2.1 Objetivo Geral

Estudar os modelos GAMLSS no contexto de análise espacial e aplicá-los a dados de ocorrência de tuberculose bovina (bTB) no estado de Minas Gerais e de dengue no estado da Paraíba.

2.2 Objetivos Específicos

1. Identificar a presença de autocorrelação espacial para os dados de bTB no estado de Minas Gerais e dengue nos municípios do estado da Paraíba;
2. Analisar os dados de dengue e bTB por meio dos modelos GAMLSS;
3. Incluir o efeito espacial nos modelos, de forma a obter possíveis melhorias no ajuste dos mesmos;
4. Examinar os resíduos dos modelos de forma a identificar possíveis irregularidades nos ajustes;
5. Avaliar a relação entre os dados de ocorrência de bTB e dados de animais abatidos com lesões sugestivas, de forma a identificar possíveis discrepâncias entre esses dados nos municípios do estado de Minas Gerais.

3 FUNDAMENTAÇÃO TEÓRICA

A seguir é apresentado parte da teoria que envolve os modelos de regressão linear, com ênfase no modelo linear normal, modelos lineares generalizados e modelos aditivos generalizados, os quais servirão como base introdutória para a teoria dos modelos GAMLSS. Em seguida é apresentada a metodologia para análise de dados espaciais, mais especificamente para análise de dados de área. Posteriormente é feita uma definição de Campos Aleatórios Markovianos Gaussianos, bem como é apresentada a teoria sob essa metodologia de análise e também é comentado sobre a inclusão de um modelo Intrínseco Autoregressivo na metodologia GAMLSS para análise de dados espacialmente correlacionados, que é foco principal deste trabalho.

3.1 O Modelo Linear Normal

A formulação geral de um modelo de regressão visa relacionar uma variável resposta Y_i com uma mais variáveis explanatórias ou independentes $\mathbf{x}_i$, de forma que seja possível estimarmos os valores da variável resposta por meio de uma combinação linear das variáveis independentes. Assim como citado anteriormente, o modelo de regressão mais simples é o modelo de regressão linear normal composto por apenas duas variáveis, em que uma delas é independente. Este modelo pode ser estendido para p variáveis independentes. Este é comumente chamado de modelo de regressão linear normal múltiplo e descrito por

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ip} + \varepsilon_i = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k + \varepsilon_i, \tag{3.1}$$

ou matricialmente,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$ é o vetor que contém os valores da variável resposta, com dimensão $n \times 1$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor de parâmetros de dimensão $(p+1) \times 1$ associado a matriz $\boldsymbol{X} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ de variáveis independentes de dimensão $n \times (p+1)$ e ε_i é o termo de erro aleatório associado à i-ésima observação da variável resposta, em que $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, ou seja, assume-se que o termo de erro é independente e identicamente distribuído segundo uma distribuição de probabilidade normal com média zero

e variância constante. Isso implica dizer que $y_i \overset{iid}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, em que μ_i é dado por $\mu_i = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k$. Matricialmente temos $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$, em que $\boldsymbol{\varepsilon}$ tem dimensão $n \times 1$. Logo, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}\sigma^2)$, sendo $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ e $\boldsymbol{I}$ uma matriz identidade de ordem n.

O modelo linear normal é bastante utilizado em muitas situações práticas. Todavia, em muitos casos, a natureza dos dados não permite sua adequação. Uma alternativa para solucionar esse problema é fazer uma transformação na variável resposta buscando uma aproximação à distribuição normal, mas nem sempre o problema é solucionado.

3.2 Modelos Lineares Generalizados - GLM

Na perspectiva de poder resolver as limitações encontradas no modelo linear normal, Nelder e Wedderburn (1972) introduziram uma classe de modelos mais geral, em que a variável resposta pode ser modelada por diversas distribuições de probabilidade, inclusive a distribuição normal, esses modelos são chamados de Modelos Lineares Generalizados, do inglês *Generalized Linear Models* (GLM). A principal prerrogativa nessa classe de modelos é que a distribuição de probabilidade utilizada para modelar a variável aleatória Y_i deve obrigatoriamente pertencer a família exponencial de distribuições.

A família exponencial engloba uma variedade de distribuições de probabilidade cujas densidades podem ser representadas de uma forma unificada. Exemplos de distribuições pertencentes à essa família são as distribuições Binomial, Binomial Negativa, Normal, Normal Inversa, Gama, Poisson, Multinomial, Beta, Rayleigh, entre outras. Dizemos que uma distribuição de probabilidade qualquer pertence à família exponencial de distribuições se, e somente se, sua função densidade ou de probabilidade pode ser expressa da seguinte forma

$$f(x;\theta) = h(x)\exp[\eta(\theta)t(x) - b(\theta)], \tag{3.2}$$

em que as funções $\eta(\theta)$, $b(\theta)$, $t(x)$ e $h(x)$ tem valores em subconjuntos dos reais e $\eta(\theta)$, $b(\theta)$ e $t(x)$ não são únicas (CORDEIRO; DEMETRIO, 2013). Se as funções $\eta(\theta)$ e $t(x)$ são iguais a função identidade, a função em (3.2) é chamada de família exponencial na forma canônica e é dada por

$$f(x;\theta) = h(x)\exp[\theta x - b(\theta)],$$

sendo θ denominado de parâmetro canônico nessa parametrização.

De acordo com McCullagh e Nelder (1989), um GLM é composto basicamente por três componentes: a componente aleatória que corresponde a variável aleatória Y_i, a componente sistemática η_i que corresponde a uma soma linear das variáveis explanatórias e seus efeitos e uma função de ligação $g(\cdot)$ que relaciona as componentes aleatória e sistemática.

A componente aleatória do modelo pode então ser representada pela família exponencial na forma canônica com a introdução de um parâmetro ϕ de perturbação. Neste caso denotamos que $Y_i \overset{\text{ind}}{\sim} \mathcal{FE}(\mu_i, \phi)$. A função densidade da distribuição é então definida da seguinte forma

$$f(y_i; \mu_i, \phi) = \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\},$$

em que, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas e ϕ ($\phi > 0$) é o parâmetro de dispersão da distribuição. Também podemos denotar a distribuição da componente aleatória em notação matricial como segue

$$\boldsymbol{Y} \overset{\text{ind}}{\sim} \mathcal{FE}(\boldsymbol{\mu}, \boldsymbol{\phi}),$$

em que $\boldsymbol{\phi} = (\phi, \phi, \ldots, \phi)^\top$ é um vetor de constantes de dimensão $n \times 1$.

A componente sistemática do modelo é representada por uma soma linear das variáveis explanatórias e seus efeitos, que produz o que chamamos de preditor linear η_i dado por

$$\eta_i = \sum_{k=1}^{p} x_{ik}\beta_k = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{ou matricialmente} \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

As componentes aleatória e sistemática são relacionadas através de uma função monótona e diferenciável $g(\cdot)$ chamada de função de ligação, que vincula a média ao preditor linear como segue

$$g(\mu_i) = \eta_i \quad \text{ou} \quad g(\boldsymbol{\mu}) = \boldsymbol{\eta}.$$

O modelo linear normal é obtido particularmente quando assumimos uma distribuição normal para a variável resposta Y_i e a função de ligação é igual à função identidade,

dessa forma, o preditor linear modela diretamente a média da distribuição como segue $\eta_i = \mu_i$. Tal como no modelo normal, alguns modelos probabilísticos possuem uma função de ligação que chamamos de ligação canônica. A seguir é apresentada uma tabela com as distribuições de probabilidades e suas respectivas ligações canônicas.

Tabela 3.1 – Ligações canônicas para algumas distribuições importantes.

Distribuição	**Ligação canônica**	**Função**
Normal	*identidade*	$\eta_i = \mu_i$
Binomial	*logística*	$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
Poisson	*logarítmica*	$\eta_i = \log(\mu_i)$
Gama	*recíproca*	$\eta_i = \mu_i^{-1}$
Inv. Gaussiana	*recíproca do quadrado*	$\eta_i = \mu_i^{-2}$

Fonte: (PAULA, 2013).

3.3 Modelos Aditivos Generalizados - GAM

A classe de modelos lineares generalizados trouxe um grande avanço na modelagem estatística, expandindo o número de distribuições probabilísticas que podem ser utilizadas para a modelagem da variável resposta. Contudo, esses modelos assumem linearidade para a relação entre a média da variável resposta e as variáveis independentes. Para relaxar essa hipótese, Hastie e Tibshirani (1990) realizaram uma adaptação nos GLM, de forma que essa relação pudesse ser feita por meio de uma função de suavização das variáveis explanatórias, e denominaram de Modelos Generalizados Aditivos, do inglês *Generalized Additive Models* (GAM). O modelo GAM é dado por,

$$g(\mu_i) = \eta_i = \sum_{k=1}^{p} x_{ik}\beta_k + \sum_{j=1}^{J} h_j(x_{ij}), \quad Y_i \overset{\text{ind}}{\sim} \mathcal{FE}(\mu_i, \phi). \tag{3.3}$$

O modelo (3.3) pode também ser escrito matricialmente como segue

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{J} h_j(\mathbf{x}_j) = \boldsymbol{X\beta} + h_1(\mathbf{x}_1) + \ldots + h_J(\mathbf{x}_J), \quad \boldsymbol{Y} \overset{\text{ind}}{\sim} \mathcal{FE}(\boldsymbol{\mu}, \boldsymbol{\phi}),$$

em que h_j é uma função de suavização avaliada na covariável $\mathbf{x}_j$, para $j = 1, \ldots, J$ (STASINOPOULOS et al., 2017). Essa classe de modelos foi estudada extensivamente por Wood (2006).

Não obstante, a classe de modelos GAM também apresenta limitações como o fato da distribuição a ser utilizada para modelagem de Y_i ter de pertencer a família exponencial. Outra limitação é que apenas o parâmetro de locação da distribuição pode ser modelado e os demais parâmetros são considerados quantidades fixas.

3.4 Modelos GAMLSS

No intuito de generalizar os modelos das classes GLM e GAM, Rigby e Stasinopoulos (2005) propuseram uma nova classe de modelos ainda mais ampla denominada de Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS), nos quais é possível utilizar uma variedade de distribuições para a modelagem da variável aleatória Y_i. Essas distribuições de probabilidades não precisam necessariamente pertencer a família exponencial de distribuições para serem utilizadas. Rigby et al. (2019) apresentam de forma detalhada as distribuições probabilísticas implementadas no pacote `gamlss.dist` (STASINOPOULOS; RIGBY, 2020) para modelagem de Y_i. Outra vantagem dos modelos GAMLSS em relação aos modelos anteriores é que é possível modelar não só o parâmetro de locação da distribuição como também os parâmetros de escala e de forma.

De acordo com Rigby e Stasinopoulos (2005) o modelo GAMLSS assume que, para $i = 1, \ldots, n$, observações independentes y_i condicionadas a $\boldsymbol{\theta}^i$ com função densidade de probabilidade $f(y_i|\boldsymbol{\theta}^i)$, em que $\boldsymbol{\theta}^i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})^\top$ é um vetor de p parâmetros relacionado às variáveis explanatórias e aos efeitos aleatórios. Assim, os autores definem que, para $k = 1, \ldots, p$, e $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ um vetor de observações da variável resposta, $g_k(.)$ é uma função de ligação monótona que relaciona $\boldsymbol{\theta_k}$ às variáveis explanatórias e aos efeitos aleatórios por meio de um modelo aditivo dado por

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta_k} = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk} \boldsymbol{\gamma}_{jk}. \tag{3.4}$$

Na Equação (3.4), $\boldsymbol{\theta_k}$ e $\boldsymbol{\eta_k}$ são vetores de tamanho n e $\boldsymbol{\theta_k} = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})^\top$, $\boldsymbol{\beta_k} = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})^\top$ é um vetor de parâmetros de tamanho J'_k, $\boldsymbol{X}_k$ é uma matriz de delineamento conhecida de ordem $n \times J'_k$, $\boldsymbol{Z}_k$ é uma matriz de delineamento fixa e conhecida $n \times q_{jk}$ e $\boldsymbol{\gamma}_{jk}$ é um vetor de variáveis aleatórias q_{jk} dimensionais, assumindo serem independentes com distribuição normal com os seguintes parâmetros, $\boldsymbol{\gamma}_{jk} \overset{\text{ind}}{\sim} \mathcal{N}_{qjk}(\mathbf{0}, \boldsymbol{\lambda}_{jk}^{-1} \mathbf{G}_{jk}^{-1})$, aqui $\mathbf{G}_{jk}^{-1}$ é uma inversa generalizada da matriz simétrica e sin-

gular $\mathbf{G}_{jk}$ de ordem $q_{jk} \times q_{jk}$. Assim, $\boldsymbol{\gamma}_{jk}$ tem uma função densidade a priori imprópria proporcional à $\exp\left(-\frac{1}{2}\boldsymbol{\lambda}_{jk}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{G}\boldsymbol{\gamma}_{jk}\right)$. Note que na formulação do modelo (3.4), a distribuição condicional da variável resposta pode assumir qualquer distribuição, pertencente ou não à família exponencial, enquanto o vetor de variáveis aleatórias $\boldsymbol{\gamma}_{jk}$ é assumido ser normalmente distribuído.

Se não houverem termos aditivos para quaisquer parâmetros da distribuição, isto é, se para $k = 1, \ldots, p$, $J_k = 0$, o modelo (3.4) se resume ao modelo paramétrico linear, dado por

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta_k} = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{3.5}$$

Se, $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade $n \times n$ e $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$, para todas as combinações de j e k, o modelo (3.4) fica definido como

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta_k} = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{3.6}$$

em que $\mathbf{x}_{jk}$, para $j = 1, \ldots, J_k$ e $k = 1, \ldots, p$ são vetores explanatórios de tamanho n assumidos ser conhecidos, h_{jk} é uma função desconhecida das variáveis explanatórias X_{jk} e $h_{jk}(\mathbf{x}_{jk})$ é um vetor que avalia h_{jk} em $\mathbf{x}_{jk}$. Rigby e Stasinopoulos (2005) e Stasinopoulos e Rigby (2007) definem o modelo (3.6) como GAMLSS semiparamétrico aditivo, sendo este, um caso particular muito importante do modelo (3.4).

O modelo (3.6) pode ser estendido de forma a permitir que termos não lineares paramétricos sejam incluídos no modelo para o vetor de parâmetros $\boldsymbol{\theta_k}$, da seguinte forma (RIGBY; STASINOPOULOS, 2006; STASINOPOULOS; RIGBY, 2007)

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta_k} = h_k\left(\boldsymbol{X}_k\boldsymbol{\beta}_k\right) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{3.7}$$

em que, h_k, para $k = 1, \ldots, p$ são funções não lineares e $\boldsymbol{X}_k$ é uma matriz de delineamento conhecida de ordem $n \times J_k''$. Stasinopoulos e Rigby (2007) se referem ao modelo (3.7) como modelo GAMLSS semiparamétrico aditivo não linear. Se para todos os parâmetros da distribuição não houverem termos aditivos, então o modelo (3.7) se reduz ao modelo GAMLSS paramétrico não linear, como segue

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta_k} = h_k\left(\boldsymbol{X}_k\boldsymbol{\beta}_k\right). \tag{3.8}$$

Ademais, se $h_k(\boldsymbol{X}_k\boldsymbol{\beta}_k) = \boldsymbol{X}_k^\top\boldsymbol{\beta}_k$, para $i = 1,\ldots,n$ e $k = 1,\ldots,p$, então, o modelo (3.8) se reduz ao modelo paramétrico linear (3.5). Note que, alguns dos termos em cada $h_k(\boldsymbol{X}_k\boldsymbol{\beta}_k)$ podem ser lineares, o que corresponde ao caso onde o modelo GAMLSS é uma combinação de termos lineares e não lineares. Para Stasinopoulos e Rigby (2007) devemos nos referir a qualquer combinação entre os modelos (3.5) e (3.7) como um modelo GAMLSS paramétrico.

Se definirmos uma distribuição com quatro parâmetros para modelagem da variável aleatória Y_i, o vetor $\boldsymbol{\theta}^i$ associado a essa distribuição é dado por $\boldsymbol{\theta}^i = (\theta_{1i},\theta_{2i},\theta_{3i},\theta_{4i})^\top = (\mu_i,\sigma_i,\nu_i,\tau_i)^\top$, em que os dois primeiros μ_i e σ_i correspondem aos parâmetros de locação e escala, respectivamente e os dois últimos parâmetros ν_i e τ_i representam a forma da distribuição, em que, cada qual pode ser uma função das variáveis explanatórias. Dessa forma, o modelo (3.4) fica dado como em Rigby e Stasinopoulos (2005) da seguinte forma,

$$\begin{aligned}
\boldsymbol{Y} &\overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu},\boldsymbol{\tau}) \\
g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1}\boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\
g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2}\boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\
g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3}\boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\
g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4}\boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4},
\end{aligned} \tag{3.9}$$

em que $\boldsymbol{Y} \overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\nu}\boldsymbol{\tau})$ é uma distribuição da família GAMLSS com quatro parâmetros.

Os modelos da família GAMLSS têm sido amplamente utilizados para solucionar problemas em que a distribuição de probabilidade assumida pela variável resposta tem um alto grau de complexidade, impossibilitando muitas vezes, o ajuste de modelos mais comuns. Devido à flexibilização da relação entre as variáveis explanatórias com a variável resposta e a possibilidade de se poder modelar todos os parâmetros da distribuição, os modelos GAMLSS têm sido utilizados cada vez mais em trabalhos científicos como ferramenta para solucionar problemas muitas vezes intratáveis.

Segundo Stasinopoulos et al. (2017) os modelos GAMLSS têm sido utilizados nas mais diversas áreas do conhecimento, apresentando resultados satisfatórios do ponto de

vista da modelagem estatística. Estes modelos tornam a modelagem estatística uma ferramenta ainda mais poderosa, com um grande avanço em contraste com os demais modelos de regressão.

Nakamura et al. (2018) afirmam que GAMLSS é uma classe muito geral de modelos de regressão univariados, em que todos os parâmetros da distribuição podem ser modelados como funções paramétricas ou funções de suavização não-paramétricas aditivas das variáveis explanatórias. Os autores apresentam em seu trabalho uma nova distribuição que pode ser utilizada em contraste com a distribuição beta para dados pertencentes ao intervalo (0,1), levando-se em consideração a assimetria e a curtose da distribuição.

3.4.1 Estimação dos Parâmetros

Sob a suposição de independência das observações, o vetor de parâmetros $\boldsymbol{\beta}_k$ e os parâmetros dos efeitos aleatórios $\boldsymbol{\gamma}_{jk}$, para $j = 1, \ldots, J_k$ e $k = 1, \ldots, p$ são estimados na abordagem GAMLSS (para valores fixos dos hiper-parâmetros λ_{jk}) maximizando a função de verossimilhança penalizada dada por (RIGBY; STASINOPOULOS, 2005)

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk},$$

em que $\ell = \sum_{i=1}^{n}\log\left[f(y_i|\boldsymbol{\theta}^i)\right]$ é a função de log-verossimilhança do modelo. Se considerarmos o modelo com quatro parâmetros dado em (3.9), a função de log-verossimilhança ℓ, fica dada como segue,

$$\ell = \sum_{i=1}^{n}\log\left[f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)\right],$$

em que $f(\cdot)$ representa a função densidade de probabilidade da variável resposta condicionada ao vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^{\top}$. Deste modo, a função de log-verossimilhança penalizada para o modelo fica dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k}\lambda_{jk}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk}. \tag{3.10}$$

Sendo assim, as estimativas obtidas serão referentes ao vetor de parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)^{\top}$ que corresponde a parte linear do modelo, aos parâmetro dos efeitos

aleatórios $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}, \ldots, \boldsymbol{\gamma}_{J_1 1}, \boldsymbol{\gamma}_{12}, \ldots, \boldsymbol{\gamma}_{J_4 4})^\top$ e ao vetor de hiper-parâmetros do modelo $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{11}, \ldots, \boldsymbol{\lambda}_{J_1 1}, \boldsymbol{\lambda}_{12}, \ldots, \boldsymbol{\lambda}_{J_4 4})^\top$ [ver De Bastiani et al. (2018)].

Rigby e Stasinopoulos (2005) fornecem dois algoritmos para maximizar a função de log-verossimilhança penalizada ℓ_p dada na equação (3.10) para a estimação dos parâmetros. São eles, o algoritmo CG, que é uma generalização do algoritmo utilizado em Cole e Green (1992) e o algoritmo RS que é uma generalização do algoritmo utilizado em Rigby e Stasinopoulos (1996a) e Rigby e Stasinopoulos (1996b).

Para estimação do modelo por meio do algoritmo CG é necessário que se tenha informações sobre as derivadas (esperadas ou aproximadas) de primeira e segunda ordem e as derivadas cruzadas da função de verossimilhança em relação ao vetor de parâmetros $\boldsymbol{\theta}$. Contudo, para muitas funções densidade de probabilidade $f(y_i|\boldsymbol{\theta}^i)$, os parâmetros têm informação ortogonal, isto é, os valores esperados das derivadas cruzadas da função de verossimilhança são nulas. Nessas situações, o algoritmo RS é mais apropriado, pelo fato de não requerer o uso das derivadas cruzadas (RIGBY; STASINOPOULOS, 2005).

No *software* R (R Core Team, 2019) existe ainda a possibilidade de ajustar os modelos usando uma mistura desses dois algoritmos. Neste caso, o processo de estimação é iniciado pelo RS e finalizado com o CG. Mais detalhes sobre o funcionamento dos algoritmos pode ser encontrado em Rigby e Stasinopoulos (2005) e Stasinopoulos et al. (2017).

De acordo com Rigby e Stasinopoulos (2005), o vetor de hiper-parâmetros $\boldsymbol{\lambda}$ pode ser fixo ou estimado. As estimativas para os hiper-parâmetros podem ser obtidas utilizando métodos locais (RIGBY; STASINOPOULOS, 2014) ou métodos globais (RIGBY; STASINOPOULOS, 2005). Os métodos locais são geralmente mais rápidos e de fácil implementação, em relação aos métodos globais (DE BASTIANI et al., 2018).

3.4.2 Seleção dos Modelos

A seleção do modelo é de fato, uma das etapas mais importantes durante a análise de regressão, uma vez que problemas de superestimação ou subestimação dos valores da variável de interesse podem ocorrer, ocasionando estimativas viciadas e não representativas para o fenômeno estudado. A seleção do modelo deve ser feita de maneira tal que, o modelo mais adequado para o fenômeno em estudo seja escolhido, de forma a evitar problemas com as estimativas. Isto pode ser feito de várias maneiras, duas delas são

bastante utilizadas na literatura: o uso de critérios de seleção e a validação cruzada do modelo (para grandes amostras).

Uma maneira muito comum de comparação entre modelos é trabalhar com a ideia de modelos encaixados ou aninhados, ou seja, é preciso encontrar um modelo mais geral de forma que os demais modelos a serem avaliados são casos particulares deste. Dessa forma, considerando dois modelos ($\mathcal{M}_1$ e $\mathcal{M}_2$), em que $\mathcal{M}_2$ é o modelo geral e $\mathcal{M}_1$ está aninhado à $\mathcal{M}_2$. A comparação desses modelo pode ser feita através do teste da razão de verossimilhanças, que consiste em avaliar o logaritmo da função de verossimilhança do modelo geral e do modelo reduzido.

Para Stasinopoulos et al. (2017), na abordagem GAMLSS, essa comparação pode ser feita baseada nos valores do desvio global de ambos os modelos. Logo, sendo $\mathcal{M}$ um modelo estatístico com um vetor de parâmetros $\boldsymbol{\theta}$. O valor do desvio global deste modelo é obtido por $\text{GDEV} = -2\ell(\hat{\boldsymbol{\theta}})$. Deste modo, para os modelos $\mathcal{M}_1$ e $\mathcal{M}_2$ com desvios GDEV_1 e GDEV_2, respectivamente, a estatística de teste é dada por

$$\Lambda = \text{GDEV}_1 - \text{GDEV}_2.$$

Sob a hipótese nula H_0 de que o modelo reduzido, isto é, $\mathcal{M}_1$ é adequado, a estatística Λ tem assintoticamente distribuição χ^2_d, em que d é dado pela diferença entre os graus de liberdade dos modelos em questão, ou seja, $d = v_1 - v_2$, em que v_1 e v_2 representam os graus de liberdade dos modelos $\mathcal{M}_1$ e $\mathcal{M}_2$, respectivamente.

De acordo com Stasinopoulos et al. (2017), para a comparação de modelos GAMLSS não encaixados, o critério de informação de Akaike generalizado pode ser utilizado. O critério é obtido adicionando uma penalidade κ ao desvio ajustado para cada grau de liberdade usado no modelo. O critério é dado por

$$\text{GAIC}(\kappa) = \text{GDEV} + \kappa v,$$

em que v denota o total de graus de liberdade do modelo. O modelo é selecionado a partir do menor valor do critério $\text{GAIC}(\kappa)$. O critério de informação de Akaike (AIC) e o critério de informação Baeysiano (BIC) são casos particulares do critério de informação de Akaike generalizado, em que, $\kappa = 2$ e $\kappa = \log(n)$ respectivamente.

Na abordagem GAMLSS, a seleção do modelo envolve várias etapas, tal como especificado em Rigby e Stasinopoulos (2005) e Stasinopoulos et al. (2017). Seja $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L}\}$, representando um modelo GAMLSS, as seguintes componentes de $\mathcal{M}$ devem ser especificadas:

i) a distribuição da variável resposta ($\mathcal{D}$);

ii) o conjunto de funções de ligação ($\mathcal{G}$);

iii) os termos que aparecem em todos os preditores para μ, σ, ν e τ ($\mathcal{T}$);

iv) o conjunto de hiper-parâmetros do modelo ($\mathcal{L}$).

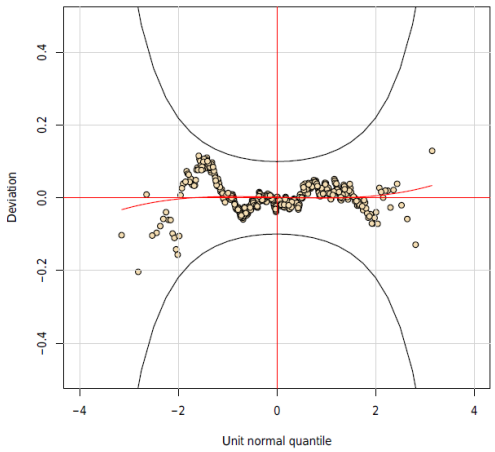
Stasinopoulos et al. (2017) afirmam que a escolha da distribuição deve ser feita em dois estágios: no estágio de ajuste do modelo e no estágio de diagnóstico do modelo. O estágio de ajuste envolve a comparação de modelos ajustados para diferentes distribuições plausíveis para Y_i usando o critério de informação de Akaike generalizado. Isto pode ser feito com auxílio da função `chooseDist` no pacote `gamlss` (STASINOPOULOS; RIGBY, 2007) que ajusta diferentes modelos para a variável resposta levando em consideração o efeito das variáveis explanatórias. Já o estágio de diagnóstico envolve a utilização de um gráfico (*worm plot*), que permite a detecção de inadequações do modelo de forma global ou em específicas amplitudes para uma ou duas variáveis explanatórias.

De maneira geral, o *worm plot* avalia a adequabilidade do modelo ajustado através da análise dos resíduos. Na Figura 3.1 é apresentado o *worm plot* obtido para um modelo bem ajustado. Os pontos no gráfico mostram o quão longe os resíduos ordenados do modelo estão de seus valores esperados (representados pela linha horizontal). Quanto mais próximos os pontos estão da linha horizontal, mais próxima a distribuição dos resíduos está de uma distribuição normal padrão (indicando um bom ajuste).

Além disso, para um bom ajuste, deve-se observar um percentual aproximado de 95% dos pontos dentro do intervalo de confiança, dado no gráfico pelas curvas elípticas. A forma da curva ajustada sob os pontos no *worm plot* pode indicar diferentes inadequações no modelo ajustado, conforme apresentado na Tabela 3.2.

A especificação da função de ligação é feita em primeiro momento de acordo com o intervalo de variação de cada parâmetro da distribuição. Porém, diferentes ligações podem ser utilizadas baseado no ponto de vista interpretativo e comparadas através do critério GAIC, como também durante a análise de diagnóstico.

Figura 3.1 – *Worm plot* para um modelo bem ajustado.



Fonte: (STASINOPOULOS et al., 2017).

Tabela 3.2 – Diferentes formas da curva ajustada no *worm plot*, irregularidades nos resíduos e no ajuste da distribuição.

Forma	**Resíduos**	**Distribuição ajustada**
nível: acima da origem	*média muito alta*	*locação estimada muito baixa*
nível: abaixo da origem	*média muito baixa*	*locação estimada muito alta*
linha: inclinação positiva	*variância muito alta*	*escala estimada muito baixa*
linha: inclinação negativa	*variância muito baixa*	*escala ajustada muito alta*
forma de U	*assimetria positiva*	*assimetria estimada muito baixa*
forma de U invertido	*assimetria negativa*	*assimetria estimada muito alta*
S com curva esquerda para baixo	*leptocurtose*	*calda da distribuição muito leve*
S com curva esquerda para cima	*platicurtose*	*calda da distribuição muito pesada*

Fonte: (STASINOPOULOS; RIGBY, 2007).

Stasinopoulos et al. (2017) consideram que a seleção de variáveis é uma das etapas mais importantes em um processo de análise de regressão. Seja $\boldsymbol{\chi}_k$ um conjunto de variáveis explanatórias disponíveis para a modelagem do vetor de parâmetros $\boldsymbol{\theta}_k$ de um modelo GAMLSS, em que $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$. $\boldsymbol{\chi}_k$ contém fatores e/ou termos quantitativos que podem entrar no modelo como termos lineares ou termos de suavização aditivos.

Na modelagem GAMLSS, a inclusão desses termos no modelo pode ser feita através dos procedimentos de seleção *backward*, *forward* e *stepwise*, levando em consideração o valor do critério GAIC em cada etapa do processo. No *backward*, o modelo é ajustado com todas as variáveis disponíveis e o processo passa a realizar a retirada de uma variável por vez, de forma que o critério de informação de Akaike generalizado seja minimizado. Em uma etapa do processo, o critério pode aumentar conforme seja retirada determinada variável. Neste momento o processo é finalizado, resultando no modelo final.

No *forward*, o procedimento é similar ao *backward*. Porém, na primeira etapa do processo o modelo é ajustado apenas com o intercepto e o processo passa a realizar a inclusão das variáveis de forma que o critério GAIC seja mínimo. Em um dado momento o critério pode aumentar conforme uma determinada variável seja incluída e neste passo, o procedimentos é finalizado.

No *stepwise*, o processo utilizado é uma uma junção do que foi visto para o *backward* e *forward*. Neste procedimento, as variáveis podem ser incluídas e/ou retiradas do modelo em cada etapa do processo, de tal forma que o critério GAIC seja minimizado. O processo pode iniciar apenas com o intercepto ou com todas as variáveis no modelo. Variáveis incluídas no modelo em etapas anteriores podem ser retiradas na atual etapa, assim como variáveis retiradas podem ser incluídas novamente. O processo é continuado até que o valor do GAIC aumente em determinada etapa, quando o processo é encerrado e o modelo final é, então, escolhido.

3.5 Estatística Espacial

Na Estatística, a grande maioria das análises realizadas leva em consideração que a variável aleatória em estudo ocorre de forma independente, porém, isso nem sempre é verídico, pois existem situações em que a variável aleatória possui uma estrutura de dependência. A Estatística Espacial é uma área da Estatística que tem como premissa a dependência da variável aleatória no espaço em que ela ocorre e, desse modo, a utilização da localização da variável aleatória se faz necessário para a análise.

Assim como dito anteriormente, a Estatística Espacial é dividida em três principais subáreas: Geosestatística, que trata de processos contínuos no espaço, Dados de Área, em que a variável aleatória é medida em regiões bem definidas e limitadas e Processo Pontual, em que se tem medidas precisas sobre a ocorrência de um evento do processo estocástico. Em alguns casos, um processo de padrões de pontos pode se estender para uma análise de dados de área, a critério do pesquisador. Cada subárea possui metodologias próprias e bem definidas para a análise do fenômeno em questão. Neste trabalho, focaremos apenas no estudo de Dados de Área.

3.5.1 Análise de Dados de Área

De acordo com Assunção (2001), análise de Dados de Área refere-se ao mapa geográfico de uma determinada região R particionada em áreas $A_1, \ldots, A_n$, de modo que $\bigcup_{i=1}^{n} A_i = R$ e $A_i \cap A_j = \emptyset$ se $i \neq j$. Em cada área A_i mede-se uma ou mais variáveis aleatórias Y_i e se possível, covariáveis que supostamente estejam relacionadas com a distribuição de Y_i. As áreas A_i, $i = 1, \ldots, n$, são delimitadas por um conjunto de coordenadas que formam um polígono. A variável aleatória medida em cada área refere-se a toda região compreendida pela área A_i e não somente a um ponto específico.

Assim como qualquer análise estatística, a análise de dados de área pode ser precedida por uma análise exploratória antes da avaliação da dependência espacial propriamente dita. Uma maneira de realizar essa análise é por meio da construção de mapas coropléticos do atributo na região de estudo, o que permite a visualização do comportamento do fenômeno na região.

3.5.2 Matriz de Proximidade Espacial

Segundo Fischer e Wang (2011) um aspecto crucial para a definição de autocorrelação espacial é a determinação de locais próximos, ou seja, locais que se encontram ao redor de um ponto de dados, que podem estar correlacionados. O primeiro passo para se avaliar essa correlação espacial é a definição de uma matriz que quantifica a relação de vizinhança entre áreas em estudo.

De acordo com Waller e Gotway (2004), a determinação da coleção de pesos espaciais é feita por meio da *matriz de proximidade espacial* (também conhecida como matriz de vizinhança, matriz de conectividade ou matriz de pesos espaciais). O (i,j)-ésimo elemento de uma matriz de proximidade espacial $\mathbf{W}$, denotado por w_{ij}, quantifica a dependência espacial entre as regiões i e j, e coletivamente, define uma estrutura de vizinhança em toda a área. De maneira geral, a matriz de proximidade é da forma

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

Em muitos estudos, é suficiente adotar a matriz de conectividade binária para caracterização de vizinhança entre as áreas. Esta matriz representa a definição mais simples de vizinhança. As áreas são então consideradas como vizinhas se apresentarem uma aresta em comum. Deste modo, para cálculo dos valores da matriz de vizinhança é utilizado o seguinte critério.

$$\begin{cases} w_{ij} = 1, & \text{se a área } i \text{ compartilha uma fronteira comum com a área } j \\ w_{ij} = 0, & \text{caso contrário.} \end{cases}$$

Pode-se perceber que a matriz de proximidade obtida é uma matriz simétrica de ordem n, pois $w_{ij} = w_{ji}$ e também, $w_{ii} = 0$. Existem diversas outras maneiras para se atribuir os pesos da matriz de proximidade espacial $\mathbf{W}$, conforme pode ser visto em Waller e Gotway (2004).

3.5.3 Autocorrelação Espacial

A autocorrelação espacial é uma das maneiras de se avaliar a dependência da variável em estudo no espaço em que ela se encontra. O objetivo é identificar se existe similaridade nos valores da variável Y_i em relação as áreas A_i. Para Fischer e Wang (2011) o conceito da análise espacial está na pressuposição de que valores de uma variável em uma determinada área são mais similares aos valores das áreas vizinhas. Isto é explicado pela primeira lei da geografia de Tobler, que afirma que, *"Todas as coisas são parecidas, mas coisas mais próximas se parecem mais que coisas mais distantes"* (TOBLER, 1970).

A autocorrelação espacial busca analisar a covariância ou correlação entre as observações de áreas vizinhas para uma determinada variável de interesse. Muitas medidas têm sido propostas na literatura para verificar a existência de dependência espacial. Duas delas são muito conhecidas e amplamente utilizadas: o Índice I de Moran (MORAN, 1950) e a Estatística c de Geary (GEARY, 1954).

O Índice de Moran avalia a dependência espacial por meio de um produto cruzado de desvios em relação à média e é dado da seguinte forma

$$I = \frac{n}{w_0} \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

com fator de normalização

$$w_0 = \sum_{i=1}^{n} \sum_{j \neq i}^{n} w_{ij}, \tag{3.11}$$

em que I é o valor do índice de Moran, w_{ij} corresponde ao valor da matriz de proximidade na i-ésima linha e j-ésima coluna, y_i é o valor da variável na área i, y_j é o valor da variável na área j e $\bar{y}$ é a média da variável aleatória em estudo.

De acordo com Almeida (2012), o valor esperado para índice de Moran, sob a hipótese de independência é $\Big[-1/(n-1)\Big]$. O valor calculado do índice deve ser igual ao valor esperado (dentro de uma significância estatística) se y_i for independente dos valores nas regiões vizinhas. Valores de I que excedem o valor esperado indicam autocorrelação espacial positiva (indicando similaridade dos valores do atributo no espaço). Já valores de I abaixo do valor esperado indicam autocorrelação espacial negativa (indicando dissimilaridade dos valores do atributo no espaço).

A estatística c de Geary usa a diferença dos quadrados para medir o grau de associação entre os valores da variável levando-se em consideração as áreas, obtida da seguinte forma

$$c = \frac{(n-1)}{2w_0} \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

em que w_0 é dado pela Equação (3.11).

O valor da estatística de Geary varia no intervalo (0,2), ao passo que seu valor esperado é 1, sob a hipótese de independência. Valores menores que seu valor esperado indicam autocorrelação espacial positiva, enquanto que valores maiores que seu valor esperado indicam autocorrelação espacial negativa. Deste modo, valores de c no intervalo (0,1) estatisticamente significativos, indicam concentração espacial dos dados, ao passo que valores de c no intervalo (1,2) indicam dispersão espacial dos valores da variável de interesse (ALMEIDA, 2012).

Se os dados estiverem concentrados (autocorrelação espacial positiva), isto significa que regiões com altos valores para a variável em estudo estão próximas à regiões também com altos valores, ou regiões com baixos valores estão próximas à regiões com baixos valores. Por outro lado, se os dados estiverem dispersos (autocorrelação espacial negativa), isto significa que regiões com altos valores da variável em estudo estão próximas à regiões com baixos valores, ou regiões com baixos valores são vizinhas de regiões com altos valores.

Apesar do índice de Moran global ser uma ótima medida para detectar a existência de autocorrelação espacial para as áreas em estudo, diferentes regimes de associação espacial podem ocorrer para determinadas localidades dentro da região de estudo, quando esta apresenta muitas áreas. Nesses casos, determinadas regiões podem apresentar um grau de associação espacial maior que as demais. Dessa forma, o índice de Moran global de forma isolada não é capaz de captar com eficácia tais associações locais. Para tanto, medidas locais devem ser definidas no intuito de captar essas associações. Uma delas é o índice de Moran local.

De acordo com Anselin (1995), o índice de Moran local é uma medida capaz de avaliar a autocorrelação espacial local entre as áreas A_i e seus vizinhos, podendo medir o grau de similaridade dos valores da variável aleatória entre essas áreas. A estatística do índice de Moran local para as áreas A_i, com $i = 1, \ldots, n$ é dada por

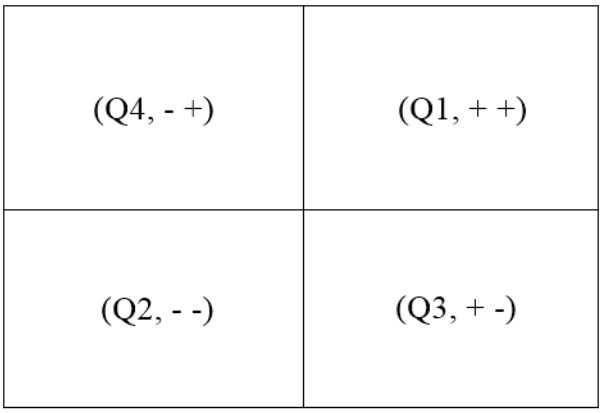
$$I_i = (y_i - \bar{y}) \sum_{j \in J_i} w_{ij}(y_j - \bar{y})^2,$$

em que J_i denota o conjunto de vizinhos da área i e o somatório em j é executado apenas nas áreas pertencentes à J_i, $\bar{y}$ denota a média dessas observações vizinhas.

Uma maneira alternativa de visualizar a autocorrelação espacial é por meio do diagrama de dispersão de Moran (Moran *Scatterplot*). Esse gráfico, mostra a defasagem espacial da variável de interesse no eixo das ordenadas e os valores da variável no eixo das abscissas. O diagrama de dispersão de Moran é dividido em quatro quadrantes conforme apresentado na Figura 3.2. No primeiro quadrante estão localizadas as áreas que possuem altos valores para a variável, rodeados por vizinhos com altos valores (Q1, $++$). No segundo quadrante estão as áreas que possuem baixos valores para a variável, rodeados por vizinhos com baixos valores (Q2, $--$). No terceiro quadrante localiza-se as áreas com altos valores, rodeadas por vizinhos com baixos valores (Q3, $+-$) e no quarto quadrante estão as áreas que possuem baixos valores da variável, rodeadas por vizinhos com altos valores (Q4, $-+$).

É também possível construir um mapa da região de estudo com os valores organizados pelos quadrantes do diagrama de dispersão de Moran, de forma que seja possível observar em que quadrante estão agrupadas determinadas áreas. É possível também construir um mapa que indique em quais regiões a autocorrelação local é ainda mais sig-

Figura 3.2 – Representação gráfica do diagrama de dispersão de Moran.



Fonte: Elaboração própria.

nificativa, baseado nos valores do índice de Moram local. Este mapa é muito conhecido como LISA (*Local Indicators of Spatial Association*) *map*.

3.5.4 Modelos de Regressão Espacial para Dados de Área

Um modelo de regressão baseia-se em uma equação matemática que estabelece uma relação entre uma variável resposta Y_i e uma ou mais variáveis explanatórias $\mathbf{x}_i$. Um modelo de regressão bastante simples é expresso na Equação (3.1). Em muitos desses modelos é comum assumirmos que as observações sejam independentes e identicamente distribuídas (*i.i.d*). Todavia, de acordo com Bivand, Pebesma e Gomez-Rubio (2013), quando estamos lidando com dados espaciais essa afirmação pode não ser verdadeira, uma vez que as observações podem não ser independentes, pois existe a possibilidade da existência de alguma correlação entre áreas vizinhas.

Yin et al. (2018) afirmam que o uso do modelo visto em (3.1) não é apropriado para avaliar a relação entre as variáveis, quando estamos trabalhando com dados espacialmente dependentes. Para tanto, especificamente para a metodologia de dados de área, dois modelos têm sido amplamente utilizados para este fim, o modelo SAR (*Spatial Autoregressive* ou *Spatial Lag Model*) e o modelo SEM (*Spatial Error Model*).

Para Fischer e Wang (2011) o modelo SAR é uma extensão do modelo de regressão tal como visto em (3.1). Esse modelo permite que as observações da variável resposta Y_i na área i $(i = 1, \ldots, n)$ dependa de observações de áreas vizinhas j $(j \neq i)$. O modelo SAR pode então ser definido da seguinte forma

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \sum_{k=1}^{p} x_{ik}\beta_k + \varepsilon_i, \tag{3.12}$$

em que o termo de erro ε_i é independente e identicamente distribuído, segundo uma distribuição normal com média zero e variância constante, ou seja, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, w_{ij} é o (i,j)-ésimo elemento da matriz de pesos espaciais ou matriz de proximidade $\mathbf{W}$ e ρ é um parâmetro (a ser estimado) que irá determinar a intensidade da relação espacial autorregressiva entre y_i e $\sum_{j=1}^{n} w_{ij} y_j$. Em notação matricial, o modelo em (3.12) é escrito como

$$\boldsymbol{Y} = \rho \mathbf{W} \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$.

Anselin e Rey (1991) consideram que uma outra forma de se avaliar a possível dependência espacial entre as observações é incorporando um processo espacial autorregressivo ao termo de erro no modelo dado em (3.1), de modo que autocorrelação dos resíduos seja então considerada. Dessa forma o modelo fica dado por

$$y_i = \sum_{k=1}^{p} x_{ik}\beta_k + \lambda \sum_{j=1}^{n} w_{ij}\varepsilon_j + \delta_i, \tag{3.13}$$

que corresponde ao modelo linear normal, porém, com o termo de erro ε_i sendo dado por $\varepsilon_i = \lambda \sum_{j=1}^{n} w_{ij}\varepsilon_j + \delta_i$, em que λ é parâmetro espacial autorregressivo e δ_i corresponde ao termo de erro aleatório, com $\delta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Em notação matricial o modelo dado em (3.13) pode ser reescrito como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \lambda \mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{\delta},$$

em que $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$.

O modelo SEM pode ser visto como uma combinação entre o modelo de regressão padrão e o modelo espacial autorregressivo em termos do erro ε_i, e portanto, tem valor esperado igual ao modelo de regressão linear normal. Em grandes amostras, as estimativas de ambos os modelos serão semelhantes, no entanto, em pequenas amostras pode haver um ganho em eficiência ao se modelar corretamente a dependência espacial no termo de erro ε_i (FISCHER; WANG, 2011). Outra representação deste modelo é apresentada na subsessão

3.6.3, em que o modelo SEM é introduzindo partindo do conceito de Campos Aleatórios Markovianos Gaussianos e é chamado de modelo CAR (*Conditional Autoregressive*).

Outra alternativa para considerar a estrutura espacial na modelagem é apresentada em Bivand, Pebesma e Gomez-Rubio (2013). Os autores fazem menção ao modelo espacial de Durbin, o qual inclui o efeito espacial não só nas observações da variável aleatória Y_i, como também nas variáveis explanatórias, como segue

$$y_i = \sum_{k=1}^{p} x_{ik}\beta_k + \rho \sum_{j=1}^{n} w_{ij} y_j + \gamma \sum_{j=1}^{n} w_{ij} x_{jk} + \varepsilon_i. \tag{3.14}$$

Assim como nos modelos SAR e SEM, o modelo (3.14) também assume normalidade para os resíduos, ou seja, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Este modelo pode ser reescrito matricialmente como

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \rho \mathbf{W}\boldsymbol{Y} + \gamma \mathbf{W}\boldsymbol{X} + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$.

3.6 Campos Aleatórios Markovianos Gaussianos

Uma maneira alternativa de lidar com dados de área é considerar a estrutura espacial presente nos dados utilizando Campos Aleatórios Markovianos Gaussianos, do inglês *Gaussian Markov Random Fields* (GMRF). Os GMRF possuem uma sólida teoria com ampla aplicabilidade nas mais variadas áreas do conhecimento. De acordo com Rue e Held (2005), o uso de GMRF no campo da estatística espacial é bastante vasto, incluindo desde processos discretos, como análise de dados de área e padrão de pontos, à aproximações de GMRF na geoestatística para Campos Gaussianos.

A seguir será exposto de forma resumida a teoria que envolve o uso de um GMRF para o estudo de uma variável aleatória definida em uma área de interesse, também definida como *lattice* irregular, de modo a dar um embasamento teórico necessário para análises de dados dessa natureza por meio dos GMRF.

3.6.1 Independência Condicional

Para Rue e Held (2005) o conceito chave para o entendimento de um GMRF está no entendimento da estrutura de independência condicional. Por resultados de probabilidade

é sabido que duas variáveis Y_1 e Y_2 são ditas independentes se, e somente se, $\pi(Y_1, Y_2) = \pi(Y_1)\pi(Y_2)$, o que pode ser reescrito da forma $Y_1 \perp Y_2$. Dada uma terceira variável Y_3, essas duas variáveis (Y_1 e Y_2) são ditas condicionalmente independentes se, e somente se, $\pi(Y_1, Y_2|Y_3) = \pi(Y_1|Y_3)\pi(Y_2|Y_3)$, ou seja, $Y_1 \perp Y_2|Y_3$. De maneira geral, é possível calcular a densidade condicional de $\mathbf{Y}_A$, dado $\mathbf{Y}_{-A}$ da seguinte forma

$$\pi(\mathbf{Y}_A|\mathbf{Y}_{-A}) = \frac{\pi(\mathbf{Y}_A, \mathbf{Y}_{-A})}{\pi(\mathbf{Y}_{-A})} \propto \pi(\mathbf{Y})$$

em que $\mathbf{Y}_A$ representa o vetor aleatório $\mathbf{Y}$ cujos elementos pertencem a um determinado conjunto A e $\mathbf{Y}_{-A}$ representa a realização do vetor $\mathbf{Y}$ com uma restrição nos elementos do conjunto A.

3.6.2 Grafo não Direcionado

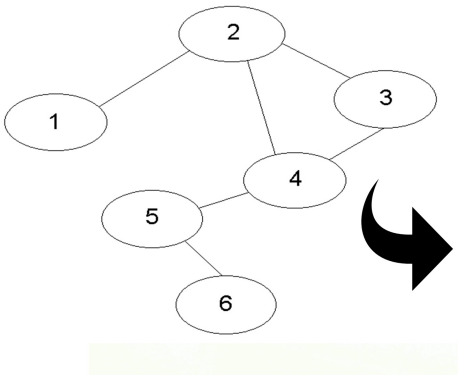
De acordo com Rue e Held (2005) um grafo não direcionado pode ser usado para representar a estrutura de dependência condicional em um GMRF. Um grafo não direcionado $\mathcal{G}$ é uma estrutura que consiste de um conjunto finito de vértices $\mathcal{V}$ e o um conjunto finito de arestas $\mathcal{E}$ desses vértices, compostas pelos pares $\{i, j\}$ em que $\{i, j\} \in \mathcal{V}$ e $i \neq j$. Dizemos que os vértices i e j são adjacentes se existe uma aresta entre eles, ou seja, $\{i, j\} \in \mathcal{E}$, caso contrário, os vértices são ditos não adjacentes (EDWARDS, 2012). Em muitos casos assumimos que $\mathcal{V} = \{1, \ldots, n\}$.

Em estudos de dados de área é comum representar a estrutura de vizinhança das áreas através da matriz de proximidade $\mathbf{W}$. Os valores dessa matriz podem ser obtidos a partir de um grafo não direcionado que representa a estrutura de vizinhança da região em estudo. Nesse caso, cada vértice do grafo representa uma área da região em estudo e $\mathbf{W}$ recebe valor 1 para o par correspondente as áreas i e j se os vértices que representam essas áreas são adjacentes e as áreas são então ditas vizinhas, caso os vértices sejam não adjacentes, as áreas não possuem vizinhança e a matriz $\mathbf{W}$ recebe valor 0. A seguir é representado na Figura 3.3 um exemplo de um grafo não direcionado e sua respectiva matriz de proximidade.

3.6.3 Modelos Autoregressivos Condicional (CAR) e Intrínseco (IAR)

Rue e Held (2005) afirmam que a grande utilidade dos GMRF deriva do fato de que muitas coisas, as quais estamos interessados em calcular são facilmente e rapidamente

Figura 3.3 – Grafo não direcionado e matriz de proximidade correspondente.



$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Fonte: Elaboração própria.

calculadas por meio de um GMRF. Os estudos pioneiros na construção dos modelos CAR (*Conditional Autoregressive*) foram introduzidos por Besag (1974) e Besag (1975). Segundo o autor, os modelos CAR são um caso particular de modelos GMRF, em que a matriz de precisão $\mathbf{G}$ é uma matriz não singular. O modelo CAR pode ser representado da seguinte forma

$$Y_i|\mathbf{y}_{-i} \sim \mathcal{N}\left(\sum_{j:j\neq i} \beta_{ij} y_j, k_i\right),$$

em que $\mathbf{y}_{-i} = \{y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\}$, $\beta_{ii} = 0$, $\beta_{ij} = -G_{ij}/G_{ii} (i \neq j)$ e $k_i = 1/G_{ii} > 0$. A simetria da matriz $\mathbf{G}$ garante que $\beta_{ij} k_j = \beta_{ji} k_i$.

De acordo com De Bastiani et al. (2018), no caso específico em que a matriz de precisão ($\mathbf{G}$) no modelo CAR é uma matriz singular, o modelo resultante é chamado de IAR (*Intrinsic Autoregressive*), um caso particular do modelo CAR.

3.7 O Modelo Autoregressivo Intrínseco na família GAMLSS

Em situações em que o fenômeno em estudo não ocorre de maneira independente, modelos de regressão que levam em consideração essa estrutura de dependência são de grande valia para lidar com esse tipo de dado, visto que, tais modelos podem incorporar em sua construção essa dependência presente nas observações antes ignorada, que muitas vezes pode ser devido ao efeito de tempo (dependência temporal), devido ao efeito do espaço em que ela ocorre (dependência espacial) ou até mesmo devido ao efeito de ambas as fontes de variação de forma simultânea (dependência espaço-temporal).

De Bastiani et al. (2018) apresentam uma formulação interessante, em que mesclam a teoria envolvida nos modelos GAMLSS com a estrutura espacial de um modelo Intrínseco Autorregressivo (IAR) para dados agrupados em regiões bem definidas. A inclusão da estrutura espacial nos modelos GAMLSS pode trazer grande melhoria nas estimativas dos parâmetros dos modelos, quando os dados apresentam uma dependência espacial. De acordo com os autores, o vetor aleatório $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top$ é chamado de GMRF, com respectivo grafo não direcionado $\mathcal{G}$, com média $\boldsymbol{\mu}$ e matriz de precisão $\lambda\mathbf{G}$, se e somente se, sua função densidade de probabilidade é dada da seguinte forma

$$\pi(\boldsymbol{\gamma}) \propto \exp\left[-\frac{1}{2}\lambda(\boldsymbol{\gamma}-\boldsymbol{\mu})^\top \mathbf{G}(\boldsymbol{\gamma}-\boldsymbol{\mu})\right] \tag{3.15}$$

e

$$G_{ij} \neq 0 \Longleftrightarrow \{i,j\} \in \mathcal{E}, \quad \text{para } i \neq j,$$

em que G_{ij} é o elemento da matriz de precisão $\mathbf{G}$ para a linha i e coluna j.

De acordo com Rue e Held (2005), γ_i e γ_j são condicionalmente independentes se, dado um γ_r, para todo r diferente de i e j, se e somente se, $G_{ij} = 0$. Devido à estrutura da matriz de precisão, os autores definem o vetor aleatório conforme um modelo CAR como segue,

$$\gamma_i | \boldsymbol{\gamma}_{-i} \sim \mathcal{N}\left(\sum_j \beta_{ij}\gamma_j, k_i\right),$$

em que $\boldsymbol{\gamma}_{-i} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_q)$, $\beta_{ii} = 0$, $\beta_{ij} = -G_{ij}/G_{ii} (i \neq j)$ e $k_i = 1/(\lambda G_{ii}) > 0$, para $i = 1, \ldots, q$. Então, conforme em De Bastiani et al. (2018), usando o Lema de Brook (BROOK, 1964), pode ser mostrado que a distribuição conjunta de $\boldsymbol{\gamma}$ é da forma como dado em (3.15) com média $\boldsymbol{\mu} = \mathbf{0}$, fornecendo, $\beta_{ij}k_j = \beta_{ji}k_i$, o que significa que a matriz $\mathbf{G}$ é simétrica.

A implementação de como um modelo IAR específico é incorporado ao modelo GAMLSS é dada em De Bastiani et al. (2018) da seguinte forma. Seja $\mathbf{W}$ uma matriz de proximidade (que assume-se ser simétrica) em que os elementos $w_{ii} = 0$ e $w_{ij} = 1$ se i e j $(i \neq j)$ compartilham uma borda comum, ou 0 caso contrário. A matriz de precisão $\mathbf{G}$ pode ser obtida como $\mathbf{D}_w - \mathbf{W}$, em que $\mathbf{D}_w$ é uma matriz diagonal com cada elemento da diagonal correspondendo a soma dos elementos da respectiva linha da matriz

de proximidade. Se tomarmos como exemplo a matriz de proximidade na Figura 3.3, teremos o seguinte o resultado para a matriz de precisão $\mathbf{G}$.

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \Longrightarrow \mathbf{D}_w = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

e portanto,

$$\mathbf{G} = \mathbf{D}_w - \mathbf{W} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

O efeito da matriz $\mathbf{G}$ é aproximar valores ajustados das regiões vizinhas juntos (em vez de reduzi-los à média geral, como é o caso de um termo simples do modelo de efeito aleatório). Perceba que a matriz $\mathbf{G}$ é tratada no GAMLSS como uma penalidade extra na função de log-verossimilhança penalizada ℓ_p (DE BASTIANI et al., 2018).

Assumindo que a variável resposta e as variáveis explanatórias sejam medidas e que suas observações pertençam espacialmente a um conjunto de áreas ou regiões, zero, uma ou mais de uma observação pode ser registrada em cada região, caracterizando um processo de análise espacial discreto agrupado por áreas.

Para incorporar modelos IAR em modelos GAMLSS, tome $\boldsymbol{Z}$ uma matriz índice que indica que observação pertence a que área, e tome $\boldsymbol{\gamma}$ um vetor de q efeitos espaciais aleatórios e assumindo que $\boldsymbol{\gamma} \sim \mathcal{N}_q\left(\mathbf{0}, \lambda^{-1}\mathbf{G}^{-1}\right)$, em que $\mathbf{G}^{-1}$ é uma inversa generalizada da matriz $\mathbf{G}_{q\times q}$. No seguinte modelo IAR baseado na Equação (3.15), a matriz $\mathbf{G}$ contém a informação sobre os vizinhos (regiões adjacentes), com elementos dados por $G_{ii} = n_i$, em que n_i representa o número total de regiões adjacentes à região i e $G_{ij} = -1$ se a região i e j são adjacentes, e zero caso contrário, para $i = 1, \ldots, q$ e $j = 1, \ldots, q$. Este

modelo tem a propriedade atrativa que depende de λ e γ_j para todo $j \neq i$, então $\gamma_i \sim \mathcal{N}\Big(\sum \gamma_j n_i^{-1}, (\lambda n_i)^{-1}\Big)$, em que o somatório é sob todas as regiões que são vizinhas da região i.

O padrão diferente de zero na matriz $\mathbf{G}$ determina o grafo $\mathcal{G}$. Um valor diferente de zero na matriz $\mathbf{G}$ indica que existe uma conexão entre as duas áreas correspondentes no grafo $\mathcal{G}$ (eles são vizinhos conectados). O valor zero na matriz $\mathbf{G}$ indica que não há conexão entre as duas áreas no grafo $\mathcal{G}$, o que implica que os efeitos espaciais correspondentes γ_i e γ_j para as duas regiões são condicionalmente independentes.

4 APLICAÇÃO 1: ESTUDO DE CASOS DE TUBERCULOSE BOVINA NO ESTADO DE MINAS GERAIS ENTRE OS ANOS 2014-2018

Resumo

A Tuberculose bovina (bTB) é uma doença infecciosa de importante impacto na saúde pública e na economia de qualquer região de sua ocorrência. A doença é causada pelos agentes pertencentes ao complexo *Mycobacterium tuberculosis* em maior escala pelo *Mycobacterium bovis*. A bTB pode acometer diversas espécies de animais domesticados e algumas espécies de animais selvagens, podendo também ser transmitida ao homem por meio destes hospedeiros. O estudo da zoonose é de suma importância, uma vez que a mesma traz diversos prejuízos para a população, envolvendo problemas na saúde animal e humana, o que vai de encontro à iniciativa *One Health* que visa o tratamento da saúde de humanos, animais e do meio ambiente de forma unificada. Neste sentido, o objetivo do presente estudo foi identificar o modelo de melhor ajuste para os casos de Tuberculose bovina no estado de Minas Gerais através dos modelos da família GAMLSS. Na análise espacial descritiva foi possível constatar que a bTB esteve mais presente na região oeste do estado de Minas Gerais, com maiores contagens nos municípios pertencentes as regiões do Alto Paranaíba, Triângulo Mineiro e Noroeste. Campo Florido apresentou o maior número de animais com diagnóstico positivo para a bTB, sendo que, provavelmente, este fato se deve ao maior número de testes realizados em uma propriedade deste município com ativa exportação de bovinos. Identificou-se que os casos de bTB no estado de Minas Gerais não apresentaram autocorrelação espacial significativa. Os modelos de regressão mostraram que, há um aumento no número de animais com diagnóstico positivo em localidades com maior efetivo bovino e maior número de vacas ordenhadas. O maior problema enfrentado neste trabalho foi sem dúvida, a falta de informação sobre a doença em muitos municípios (*missing value*), o que dificulta a avaliação da real situação da doença no estado. Além disso, há uma grande chance de que o número de animais com resultado positivo em muitos destes municípios esteja subnotificado, uma vez que as notificações são voluntárias.

Palavras-chave: Modelagem estatística. *One health*. Zoonose.

4.1 Introdução

A Tuberculose bovina é uma doença infecciosa causada por membros do complexo *Mycobacterium tuberculosis*, em maior grau causada pelo agente *Mycobacterium bovis*, acometendo diversas espécies de animais domésticos e algumas espécies de animais selvagens. A doença pode ser transmita ao homem através destes hospedeiros, sendo portanto, caracterizada como uma doença de forte impacto na saúde animal e humana. O estudo da doença está no contexto da iniciativa *One Health*[1], que tem como premissa a interação entre a medicina veterinária e humana, de forma que os cuidados à saúde humana, de animais e do meio ambiente sejam tratados de maneira unificada.

[1] https://onehealthinitiative.com/about/

Para Santos et al. (2018), a Tuberculose bovina é uma doença de alta relevância na saúde pública e na economia de um país. Isto deve-se aos gastos relacionados à programas de controle e erradicação da doença, como o Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose Animal (PNCEBT) no Brasil. Além disso, têm-se grandes perdas nas produções pecuárias do país devido ao sacrifício de animais doentios, haja vista que não há tratamento eficaz para mesma.

A Tuberculose bovina tem sido alvo de muitos estudos científicos que visam investigar o comportamento da doença, as diversas formas de contágio e identificar fatores associados ao aumento da mesma no intuito de contribuir com as autoridades responsáveis, propondo medidas mitigadoras para o controle e erradicação da doença. Dentre os vários estudos, pode-se citar: Belchior et al. (2016), Srinivasan et al. (2018), Campos (2019), Souza Filho (2019), Heijden et al. (2020), Conceição et al. (2020) e Lima et al. (2020). No entanto, não foi encontrado na literatura estudos voltados à investigação da Tuberculose bovina por meio dos modelos GAMLSS, especialmente no Brasil e no estado de Minas Gerais.

Por se tratar de uma doença assistida por um programa de controle oficial e com baixa prevalência de animais positivos (0,56%) no estado de Minas Gerais (BARBIERI et al., 2016), a base de dados de bovinos acometidos com Tuberculose bovina nos municípios do estado apresenta diversas características que podem trazer problemas durante a análise de regressão, tais como elevada frequência de zeros na amostra e uma variabilidade extremamente alta. O que torna muito difícil o tratamento adequado desses dados através de modelos mais comuns.

Neste sentido, o presente trabalho se justifica por apresentar uma proposta de tratamento dos dados de Tuberculose bovina nos municípios do estado de Minas Gerais por meio dos modelos da família GAMLSS, buscando uma modelagem mais adequada dos dados frente aos diversos problemas que podem ser enfrentados devido à complexidade da distribuição assumida para variável resposta. Estes modelos podem também ser aplicados não só a diversos outros problemas da epidemiologia, como também nas mais variadas áreas do conhecimento.

4.2 Revisão Bibliográfica

Em 1882 era descoberta a bactéria causadora da Tuberculose humana (TB), pelo Dr. Heinrich Hermann Robert Koch, a qual ele chamou de bacilo da Tuberculose. Em 1905, Koch recebera o prêmio nobel pela descoberta que inovaria o modo como a medicina tratava a doença na época. Em 1898, Theobald Smith descobriu que havia uma diferença na bactéria causadora da doença em humanos e bovinos. Em sua publicação, Smith demonstrou as diferenças entre esses dois agentes, atualmente conhecidos como *Mycobacterium tuberculosis (M. tuberculosis)* agente causador da Tuberculose humana e *Mycobacterium bovis (M. bovis)* agente causador da Tuberculose bovina (bTB).

A Organização Mundial da Saúde Animal (OIE)[2] define a Tuberculose Bovina como uma doença crônica de animais causada por membros do complexo *Mycobacterium tuberculosis* principalmente por *M. bovis* , mas também por *M. caprae* e, em menor grau, por *M. tuberculosis* . É uma das principais doenças infecciosas do gado, e também afeta outros animais domesticados e certas populações de animais silvestres, causando um estado geral de doença, pneumonia, perda de peso e eventual morte. A bTB é caracterizada como uma zoonose, devido ao fato de poder ser transmitida dos animais para o homem.

Ainda segundo a OIE, o *Mycobacterium bovis* foi isolado de inúmeras espécies selvagens, incluindo búfalos africanos, búfalos asiáticos domésticos, bisontes, ovelhas, cabras, equinos, camelos, porcos, javalis, veados, antílopes, cães, gatos, raposas, martas, texugos, furões, ratos primatas, lhamas, kudus, elands, antas, alces, elefantes, sitatungas, órixes, addaxes, rinocerontes, gambás, esquilos, lontras, focas, lebres, toupeiras, guaxinins, coiotes e vários felinos predadores, incluindo leões, tigres, leopardos e lince.

[1]A doença é contagiosa e pode ser transmitida de forma direta, pelo contato com animais infectados (domésticos e selvagens) ou indiretamente, pela ingestão de material contaminado. Nos bovinos a principal forma de transmissão é pelas vias respiratórias, contudo, bezerros podem ser infectados pela ingestão de leite de vacas contaminadas. Nos humanos a infecção pode acontecer pela ingestão de leite cru ou pelo contato com tecidos infectados. Os sinais clínicos usuais são: fraqueza, perda do apetite e peso, febre flutuante, dispneia e tosse intermitente, sinais de pneumonia de baixo grau, diarreia, linfonodos proeminentes e aumentados.

[2] https://www.oie.int/en/animal-health-in-the-world/animal-diseases/Bovine-tuberculosis/

Souza Filho (2019) infatiza que a Tuberculose bovina gera diversos prejuízos aos produtores de rebanho bovino, como queda na produção de leite, perdas na produção de carne e perda referente à condenação de carcaças com lesões de tuberculose no abatedouro, além de representar uma barreira no livre comércio de animais e produtos de origem animal. Esses problemas de forma generalizada e aliados aos gastos relacionados aos programas de controle da bTB se estendem obviamente para um problema econômico no país.

4.2.1 Cenário Mundial

A Tuberculose bovina é encontrada no mundo todo, apesar de alguns países nunca terem detectado a doença. A bTB é uma doença listada na OIE e deve ser relatada à mesma, conforme seu Código Sanitário para Animais Terrestres. A doença apresenta maior prevalência nos países da África e da Ásia, podendo também ser encontrada em países da Europa e das Américas.

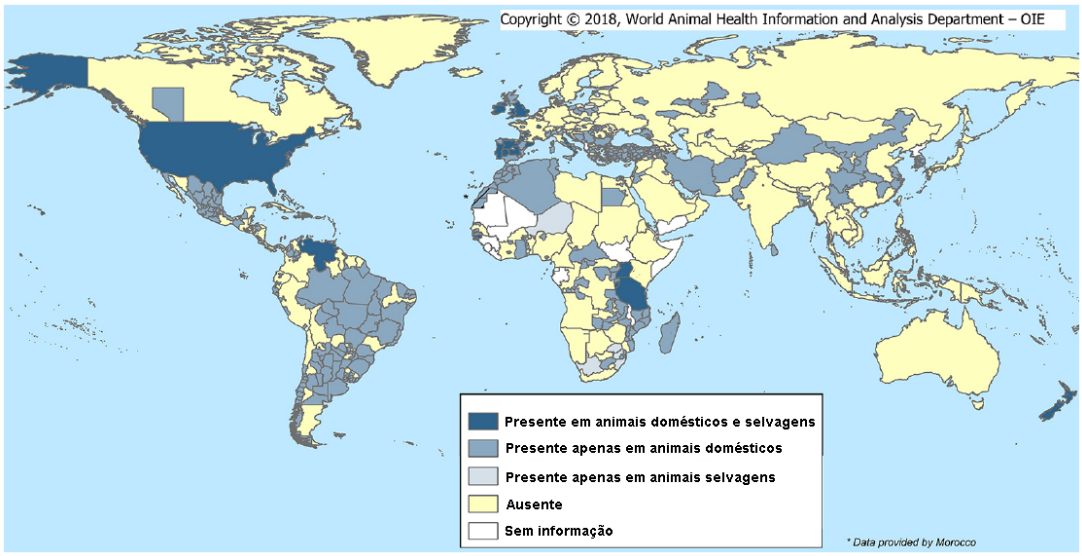
Entre janeiro de 2017 e junho de 2018, 188 países e territórios informaram à OIE sua situação em relação à Tuberculose bovina. Destes, 82 (que representa um percentual de 44%) relataram a presença da doença. O gado é o principal reservatório do *Mycobacterium bovis*, representando a principal fonte de infecção em humanos. Todavia, há relatos da presença da doença em outros animais, domesticados e não domesticados, tendo sido encontrada pela primeira vez em texugos selvagens, segundo Cassidy (2019).

Na Figura 4.1 é possível observarmos a distribuição geográfica da incidência global da Tuberculose bovina. Pode-se notar que, entre os anos de 2017 e 2018 muitos países não apresentaram incidência da zoonose e que em grande maioria deles a doença foi detectada apenas em animais domesticados.

De acordo com a OIE (2019), entre os 82 países afetados, 29 (35,4%) relataram a presença de bTB no rebanho de gado e em animais selvagens. Dois países (2,4%) relataram que o bTB está presente apenas na vida selvagem, enquanto que 51 desses países (62,2%) indicaram que apenas o gado foi afetado.

Embora a infecção em rebanhos bovinos tenha sido controlada em vários países, a eliminação completa da doença é muito complicada, devido ao impacto significativo da infecção persistente de animais selvagens, configurando um cenário de potenciais reserva-

Figura 4.1 – Distribuição global da Tuberculose bovina (bTB) entre 2017 e 2018.



Fonte: (OIE, 2019)

tórios na ausência da doença em bovinos e por isso, a bTB representa um sério problema para a saúde animal e humana.

De acordo com Srinivasan et al. (2018), em muitos países desenvolvidos, onde programas de controle nacional contra o Tuberculose bovina foram implementados, há um controle satisfatório da doença, embora a erradicação completa da bTB seja bastante desafiadora, devido ao potencial transbordamento de hospedeiros da vida selvagem. Em contrapartida, em alguns países em desenvolvimento, como a Índia por exemplo, a bTB permanece endêmica e muitas dificuldades são enfrentadas para o controle da doença, tais como os custos econômicos associados aos programas de controle da bTB.

Em um estudo sobre a prevalência da Tuberculose bovina em búfalos africanos, Heijden et al. (2020) constataram que o impacto de um programa de monitoramento da bTB a longo prazo com a remoção de animais que testam positivo é de fato significativo, resultando na diminuição da prevalência da doença no rebanho.

Para Conceição et al. (2020) a Tuberculose é ainda um dos maiores problemas de saúde pública, estimando que causa cerca de 140.000 novos casos e mais de 12.000 mortes em humanos em todo o mundo. Segundo os autores, um importante fator no controle da zoonose é o fato de que casos de Tuberculose humana causada pelo *M. bovis* é provavelmente subestimada.

4.2.2 A Tuberculose bovina no Brasil

O Brasil é um dos maiores países em extensão territorial, com alta representatividade na agropecuária no mercado internacional. De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE)[3], no ano de 2019 o Brasil apresentou um efetivo de rebanho bovino de 213.523.056 cabeças e 16.357.485 de vacas ordenhadas neste período. Ainda segundo o IBGE, o Departamento de Agricultura dos Estados Unidos (*United States Department of Agriculture*) estimou que em 2018 o Brasil apresentou o segundo maior rebanho mundial de bovinos, sendo o principal país exportador e o segundo maior produtor de carne bovina do mundo.

O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) instituiu em 2001 o Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT) que foi implementado com o objetivo de reduzir a prevalência de Brucelose e Tuberculose bovina no país, visando a erradicação dessas doenças. De acordo com o MAPA[4], a estratégia de atuação do programa é baseada na classificação das unidades federativas quanto ao grau de risco para essas doenças e na definição e aplicação de procedimentos de defesa sanitária animal, de acordo com a classificação de risco.

O MAPA[3] frisa ainda que a eficácia de um programa nacional de promoção de saúde animal está diretamente relacionada à qualidade e padronização dos meios de diagnósticos utilizados. Para o diagnóstico da bTB, o PNCEBT leva em consideração o Teste Cervical Simples (TCS), que é adotado como prova de rotina, o Teste da Prega Caudal (TPC), utilizado exclusivamente em gado de corte, também como prova de rotina e o Teste Cervical Comparativo (TCC) que pode ser utilizado como teste confirmatório dos dois anteriores, visando maior especificidade do diagnóstico da doença.

Apesar da instituição do PNCEBT em 2001, a situação do Brasil em relação a presença da Tuberculose bovina é bastante preocupante. O país vem acumulando altos índices da doença ao longo dos anos desde o início do seu acompanhamento em 1999, com destaque para os anos 2006, 2007, 2009, 2015 e 2019 que apresentaram os maiores números

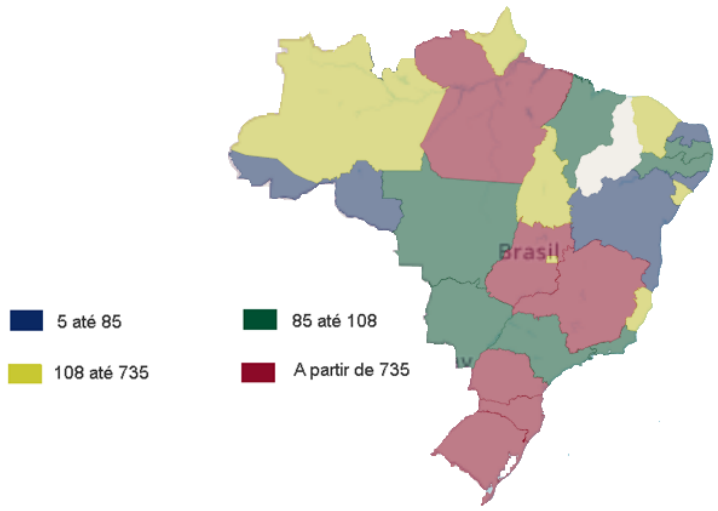
[3] https://cidades.ibge.gov.br/brasil/panorama

[4] https://www.gov.br/agricultura/pt-br/assuntos/sanidade-animal-e-vegetal/saude-animal/programas-de-saude-animal/controle-e-erradicacao-da-brucelose-e-tuberculose-pncebt

de casos confirmados da doença conforme pode ser visto no Sistema de Informação e Saúde Animal[5] do MAPA.

Na Figura 4.2, pode-se perceber que as regiões que concentraram a maior incidência da doença entre os anos de 1999 a 2019 foram as regiões Sul, Sudeste, Centro-Oeste e parte da região Norte, sendo estas as regiões com maior número de casos registrados (≥ 735). O número de animais contaminados com a bTB para ano de 2019 atingiu o maior patamar da série histórica com 7.172 animais infectados em todo o país.

Figura 4.2 – Distribuição da bTB no Brasil entre os anos de 1999 e 2019 (número de casos registrados).



Fonte: (MAPA, 2020)

Souza Filho (2019) afirma que no Brasil, a Tuberculose bovina é mais prevalente em rebanhos de leite, principalmente em fazendas com alta produção. Ainda de acordo com o autor, estudos conduzidos em 13 unidades federativas do Brasil mostraram que a doença é heterogênea dentro e entre os estados e que as maiores prevalências foram observadas no Espírito Santo, norte de São Paulo, sul de Minas Gerais e sudeste de Goiás. Nestes estados, a produção de leite foi identificada como maior fator de risco para a doença.

Segundo Lima et al. (2020) o diagnóstico e controle da Tuberculose bovina no Brasil são norteados pelo PNCEBT, contudo a presença de fatores inerentes aos testes utilizados para o diagnóstico da doença, como sensibilidade, especificidade, custo, tempo de execução, entre outros fatores, têm motivado estudos voltados à procura de ferramentas alternativas para o diagnóstico da bTB.

[5] http://antigo.agricultura.gov.br/assuntos/sanidade-animal-e-vegetal/saude-animal/sistema-informacao-saude-animal

4.2.3 Situação no Estado de Minas Gerais

Em 2018, o estado de Minas Gerais foi um dos 3 estados da federação que obteve o maior número para o efetivo de rebanho bovino com 21.810.311 cabeças, atrás apenas dos estados de Goiás com 22.651.910 cabeças e Mato Grosso com 30.199.598 cabeças, sendo este último estado, destaque nacional na criação de bovinos desde 2004 e apresentado no ano de referência o segundo maior valor da série, perdendo apenas para o ano de 2016, segundo dados do IBGE.

De acordo com o relatório da Pesquisa Pecuária Municipal (PPM, 2019), no ano de 2018, os estados do Mato Grosso, Goiás, Minas Gerais, Mato Grosso do Sul e Pará apresentaram respectivamente 14,1%, 10,6%, 10,2%, 9,8% e 9,7% do total nacional de rebanho bovino, representando 54,4% do efetivo nacional. Para o número de vacas ordenhadas, Minas Gerais foi o estado brasileiro com maior número de cabeças (3.147.732), seguido pelos estados de Goiás com 1.930.594 cabeças e Paraná com 1.356.589 cabeças, sendo o quarto estado brasileiro com maior produtividade de leite (2.840 litros/vaca/ano), atrás dos estados de Santa Catarina, Rio Grande do Sul e Paraná que apresentaram produtividades superiores à 3.200 litros/vaca/ano.

Em um estudo para avaliar a ocorrência de zoonoses em carcaças bovinas de animais abatidos em Uberaba-MG por meio de exames *post mortem*, entre 2006 e 2016, Campos (2019) verificou que a Tuberculose foi a zoonose mais presente, entre a doenças avaliadas, representando mais da metade das ocorrências (58,10%). De acordo com Instituto Mineiro de Agropecuária (IMA)[6], dos 853 municípios do estado, apenas 16 têm certificação de livres da Tuberculose animal.

Belchior et al. (2016) constataram que, entre 1.586 rebanhos de gado avaliados para 7 regiões do estado de Minas Gerais, 75 foram classificados como positivos para a Tuberculose bovina, apresentando uma prevalência moderada de 5,04% a nível de rebanho e 0,81% a nível de animal para a região em estudo. A região do Alto Paranaíba apresentou a maior prevalência (9,66%), sendo esta uma região leiteira recém formada. O Triângulo Mineiro, região com extensa produção de carne bovina foi a área com menor prevalência, com 2,08%.

Barbieri et al. (2016) investigaram os fatores de risco associados a ocorrência de animais positivos para a Tuberculose bovina no estado de Minas. O estudo avaliou 2.182 re-

[6] http://www.ima.mg.gov.br

banhos, em que 31.832 animais foram submetidos ao Teste Cervical Comparativo (TCC), onde 93 rebanhos testaram positivo para bTB, derivados de 188 animais positivos. O estudo encontrou prevalência moderada de bTB a nível de rebanho e baixa prevalência em animais. Os autores identificaram uma relação significativa entre rebanhos com 30 ou mais vacas, gado comprado de comerciantes de animais e o tipo de rebanho em relação à sua produção (carne ou leite) com o aumento na chance de observar animais positivos no rebanho.

4.3 Material e Métodos

A seguir será apresentada a metodologia utilizada para dar prosseguimento ao estudo dos casos de Tuberculose bovina no estado de Minas Gerais, bem como são apresentados a delimitação da área de estudo, a base de dados utilizada e os métodos estatísticos que foram empregados para a análise dos dados.

4.3.1 Área de Estudo

[7]O Estado de Minas Gerais é uma das 27 unidades da República Federativa do Brasil, na América do Sul. Está localizado na região Sudeste do Brasil, juntamente com os Estados do Espírito Santo, Rio de Janeiro e São Paulo. Seu território fica entre os paralelos 14º13'58"e 22º54'00"de latitude sul e os meridianos de 39º51'32"e 51º02'35"a oeste de Greenwich. Ocupa um fuso horário correspondente a -3 horas em relação a Greenwich.

Minas Gerais é o estado brasileiro com o maior número de municípios, somando 853 municípios em todo o estado. De acordo com o Instituto Brasileiro de Geografia Estatística (IBGE), possui uma área total de 586.521,121 km^2. O estado faz divisa com os estados de São Paulo, Rio de Janeiro, Espírito Santo, Bahia, Mato grosso do Sul, Goiás e Distrito Federal. No último censo realizado pelo IBGE (2010), tinha uma população de 19.597.330 habitantes e em 2019 foi estimada em 21.168.791 habitantes.

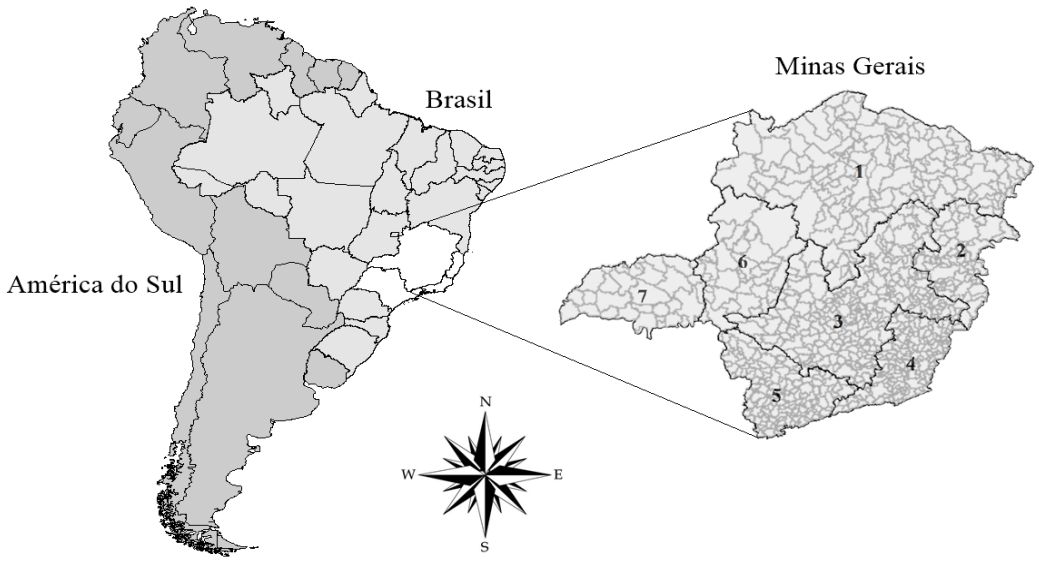
[8]O IBGE divide Minas Gerais em 12 mesorregiões e 66 microrregiões. De acordo com o órgão, este sistema de divisão tem aplicações importantes na elaboração de políticas públicas e no subsídio ao sistema de decisões quanto à localização de atividades econô-

[7] https://www.mg.gov.br/conteudo/conheca-minas/geografia/localizacao-geografica-0
[8] https://www.mg.gov.br/conteudo/conheca-minas/geografia/localizacao-geografica

micas, sociais e tributárias. Contribuem também, para as atividades de planejamento, estudos e identificação das estruturas espaciais de regiões metropolitanas e outras formas de aglomerações urbanas e rurais.

Gonçalves et al. (2009) dividem o estado em sete regiões que são caracterizadas de acordo com diferentes parâmetros relacionados à atividades da pecuária voltada a criação de bovinos, como sistemas de produção, finalidade da produção, manejo de animais, tamanho de rebanho e sistemas de comercialização. Esta mesma estratificação é apresentada em Barbieri et al. (2016). As regiões podem ser consultadas na Figura 4.3. A saber: 1 - Noroeste, Norte e Nordeste, 2 - Leste, 3 - Central, 4 - Zona da Mata, 5 - Sul e Sudeste, 6 - Alto Paranaíba e 7 - Triângulo Mineiro.

Figura 4.3 – Localização geográfica do Estado de Minas Gerais, divisões e seus respectivos municípios.



Fonte: Elaboração própria.

4.3.2 A Base de Dados

O trabalho está sendo desenvolvido em parceria com o Departamento de Medicina Veterinária da Universidade Federal de Lavras e o Instituto Mineiro de Agropecuária (IMA). O fenômeno em estudo é apresentado por meio da variável Controle de Focos da Tuberculose Animal (CFTA), que se refere ao número de bovinos com resultado positivo para a doença nos testes de diagnóstico nos municípios de Minas Gerais entre os anos de 2014 e 2018. Esses dados são provenientes das notificações recebidas pelo Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT)

fornecidos pelo IMA. Os animais foram considerados positivos para a Tuberculose bovina quando apresentaram resultado positivo nos testes de triagem e confirmatório.

Vale salientar que o número de casos de bTB nos municípios do estado de Minas Gerais obtidos dessa base de dados pode não retratar a real situação do estado em relação à Tuberculose bovina, posto que não há uma obrigatoriedade por parte das propriedades com bovinos em realizarem os exames nos animais, salvo em casos específicos como exportação ou venda de animais para outras localidades e atividades econômicas que envolve o manejo de leite bovino. Com isso, além de haver a possibilidade de subnotificação do número de bovinos com diagnóstico positivo para a bTB, este fato também gera um número alto de valores ausentes para muitos municípios na base de dados.

Também foram utilizadas as variáveis efetivo de rebanho bovino (ER) e número de propriedades com bovinos (PB) provenientes do relatório de informações gerais de bovinos, durante as etapas de vacinação contra a febre aftosa realizadas pelo Sistema de Defesa Agropecuária do estado de Minas Gerais. Estes dados foram cedidos pelo IMA.

O número de vacas ordenhadas (VO) nos anos 2014-2018, foi obtida da base de dados da Pesquisa da Pecuária Municipal (PPM)[9] realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Finalmente, também trabalhamos com o número de animais abatidos com lesões sugestivas à bTB (AIA) entre os anos 2014 e 2018, cujos dados foram cedidos pelo Serviço de Inspeção Federal (SIF), vinculado ao Departamento de Inspeção de Produtos de Origem Animal (DIPOA), da Secretaria de Defesa Agropecuária (SDA/MAPA).

Para as variáveis CFTA e AIA foi calculada a soma dos animais com resultado positivo para a bTB e animais abatidos com lesões sugestivas, respectivamente, nos municípios do estado entre os anos 2014-2018. Já para as variáveis efetivo de rebanho bovino, número de propriedades com bovinos e número de vacas ordenhadas optou-se por calcular a média nos municípios e nos anos em estudo.

4.3.3 Análise Estatística

A análise estatística se deu primeiramente por meio de uma análise exploratória dos dados, bem como observação da distribuição das variáveis e cálculo de medidas descritivas. Foram elaborados mapas coropléticos para observar a distribuição espacial dos casos de

[9] https://sidra.ibge.gov.br/pesquisa/ppm/tabelas

bTB nos municípios do estado de Minas Gerais, assim como, para avaliar a relação entre as variáveis CFTA e AIA, de modo a identificar possíveis discrepâncias desses dados nos municípios. Posteriormente, foi calculado o índice de I de Moran para ambas as variáveis, no qual identificou-se que não há autocorrelação espacial dos valores dessas variáveis no estado de Minas Gerais.

Foram ajustados modelos da família GAMLSS, no intuito de verificar a possível influência das covariáveis ER e VO na variabilidade da variável resposta CFTA. Para verificar a presença de correlação entre as variáveis explanatórias, foi utilizado um teste para o coeficiente τ de Kendall.

As análises desse estudo foram realizadas no *sotware* R (R Core Team, 2019) na versão 3.6.1, de código livre e gratuito com o auxílio do pacote `gamlss` (STASINOPOULOS; RIGBY, 2007) e `gamlss.dist` (RIGBY et al., 2019) e (STASINOPOULOS; RIGBY, 2020) para o ajuste dos modelos GAMLSS.

4.4 Resultados

Nesta sessão serão apresentados os resultados encontrados na análise dos dados de bovinos com resultado positivo nos testes de diagnóstico para a Tuberculose bovina nos municípios de Minas Gerais. Inicialmente, uma análise exploratória foi conduzida para a variável resposta, assim como para as variáveis explanatórias. Posteriormente, é realizado a investigação dos fatores associados aos casos de Tuberculose bovina por meio dos modelos GAMLSS.

4.4.1 Análise Exploratória

No total, foram registrados 2.246 animais diagnosticados como positivos nos testes para a bTB distribuídos em 152 municípios no estado de Minas Gerais entre os anos de 2014 e 2018. Campo Florido foi o município com o maior número de animais positivos, com 422, seguido pelos municípios de Unaí com 196, Patos de Minas com 127, Lagoa Formosa com 100 e Coromandel com 97. Os demais municípios tiveram contagens abaixo de 85 animais positivos. O alto número de animais diagnosticados como positivos no município de Campo Florido deve-se provavelmente à presença de uma propriedade exportadora de bovinos vivos, já que esse tipo de propriedade deve obrigatoriamente apresentar resultados em relação a ausência de bTB nos animais.

Tabela 4.1 – Estatísticas descritivas para as variáveis número de animais com diagnóstico positivo para a bTB (CFTA), número de propriedades com bovinos (PB), efetivo de rebanho bovino (ER), número de vacas ordenhadas (VO) e número de animais abatidos com lesões sugestivas (AIA).

Estatísticas	**CFTA**	**PB**	**ER**	**VO**	**AIA**
Mínimo	0,00	38,00	1441,00	18,00	1,00
Máximo	422,00	3114,20	352934,00	2500,00	367,00
1º quartil	1,00	300,75	15233,25	900,00	2,00
Mediana	2,00	516,50	28281,75	1402,0	6,00
3º quartil	10,25	767,50	48753,00	1925,67	28,00
Média	14,78	641,18	50737,84	1368,80	28,45
Desvio	41,73	518,97	63332,01	703,99	58,99
Assimetria	6,94	1,92	2,50	-0,21	3,71
Curtose	59,66	4,61	6,78	-0.97	15,41

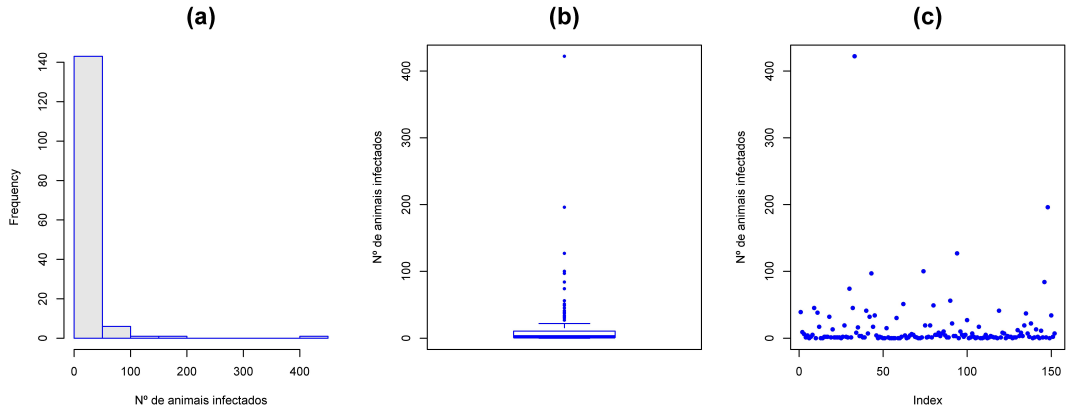
Fonte: Elaboração própria.

Conforme pode ser visto na Tabela 4.1, 75% dos municípios tiveram contagens inferiores a 10 para o número de animais diagnosticados como positivo para a bTB e 50% deles apresentaram contagens inferiores à 2, o que dá indícios de que a doença parece não apresentar alta prevalência no estado. Todavia, as mesmas estatísticas obtidas para o número de animais abatidos com lesões sugestivas a bTB sugere uma possível subnotificação do verdadeiro número de animais com diagnóstico positivo para a bTB.

A distribuição do número de animais positivos para a Tuberculose bovina (CFTA) apresenta assimetria a direita e características leptocúrticas. As variáveis explanatórias propriedades com bovinos (PB) e Efetivo de rebanho (ER) parecem apresentar distribuições levemente assimétricas e leptocúrticas. Já o número de vacas ordenhadas (VO) parece ter uma distribuição assimétrica, porém com característica platicúrtica. A seguir é exposto na Figura 4.4 a visualização gráfica da distribuição para a variável resposta (CFTA).

Na Figura 4.4(a) fica visível a alta frequência de valores no intervalo de zero a cinquenta, sendo o percentual de municípios com contagens dentro deste de intervalo igual a 94,08%, o que corresponde a 143 municípios. Enquanto que o percentual de municípios com informações nulas, ou seja, que não apresentaram diagnósticos positivos para a bTB dentro dos anos estudados foi de 19,08% aproximadamente, que corresponde a 29 municípios. Isto corrobora com as informações obtidas a partir da Tabela 4.1 em que grande parte dos municípios apresentam baixos valores para o número de animais diagnosticados como positivos para a bTB.

Figura 4.4 – Visualização gráfica para o número de animais com diagnóstico positivo para a bTB (CFTA) no estado de Minas Gerais.



Fonte: Elaboração própria.

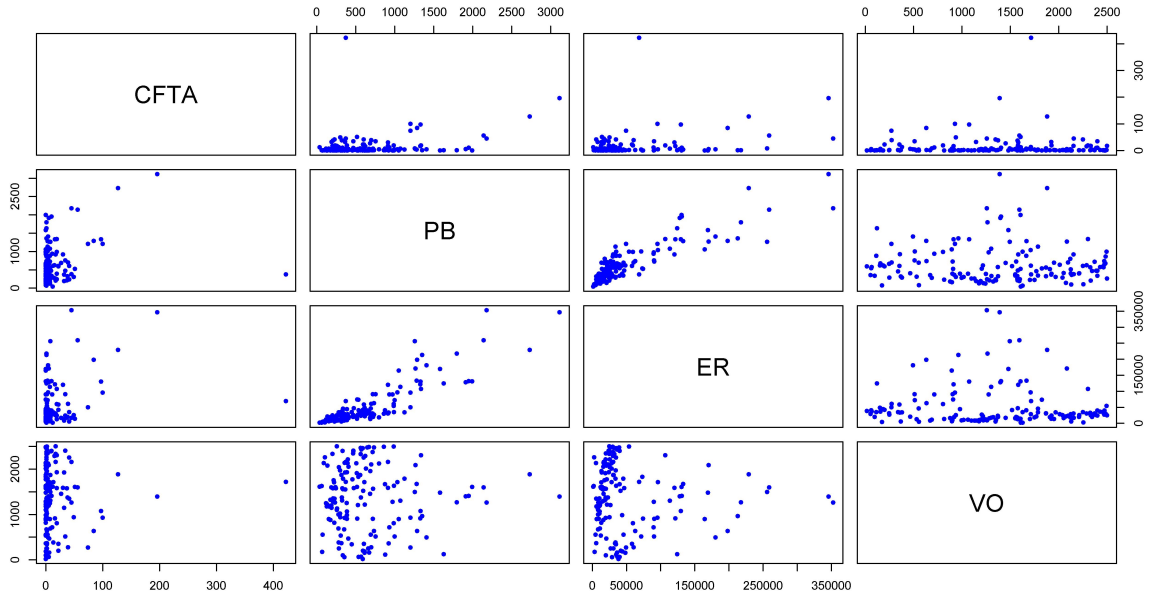
Na Figura 4.4(b) pode-se observar o alto número de observações atípicas, com destaque para o município de Campo Florido que apresentou a maior contagem de animais diagnosticados com a bTB. No gráfico de dispersão (Figura 4.4(c)) também fica visível o alto número de municípios com observações nulas e alta variabilidade para esta variável, impulsionada pelo alto número de observações atípicas.

A relação entre as variáveis explanatórias e o número de animais com diagnóstico positivo para a bTB foi realizada primeiramente através do gráfico de dispersão (Figura 4.5). Neste, é possível observar que, as variáveis explanatórias PB e ER parecem se comportar de forma similar para a relação com o número de animais positivos (CFTA). Nota-se também uma aparente correlação linear positiva entre as variáveis PB e ER. Isto pode acarretar em um possível problema de multicolinearidade durante a análise de regressão, isto é, uma das variáveis independentes pode ser obtida como combinação linear de outra. Para a relação entre o número de vacas ordenhadas e o número de animais positivos para a bTB não se pode enxergar um padrão específico.

Para confirmar a hipótese sob a presença de correlação linear entre as variáveis foi realizado um teste para o τ de Kendall, sob a hipótese nula de que as variáveis não são correlacionadas. O teste foi avaliado para a relação entre a resposta e as variáveis explanatórias, bem como para a relação apenas entre as variáveis explanatórias. Os resultados estão exibidos na Tabela 4.2.

Tal como sugerido pelo gráfico de dispersão na Figura 4.5, as variáveis independentes PB e ER apresentam de fato correlação linear significativa. Diante disto faz-se

Figura 4.5 – Gráfico de dispersão para a relação entre as variáveis PB, ER e VO com a variável CFTA.



Fonte: Elaboração própria.

necessário a escolha de uma delas para incluir durante a análise de regressão. Assim sendo, optou-se neste trabalho pela escolha da variável efetivo de rebanho bovino para compor a análise de regressão posteriormente. Ainda na Tabela 4.2 pode-se ver que apenas as variáveis PB e ER apresentam correlação significativa com a variável CFTA.

Tabela 4.2 – Teste para o τ de Kendall para a relação entre as variáveis em estudo.

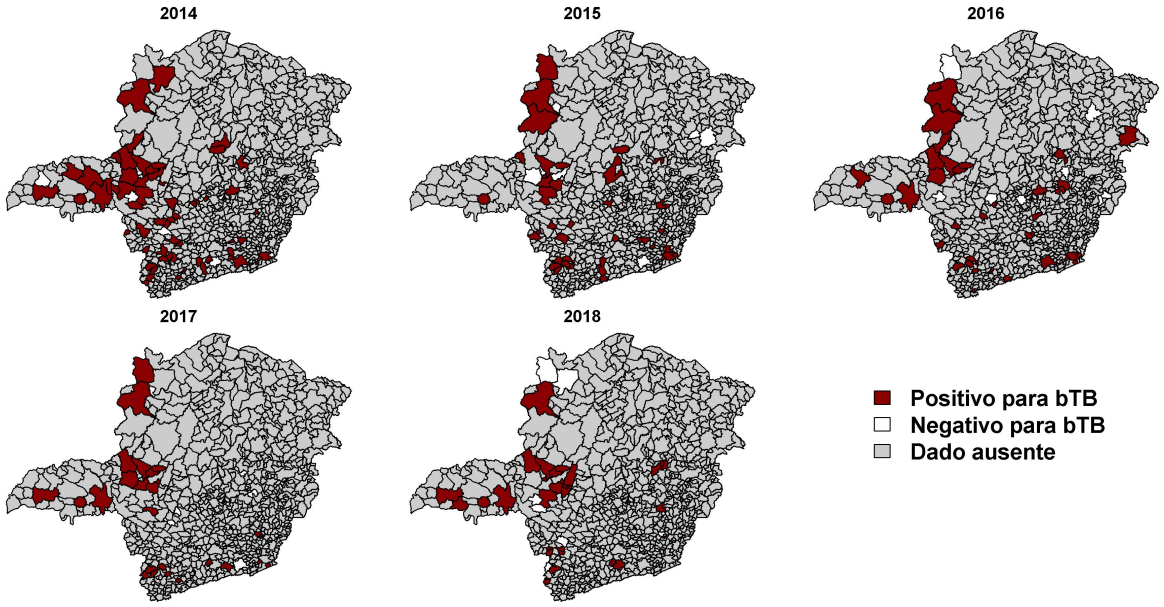
Teste para a relação com a variável resposta		
Teste	$\boldsymbol{\tau}$	**Valor** $\boldsymbol{p}$
CFTA × PB	0,1355	0,0176
CFTA × ER	0,1343	0,0186
CFTA × VO	0,0196	0,7308
Teste para a relação entre as variáveis explanatórias		
Teste	$\boldsymbol{\tau}$	**Valor** $\boldsymbol{p}$
PB × ER	0,6572	$< 0,001$
PB × VO	0,0017	0,9759
ER × VO	$-0,0055$	0,9251

Fonte: Elaboração própria.

Para verificar a forma de comportamento do número de animais com diagnóstico positivo para bTB na região de estudo de acordo com os anos estudados foram construídos mapas coropléticos (Figura 4.6), em que pode ser observado em quais municípios houve ao menos um resultado positivo para a Tuberculose bovina de acordo com cada

ano. De acordo com a Figura 4.6, fica claro que 2014 foi o ano com maior número de municípios com animais apresentando diagnóstico positivo para a Tuberculose bovina no estado mineiro. Todavia, é difícil realizar uma comparação com os demais anos posteriores, devido ao aumento no número de municípios com observações ausentes. Ver-se que, quando visualizamos os dados individualmente para cada ano, o número de municípios com contagens nulas, ou seja, que não apresentaram diagnósticos positivos para a bTB é bastante reduzido. Percebe-se também que, os municípios com bovinos positivos para a bTB concentraram-se na região oeste do estado, distribuindo-se com maior frequência nos estratos Noroeste, Alto Paranaíba e Triângulo Mineiro.

Figura 4.6 – Resultados positivos e negativos para os testes de Tuberculose bovina nos municípios do estado de Minas Gerais entre os anos 2014 e 2018.

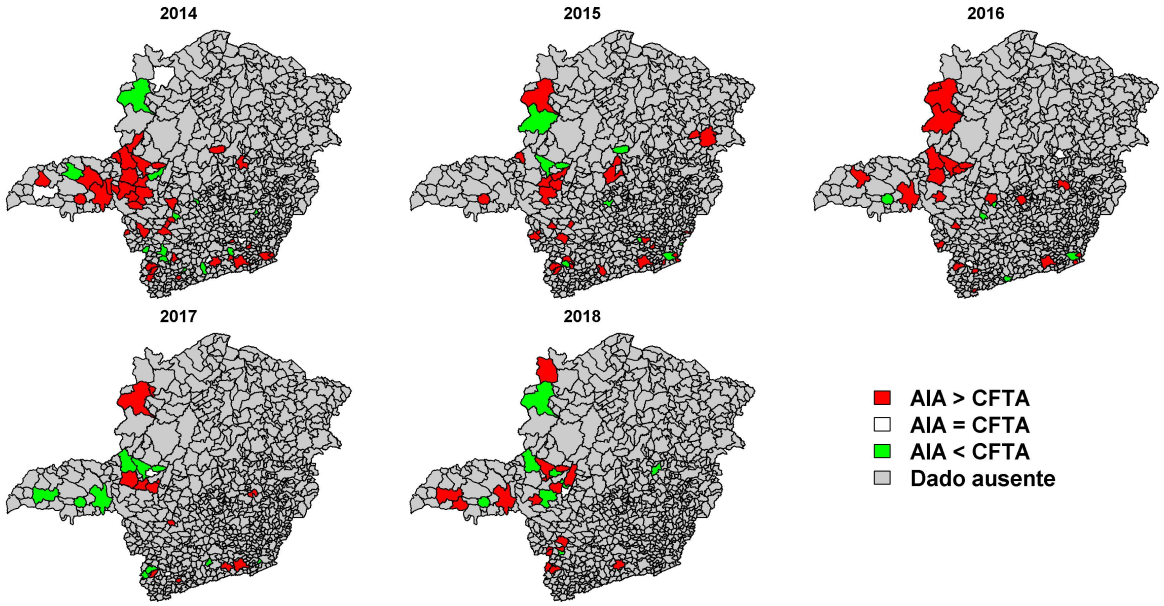


Fonte: Elaboração própria.

Um dos objetivos do presente trabalho é avaliar a relação entre os dados provenientes dos focos de Tuberculose bovina, ou seja, dos animais que testaram positivo para a doença e os dados provenientes dos abatedouros, que se refere ao número de animais abatidos com lesões sugestivas da Tuberculose bovina. A preocupação maior neste sentido é de um cenário de subnotificação do número de animais com diagnóstico positivo, devido a notificação desses dados ocorrer de forma voluntária, não retratando a real situação do estado em relação à doença. Neste sentido, foram construídos mapas coropléticos (Figura 4.7) com a comparação destes dados.

Os mapas foram gerados de forma individual para cada ano, para que se possa avaliar em quais anos houve maior discrepância entre esses dados. A cor a vermelha representa as localidade que tiveram maior número de animais abatidos com lesões sugestivas do que animais com diagnóstico positivo, enquanto que os municípios identificados pela cor verde são aqueles em que o número de animais abatidos foi menor que o número de animais positivos.

Figura 4.7 – Possíveis casos de subnotificação do número de animais positivos para a bTB nos municípios do estado de Minas Gerais entre os anos 2014 e 2018 a partir da relação AIA (número de animais abatidos com lesões sugestivas) e CFTA (número de animais com diagnóstico positivo).



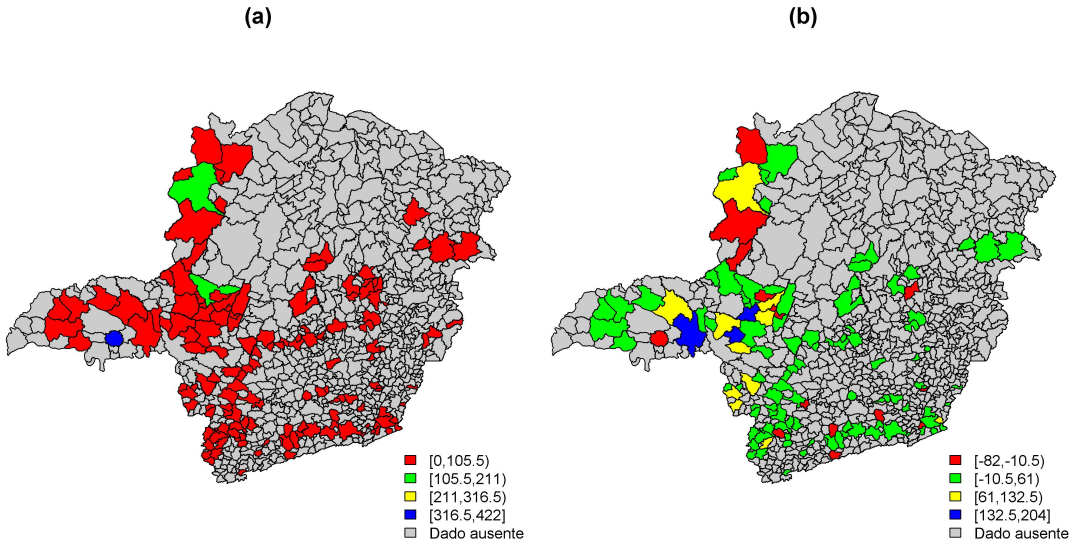
Fonte: Elaboração própria.

Na Figura 4.7, pode-se perceber que para os anos 2014, 2015 e 2016 parece ter havido maior subnotificação nos dados de CFTA. O que é bastante preocupante, uma vez que isto dificulta de forma significativa o combate à doença, haja vista o não conhecimento do real número de animais positivos nas regiões, facilitando o contágio de animais sadios. Além disso, os municípios com falta de informações também favorecem para o aumento nos casos de bTB.

Fica claro, na Figura 4.8(a) que, os municípios da região oeste do estado tiveram as maiores contagens de animais com resultado positivo. Com destaque para os municípios de Patos de Minas, Lagoa Formosa e Coromandel na região do Alto Paranaíba, que estiveram em 3º, 4º e 5º lugar, respectivamente, em relação a contagem de animais com diagnóstico positivo. O município de Campo Florido na região do Triângulo Mineiro apresentou o

maior número de animais com resultado positivo para a bTB, seguido pelo município de Unaí na região Noroeste que apresentou a segunda maior contagem.

Figura 4.8 – Visualização geográfica para o número de animais com diagnóstico positivo para a bTB entre os anos 2014-2018 Figura 4.8(a) e a diferença entre o número de animais abatidos e o número de animais positivos Figura 4.8(b).



Fonte: Elaboração própria.

Ainda na Figura 4.8(b), pode-se observar que os municípios da região oeste têm maiores tendências a terem valores subnotificados para os dados de focos, sendo estes os municípios que concentram maior diferença entre o número de animais abatidos com lesões sugestivas e o número de animais que testaram positivo para a bTB. Entre os municípios com maiores discrepâncias podemos dar maior destaque para Uberaba, Araxá e Serra do Salitre, destacados na cor azul. Já os municípios de Cajuri, Boa Esperança, Bandeira do Sul, Cruzeiro da Fortaleza, Rio Vermelho e São Gotardo não apresentaram diferenças entre esses valores.

4.4.2 Ajuste dos modelos

Nesta etapa, foram ajustados modelos GAMLSS tendo como resposta a variável número de animais com diagnóstico positivo para a bTB (CFTA), considerando as variáveis efetivo de rebanho bovino (ER) e número de vacas ordenhadas (VO) como explanatórias. A escolha da distribuição para a variável resposta foi feita através da função `chooseDist` do pacote `gamlss` (STASINOPOULOS; RIGBY, 2007), considerando uma resposta de contagem e o efeito das duas variáveis explanatórias. Neste sentido, foram utilizadas 8

distribuições candidatas de acordo com os melhores valores do GAIC. A Tabela 4.3 mostra os valores do critério GAIC obtidos para cada distribuição testada.

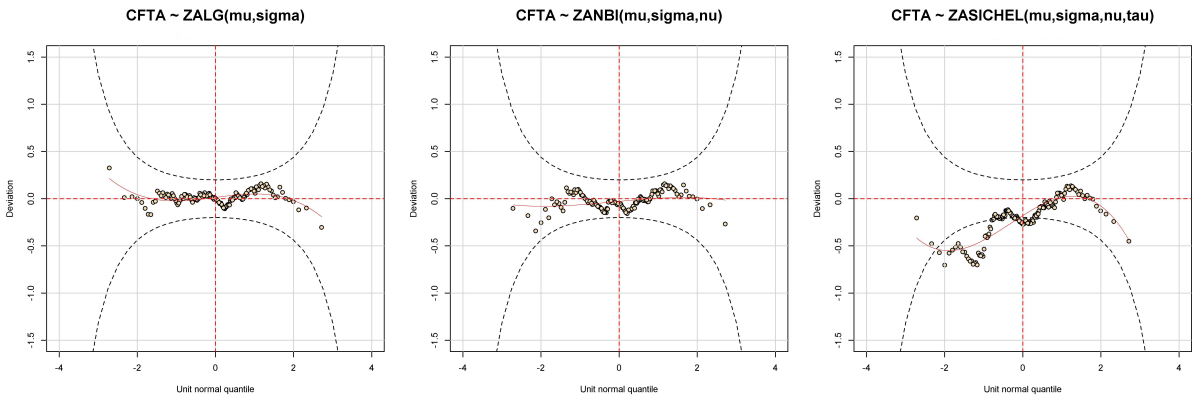
Tabela 4.3 – Possíveis modelos e valores do Critério de Informação de Akaike Generalizado para o número de animais com diagnóstico positivo para a bTB no estado de Minas Gerais.

Distribuições	GAIC
PIG	974,32
ZALG	**968,64**
ZIPIG	976,91
ZANBI	**970,66**
ZAPIG	984,82
ZASICHEL	**969,34**
GPO	976,60
BNB	980,47

Fonte: Elaboração própria.

Observando a Tabela 4.3, ver-se que as distribuições *Zero Adjusted Logarithmic* (ZALG), *Zero Altered Negative Binomial Type I* (ZANBI) e *Zero Adjusted Sichel* (ZASICHEL) apresentaram os menores valores para o Critério de Informação de Akaike Generalizado, sendo estas, as distribuições mais apropriadas para modelagem da variável resposta, dado o efeito das variáveis explanatórias. Para cada modelo presente na Tabela 4.3 foram construídos gráficos para avaliação dos resíduos, no intuito de identificar possíveis inadequações. A Figura 4.9 apresenta os resultados obtidos para os três melhores (ZALG, ZANBI e ZASICHEL).

Figura 4.9 – *Worm plot* para os resíduos dos modelos ZALG, ZANBI e ZASICHEL para a modelagem do número de bovinos com diagnóstico positivo para a bTB no estado de Minas Gerais.



Fonte: Elaboração própria.

É possível observar que o modelo ZASICHEL parece não ser adequado para modelagem deste conjunto de dados. Já os modelos ZALG e ZANBI apresentaram ajustes satisfatórios. Como estes modelos não apresentaram diferença significativa nos valores do GAIC, optou-se neste trabalho pelo modelo ZANBI, por critério de interpretabilidade, uma vez que o parâmetro de locação deste modelo representa a média da distribuição antes do modelo ser truncado no zero.

O modelo Binomial Negativo Alterado no Zero (ZANBI)

Rigby et al. (2019) definem a distribuição Binomial Negativa Alterada no Zero como uma distribuição discreta mista composta por duas componentes: o valor zero ($Y = 0$), com probabilidade ν e uma distribuição Binomial Negativa truncada no zero ($Y = Y_0$), denotada por $Y_0 \sim NBItr(\mu,\sigma)$ com probabilidade $1-\nu$. Deste modo, Y tem distribuição ZANBI, denotada por $Y \sim ZANBI(\mu,\sigma,\nu)$ com a seguinte função densidade de probabilidade

$$P(Y = y|\mu,\sigma,\nu) = \begin{cases} \nu, & \text{se} \quad y = 0 \\ cP(Y_1 = y|\mu,\sigma), & \text{se} \quad y = 1,2,3,\ldots \end{cases}$$

para $y = 0,1,2,3,\ldots$, em que $\mu > 0$, $\sigma > 0$, $0 < \nu < 1$ e $c = (1-\nu)/(1-p_0)$, em que $p_0 = P(Y_1 = 0|\mu,\sigma) = (1+\sigma\mu)^{-1/\sigma}$, com $Y_1 \sim NBI(\mu,\sigma)$. Portanto

$$P(Y_1 = y|\mu,\sigma) = \frac{\Gamma\left(y+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}},$$

para $y = 0,1,2,3,\ldots$

O parâmetro μ da distribuição representa a média da componente binomial negativa antes de ser truncada no zero. O parâmetro σ representa a dispersão da distribuição antes de ser truncada no zero e o parâmetro ν é a probabilidade exata de se observar um valor nulo, ou seja, $Y = 0$. A média e a variância da distribuição são obtidas respectivamente por

$$E(Y) = c\mu \quad \text{e} \quad Var(Y) = c\mu + c\mu^2(1+\sigma-c).$$

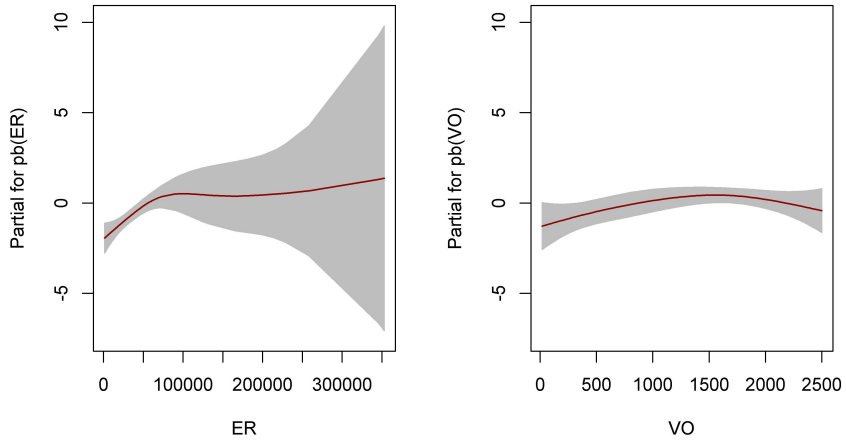
A escolha das variáveis foi realizada por meio da seleção *stepwise* para cada parâmetro da distribuição, considerando os valores do GAIC. Na expressão do modelo (Equação (4.1)) é possível observar quais variáveis foram incluídas nos preditores do modelo sele-

cionado. Observa-se que, para o efeito do parâmetro de locação, ambas as variáveis (ER e VO) se mostram significativas no modelo. Já para os parâmetros σ e ν, nenhuma das variáveis explanatórias apresentou efeito significativo.

$$\begin{aligned} \boldsymbol{CFTA} &\overset{\text{ind}}{\sim} ZANBI(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}) \\ \boldsymbol{\eta}_1 &= \log(\hat{\boldsymbol{\mu}}) = \beta_{11} + h_{21}ER + h_{31}VO \\ \boldsymbol{\eta}_2 &= \log(\hat{\boldsymbol{\sigma}}) = \beta_{12} \\ \boldsymbol{\eta}_3 &= \text{logit}(\hat{\boldsymbol{\nu}}) = \beta_{13} \end{aligned} \tag{4.1}$$

Termos de suavização não linear foram encontrados para a relação entre as variáveis explanatórias e o preditor do parâmetro μ, tal como apresentado na Figura 4.10. Para isto, foi utilizada a função *P-Spline*. Pode-se perceber que para um efetivo bovino com até 100 mil cabeças há um aumento no número de animais com diagnóstico positivo para a bTB e para valores acima de 100 mil o número de animais positivos se mantém aproximadamente constante. Observa-se também um leve aumento no número de animais positivos para um número de vacas ordenhadas até 1500, para valores maiores há um leve decaimento no número de animais positivos. As estimativas dos parâmetros do modelo podem ser consultadas na Tabela 4.4.

Figura 4.10 – Termos de suavização para as variáveis explanatórias em relação ao preditor de μ.



Fonte: Elaboração própria.

De acordo com os dados apresentados na Tabela 4.4 é possível observar que o parâmetro de dispersão σ não é afetado pelo efeito das covariáveis, assim como o parâmetro ν que representa a probabilidade de uma informação ser nula, ou seja, a probabilidade de

não haver animais positivos nos municípios. Essa probabilidade foi estimada em 0,1908 ($e^{\hat{\beta}_{13}}/1+e^{\hat{\beta}_{13}}$).

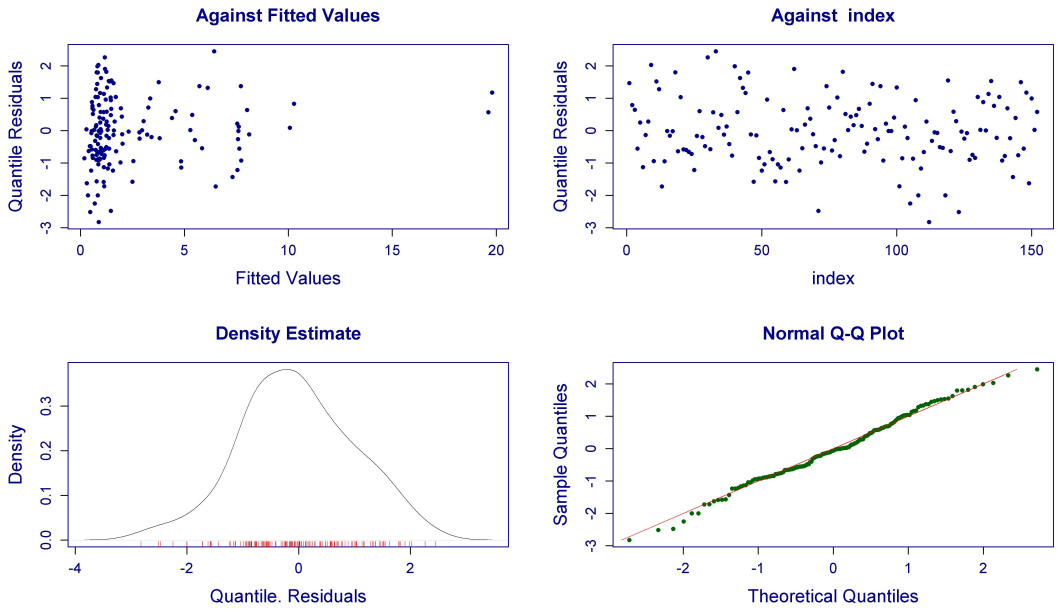
Tabela 4.4 – Estimativas dos parâmetros para o modelo ZANBI para o número de animais com diagnóstico positivo para a bTB.

Coeficiente	**Estimativa**	**Erro padrão**	**Estatística t**	**Valor p**
μ com ligação log				
Intercepto	$-2,624$	3,479	-0.754	0,452
pb(ER)	—	—	—	—
pb(VO)	—	—	—	—
σ com ligação log				
Intercepto	3,425	3,683	0,930	0,354
ν com ligação logit				
Intercepto	$-1,445$	0,206	-6.999	$<$ 0,001

Fonte: Elaboração própria.

Na Figura 4.11 pode-se observar que o modelo gerou baixos valores para os resíduos, e que estes não parecem apresentar autocorrelação. Também fica claro o comportamento normal dos resíduos o que indica uma boa adequação do modelo ajustado em concordância à Figura 4.9.

Figura 4.11 – Análise dos resíduos para o modelo ZANBI.



Fonte: Elaboração própria.

4.5 Discussão

Neste estudo, pode-se identificar que os municípios da região oeste do estado tiveram as maiores contagens para o número de bovinos com resultado positivo para a bTB, em concordância com o trabalho de Belchior et al. (2016) que apontaram o Alto Paranaíba como a região de maior prevalência da Tuberculose bovina no estado de Minas Gerais. Em contrapartida, os autores identificaram que a região com menor prevalência corresponde ao Triângulo mineiro. Todavia, foi identificado altas contagens de animais diagnosticados como positivos para a bTB em municípios dessa região, com destaque para o município de Campo Florido que apresentou o maior número de animais positivos, fato que se deve ao maior número de testes tuberculínicos realizados em uma propriedade exportadora de bovinos vivos, conforme discutido durante a análise explanatória.

Constatou-se um aumento na contagem de animais com resultado positivo para a bTB nos municípios do estado de Minas em localidades com maior número de vacas ordenhadas. Corroborando com vários outros estudos, como o trabalho de Barbieri et al. (2016) no estado de Minas Gerais, que mostraram uma relação entre o aumento na chance de se observar animais positivos em rebanhos com maior número de vacas. Belchior et al. (2016) também identificaram alta prevalência da Tuberculose bovina em uma região leiteira do estado de Minas. Souza Filho (2019) também afirma que a bTB é mais prevalente em fazendas com alta produção de leite. O mesmo foi encontrado por Perez et al. (2002) na Argentina, que identificaram agrupamentos de maiores prevalências da bTB em regiões com alta produção de leite.

Martínez et al. (2007) salientam que o gado leiteiro é o principal fator de risco para persistência e disseminação da bTB, e que essa população de bovinos deve ser o foco principal da atenção para eliminar a doença. Apesar disto, os autores deixam claro que, informações adicionais como tamanho de rebanho, densidade populacional, dinâmica da população, entre outros fatores poderiam contribuir para a precisão das previsões realizadas sobre a doença.

Neste trabalho, o efetivo bovino também se mostrou como fator influenciador para o aumento no número de animais com diagnostico positivo para a bTB em Minas Gerias. Visto que, foi observado maior número de animais positivos em regiões com maior efetivo de rebanho. Milne et al. (2019) também verificaram que a população de bovinos foi considerada um importante fator para aumento dos casos de bTB na Irlanda no Norte. Bastida

et al. (2017) também identificaram agrupamentos no número de rebanhos positivos para bTB no Estado do México entre os anos 2005-2010 localizados em regiões com grande número de animais.

Uma das limitações deste trabalho está relacionado à dificuldade de coleta de informações que poderiam ser importantes para explicar o comportamento da bTB na região de estudo. Vale também ressaltar que a falta de informações a acerca da doença e os possíveis casos de subnotificação do número de animais positivos em muitos municípios do estado de Minas Gerais devido ao fato da não obrigatoriedade de realização dos testes de diagnóstico, dificulta o estudo e o entendimento do comportamento da doença, tornando ainda mais difícil o controle e erradicação da mesma.

A análise de autocorrelação espacial é também prejudicada neste trabalho, no sentido de haver muitos valores ausentes na base de dados, devido as características dos dados já discutidas aqui. Neste sentido, os municípios que não apresentaram informações foram desconsiderados e não foram levados em consideração na estrutura de vizinhança. Portanto, um estudo com uma base de dados que apresente maior completude seria necessário para avaliar a dependência espacial dos casos de bTB nessa região.

4.6 Conclusões

A falta de informação sobre os casos de bovinos com Tuberculose bovina nos municípios do estado de Minas Gerais torna difícil uma avaliação assertiva acerca da real situação do estado em relação a ocorrência da doença. Essa situação se agrava ainda mais, diante dos possíveis casos de subnotificação destacados no presente estudo. Esses dados tem sido usados para direcionar políticas de controle da doença, mas é preciso olhar com cautela para eles. O ideal seria aplicar alguma técnica para corrigir o viés de seleção desses dados ou realizar estudos com amostragem probabilística.

A região oeste do estado concentrou os municípios com maiores contagens para o número de animais com diagnóstico positivo para a Tuberculose bovina. Na região do Alto do Paranaíba destacaram-se os municípios de Patos de Minas, Lagoa Formosa e Coromandel. Campo Florido foi o município mineiro com maior contagem, localizado na região do Triângulo Mineiro. Na região Noroeste, o município de Unaí apresentou a segunda maior contagem do estado.

Os modelos de regressão da família GAMLSS mostraram-se bastante eficazes na modelagem desse conjunto de dados, contornando o problema de assimetria, curtose elevada e superdispersão, gerando estimativas mais precisas em relação ao fenômeno. A partir do modelo ZANBI, o qual apresentou melhor adequabilidade, foi possível identificar que, conforme há um aumento no número de bovinos no rebanho, há também um aumento no número de animais positivos. De forma similar, essa contagem também aumenta quando há uma aumento no número de vacas ordenhadas no rebanho. O modelo estimou uma baixa probabilidade de se observar municípios que não apresentaram diagnóstico positivo para a bTB.

5 APLICAÇÃO 2: ESTUDO DOS CASOS NOTIFICADOS DE DENGUE NO ESTADO DA PARAÍBA ENTRE OS ANOS 2008-2018

Resumo

A dengue é uma doença infecciosa causada por 4 sorotipos distintos pertencentes ao gênero *flavivirus* e a família *flaviviridae*. A enfermidade é transmitida principalmente pela picada do mosquito fêmea do gênero *Aedes Aegypti*. A doença ainda representa um sério problema de saúde pública que assola o mundo todo e que tem se agravado ainda mais nos últimos anos, causando prejuízos na saúde e na economia dos países. O objetivo deste trabalho foi estudar o comportamento da dengue no estado da Paraíba, identificando possíveis fatores que potencializam o número de casos notificados nos municípios do estado. Para tanto, as ferramentas da estatística espacial voltadas ao estudo de dados de área aliadas ao ajuste de modelos de regressão podem auxiliar no estudo do comportamento da doença, bem como avaliar possíveis fatores relacionados ao aumento dos casos de dengue no estado da Paraíba. O estudo revelou que a taxa de incidência de dengue no estado da Paraíba não ocorre de maneira aleatória, sendo que na região oeste do estado, mais especificamente nas mesorregiões do Sertão Paraibano e Borborema, foi identificado um agrupamento de municípios com altas taxas de incidência por dengue rodeados de municípios vizinhos que apresentam a mesma situação. Princesa Isabel, Monte Horebe, Monteiro e Zabelê foram os municípios com as maiores taxas de incidência por dengue. Foi possível constatar que a ocorrência da doença está relacionada à combinação de uma série de fatores ligados ao contexto socioeconômico e ambiental da região, que podem favorecer para o aumento dos casos notificados de dengue no estado.

Palavras-chave: *Aedes Aegypti*. Autocorrelação espacial. Índice de Moran. Modelos GAMLSS.

5.1 Introdução

A dengue é uma doença infecciosa de etiologia viral com importante impacto na saúde mundial. Segundo Silva et al. (2020) há quatro tipos distintos de sorotipos que causam a enfermidade, (DENV1, DENV2, DENV3 e DENV4) pertencentes ao gênero *flavivirus* e a família *flaviviridae*. A forma de transmissão da doença ocorre por meio da picada do mosquito fêmea do gênero *Aedes* contaminado com um dos sorotipos, sendo o *Aedes Aegypti* o vetor primário no Brasil, podendo ser encontrado em regiões tropicais e subtropicais do mundo.

A doença pode apresentar duas formas clínicas: a dengue clássica ou febre da dengue (FD), que não apresenta maiores riscos a saúde do indivíduo e a febre hemorrágica da dengue (FHD), que pode levar o indivíduo a óbito. De acordo com Bhatt et al. (2013), a infecção pelo vírus da dengue em humanos pode não apresentar sintomas claros como

também pode levar a uma série de manifestações clínicas, podendo evoluir para um caso fatal. A Organização Pan Americana de Saúde (OPAS/OMS)[1] afirma que a imunidade vitalícia desenvolvida após a infecção do vírus da dengue é válida apenas para o tipo específico contraído, e que seguidas infecções por outros sorotipos aumentam as chances de evolução para o quadro mais grave da doença.

A dengue é ainda, um grande problema de saúde pública em escala mundial. O número de casos positivos da doença vem crescendo ao longo dos anos, segundo dados da Organização Mundial da Saúde (OMS)[2]. Este fato, deve-se não só a fatores ambientais como também ao crescimento desordenado da população e falta de políticas públicas adequadas para o combate a doença.

Para Barbosa e Silva (2015), fatores climáticos aliados à desestruturação urbana são fundamentais para a proliferação do *Aedes Aegypti*, sendo a precipitação pluviométrica um importante fator influenciador do surgimento de criadouros do vetor. Além de altos níveis de precipitação, temperaturas elevadas a altos valores para umidade relativa do ar também estão relacionadas com aumento no registro da incidência de dengue, como visto nos estudos (HONÓRIO; OLIVEIRA, 2001; RIBEIRO et al., 2006; CAO et al., 2017; PHANITCHAT et al., 2019).

Para além das variáveis climáticas, muitos estudos associam algumas variáveis socioeconômicas com a taxa de incidência de dengue. Messina et al. (2019) consideraram variáveis climáticas e socioeconômicas em seu modelo para realizar previsões da distribuição global da dengue para anos 2020, 2050 e 2080. Neste estudo, os autores identificaram que as variáveis precipitação, temperatura, umidade e PIB se mostraram importantes para descrever o comportamento de transmissão da doença.

Vários outros trabalhos buscaram identificar a influência de variáveis socioeconômicas na ocorrência da dengue. Dentre eles, Almeida, Medronho e Valencia (2009) observaram uma relação entre a taxa de incidência de dengue e as variáveis: percentual de domicílios ligados à rede geral de esgoto, densidade populacional e percentual de domicílios com lavadoura de roupas no município do Rio de Janeiro. Gomes, Bastos e Nascimento (2017) verificaram influência do PIB per capita, densidade populacional e saneamento básico na taxa de incidência de dengue no estado de Minas Gerais. LIMA et al. (2017) também identificaram relação entre densidade populacional e casos positivos de dengue

[1] https://www.paho.org/pt/topicos/dengue
[2] https://www.who.int/health-topics/dengue-and-severe-dengue#tab=tab_1

no interior de São Paulo. Ainda nesse estudo, os autores destacaram uma redução no número de casos em virtude do aumento da renda bruta per capita.

Poucos trabalhos são encontrados na literatura para o estado da Paraíba avaliando a propagação da dengue utilizando ferramentas da estatística espacial aliadas ao ajuste de modelos de regressão. Neste seguimento, Silva et al. (2020) identificaram uma dependência espacial na taxa de dengue no estado entre os anos 2007-2016. Os autores ajustaram modelos de regressão espaciais em comparação com o modelo linear normal e verificaram melhor adequação. Os modelos identificaram influência de variáveis socioeconômicas e climáticas na taxa de dengue do estado.

Os modelos da família GAMLSS são ainda pouco explorados em vários seguimentos da ciência. Contudo, estes modelos trazem uma ferramental estatístico muito poderoso e que pode ser bastante útil na descrição do comportamento da dengue no estado da Paraíba, dado que apresentam um número considerável de distribuições probabilísticas que podem ser usadas para modelar a distribuição da doença. Além disso, a relação das covariáveis com a variável resposta é avaliada não só no contexto linear, mas também por meio de suavizadores. Além disso, o modelo não se limita na modelagem do parâmetro de locação da distribuição, estendendo-se aos demais parâmetros (como de escala e forma).

Diante disto, o presente estudo se mostra de grande valia para verificar o comportamento da dengue no estado da Paraíba por meio dos modelos da família GAMLSS, os quais são bastante flexíveis e capazes de incorporar as características da distribuição da Taxa de Incidência de Dengue. Além disso, tais modelos podem considerar a estrutura de dependência espacial presente nas observações, o que pode trazer melhorias nas estimavas dos parâmetros, fornecendo estimativas mais adequadas.

5.2 Material e Métodos

Esta sessão destina-se à exposição da metodologia utilizada para execução do estudo voltado aos dados dengue. Neste sentido, a área de estudo é apresentada, bem como descrição da base de dados e as análises estatísticas empregadas.

5.2.1 Área de Estudo

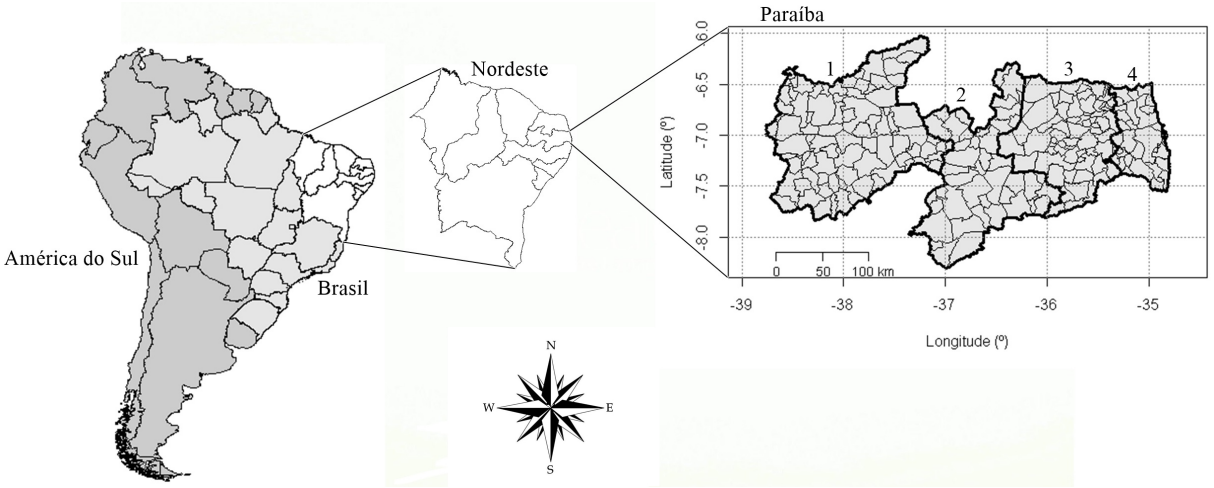
A Paraíba é uma das 27 unidades federativas do Brasil. Localiza-se no leste da região Nordeste, em conjunto com os estados de Alagoas, Bahia, Ceará, Maranhão, Piauí,

Pernambuco, Rio Grande do Norte e Sergipe. Seu território encontra-se entre os paralelos 6º02'12"e 8º19'18"S, e entre os meridianos de 34º45'54"e 38º45'45"W.

A Paraíba é o 15º estado mais populoso do Brasil, de acordo com estimativas do IBGE em 2020. O estado limita-se com o Rio Grande do Norte ao norte, Pernambuco ao sul, Ceará à oeste e o Oceano Atlântico à leste. Seu território é dividido em 223 municípios e apresenta uma área territorial de 56.467,242 km^2. No último censo realizado pelo IBGE tinha uma população de 3.766.528 habitantes, que em 2020 foi estimada em 4.039.277 habitantes, com uma densidade populacional de 66,70 hab/km^2.

O IBGE divide o estado em quatro mesorregiões, como visto na Figura 5.1. A saber: 1 - Sertão Paraibano, 2 - Borborema, 3 - Agreste Paraibano e 4 - Mata Paraibana. A Paraíba apresenta um clima quente com temperaturas elevadas que variam de acordo com relevo local.

Figura 5.1 – Localização geográfica do estado da Paraíba, mesorregiões e seus respectivos municípios.



Fonte: Elaboração própria.

5.2.2 A Base de Dados

Para o estudo do comportamento da dengue no estado da Paraíba foram utilizados dados de casos notificados de dengue entre os anos 2008-2018, provenientes do Sistema de Informação de Agravos de Notificação (Sinan), disponibilizados pela Gerência Executiva de Vigilância em Saúde da Secretaria de Estado da Saúde da Paraíba. Posteriormente foi calculada a Taxa de Incidência de Dengue por 100.000 habitantes, que procedeu com a

soma dos casos notificados entre os anos nos 223 municípios do estado, dividido pelo total populacional dos respectivos municípios e multiplicado por 100.000.

Como independentes foram utilizadas as variáveis Índice de Desenvolvimento Humano Municipal (IDHM), Índice de Gini (IG), comumente utilizado para medir a desigualdade econômica de uma região, precipitação média anual em milímetros (PRECIP), temperatura média anual em graus Celsius (TEMP), proporção de domicílios com mais de oito residentes (PDR8), proporção de domicílios sem renda mensal (PDSREND), proporção de domicílios com eletricidade desconhecida (PDEOD), número de domicílios em que o lixo é jogado em terreno baldio ou logradouro público (JTB); número de domicílios com coleta de lixo (CL); número de domicílios ligados à rede geral de esgoto ou com fossa séptica (RGE); proporção de domicílios com saneamento básico (PSB); número de domicílios particulares ligados à rede geral de água (AA); população alfabetizada (PA) e taxa de desemprego (TD).

As variáveis referentes às condições socioeconômicas são provenientes do censo demográfico de 2010 realizado pelo IBGE. Já entre as variáveis climatológicas, a variável precipitação foi obtida no site da Agência Executiva de Gestão das Águas do estado da Paraíba (AESA) e a variável temperatura foi obtida no site CLIMATE[3].

5.2.3 Análises Estatísticas

Para estes dados foi realizada inicialmente uma análise exploratória, como cálculo de medidas descritivas para os casos notificados de dengue nos anos em estudo e para os municípios do estado. Também foram construídos mapas coropléticos para observar o comportamento dos casos de dengue no estado da Paraíba.

A distribuição da Taxa de Incidência de Dengue foi avaliada. Foram utilizados o Índice I de Moran Global e a estatística c de Geary para verificar a existência de autocorrelação espacial para a Taxa de Incidência de Dengue. O Índice de Moran Local também foi utilizado para identificar áreas com valores similares e regiões em que a autocorrelação espacial é significativa. Posteriormente foram ajustados os modelos da família GAMLSS com a introdução da estrutura espacial para verificar a possível influência das covariáveis na Taxa de Incidência de Dengue no estado.

[3] www.climate-data.org

As análises foram realizadas no *software* R (R Core Team, 2019) na versão 3.6.1. Para tanto, foram utilizados os pacotes `spdep` (BIVAND; WONG, 2018) e `spatialreg` (BIVAND; PIRAS, 2015) para o auxílio nas análises espaciais. `gamlss` (STASINOPOULOS; RIGBY, 2007), `gamlss.dist` (RIGBY et al., 2019) e (STASINOPOULOS; RIGBY, 2020) e `gamlss.spatial` (DE BASTIANI; STASINOPOULOS, 2015) para o ajuste dos modelos GAMLSS com a inclusão do efeito espacial.

5.3 Resultados

Nesta sessão serão apresentados os resultados encontrados nas análises dos dados comentados na sessão 5.2.2, referentes a ocorrência de dengue no estado da Paraíba. Primeiramente é apresentada a análise explanatória dos dados, seguida pela análise de autocorrelação espacial e por fim o ajuste dos modelos de regressão.

5.3.1 Análise Exploratória

Entre os anos de 2008-2018, o estado da Paraíba apresentou um total de 167.462 casos notificados por dengue, sendo o ano de 2016, o que apresentou o maior número de casos notificados, com 44.527 casos registrados, seguido pelos anos 2015 (30.174), 2013 (18.502) e 2011 (16.314). Na Tabela 5.1 é possível observar que em todos os anos houve ao menos um município que não apresentou casos notificados. O ano de 2009 foi o que apresentou menor número de casos de dengue nos municípios do estado, com uma média de 7,25 casos, máximo de 269 e um total de 1.616 casos. Em todos os anos, João Pessoa foi o município paraibano com maior número de casos notificados por dengue, com exceção para o ano de 2008, no qual o município de Patos se sobressaiu dos demais.

Na Figura 5.2 é possível observar em quais municípios houve presença de casos notificados por dengue no estado da Paraíba, conforme ano de notificação. Pode-se observar que 2009 e 2017 foram os anos com menor número de municípios com notificação da série, em contraste com os anos de 2015 e 2016, que apresentaram o maior números de municípios com ao menos um caso notificado em seu território.

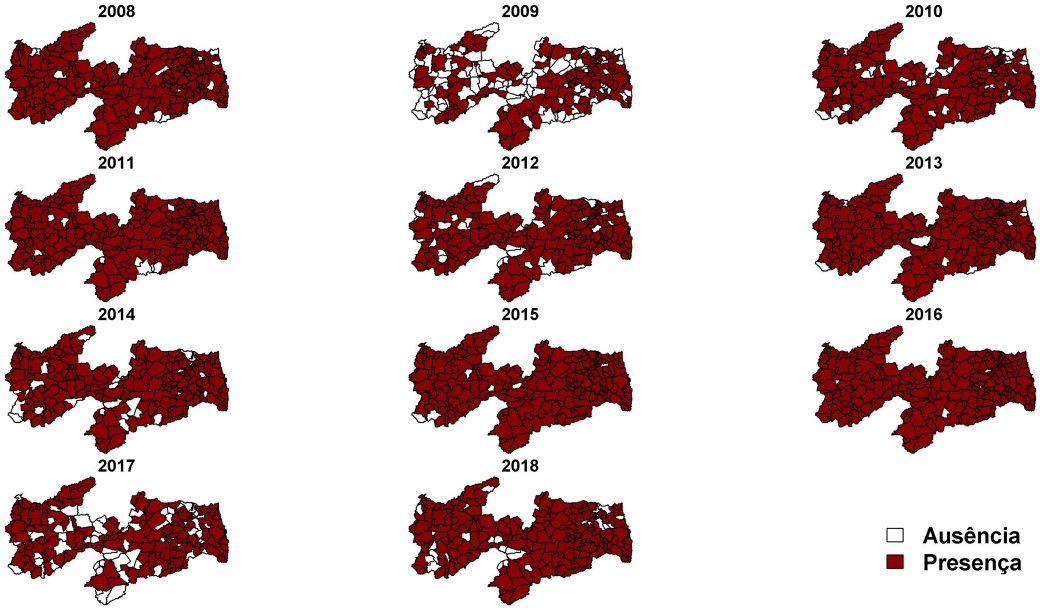
Também foi realizada uma análise descritiva para a taxa de incidência de dengue, calculada conforme comentado na subsessão 5.2.2. Neste sentido, na Tabela 5.2 apresentam-se algumas medidas descritivas para esta variável. Pode-se observar que a distribuição da taxa de incidência de dengue parece ser leptocúrtica com uma leve assimetria

Tabela 5.1 – Estatísticas descritivas para os casos notificados por dengue nos municípios do estado da Paraíba entre os anos 2008-2018.

Período	***min***	***max***	***média***	***desv***	***Total***
2008	0	1033	50,57	120,08	11277
2009	0	269	7,25	24,12	1616
2010	0	1366	39,72	122,96	8858
2011	0	4548	73,16	322,35	16314
2012	0	4636	51,80	328,60	11552
2013	0	3336	82,97	275,24	18502
2014	0	2395	34,29	167,92	7648
2015	0	4907	135,31	436,06	30174
2016	0	6982	199,67	569,33	44527
2017	0	2523	21,04	170,19	4692
2018	0	2860	55,16	226,12	12302
2008-2018	6	34740	750,95	2510,75	167462

Fonte: Elaboração própria.

Figura 5.2 – Municípios com ausência e presença de casos notificados por dengue no estado da Paraíba, segundo ano de notificação.



Fonte: Elaboração própria.

positiva. Entre o período de estudo, houve uma incidência média de aproximadamente 4.628 casos notificados a cada 100 mil habitantes. Ainda na Tabela 5.2, observa-se que 50% dos municípios apresentaram uma taxa de incidência abaixo de 3.530 casos/100 mil habitantes e 25% destes tiveram uma taxa inferior à 1.714 casos/100 mil habitantes.

Princesa Isabel, localizado na região do Sertão Paraibano, foi o município com maior taxa de incidência (19.698 casos/100 mil hab.), seguido por Monte Horebe com 19.211 casos/100 mil hab, também no Sertão do estado, Monteiro com 18.372 casos/100

mil hab. e Zabelê com 16.404 casos/100 mil hab, localizados na região da Borborema. Enquanto que Marcação, na região da Mata Paraibana, foi o município com menor taxa de incidência (213 casos/100 mil hab.).

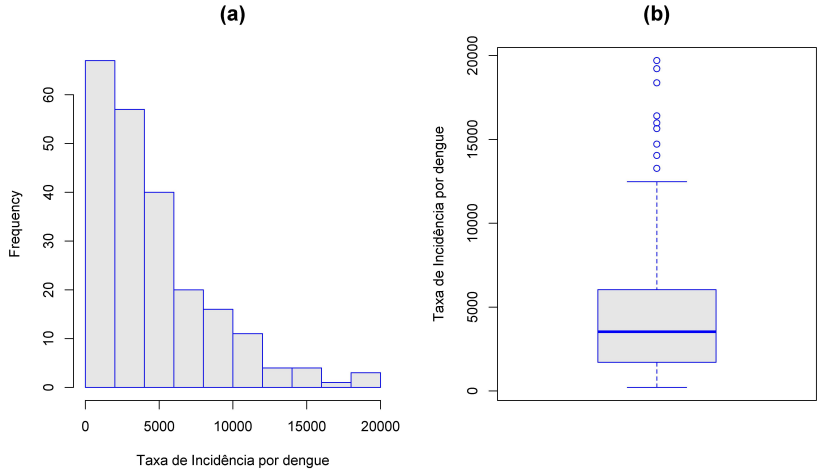
Tabela 5.2 – Estatísticas descritivas para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018.

Estatística	**Taxa de Incidência**
min	212,77
max	19698,47
1º quartil	1713,48
3º quartil	6040,13
mediana	3529,61
média	4627,94
desvio	3868,45
Simetria	1,45
curtose	2,09

Fonte: Elaboração própria.

Na Figura 5.3(a) fica claro o padrão assimétrico da distribuição da taxa de incidência de dengue nos municípios do estado da Paraíba, em concordância com os dados apresentados na Tabela 5.2. Um número razoável de observações atípicas pode ser observado na Figura 5.3(b), sendo que essas obervações apresentaram valores bem acima da média para a taxa de incidência de dengue.

Figura 5.3 – Visualização gráfica da distribuição da taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018.

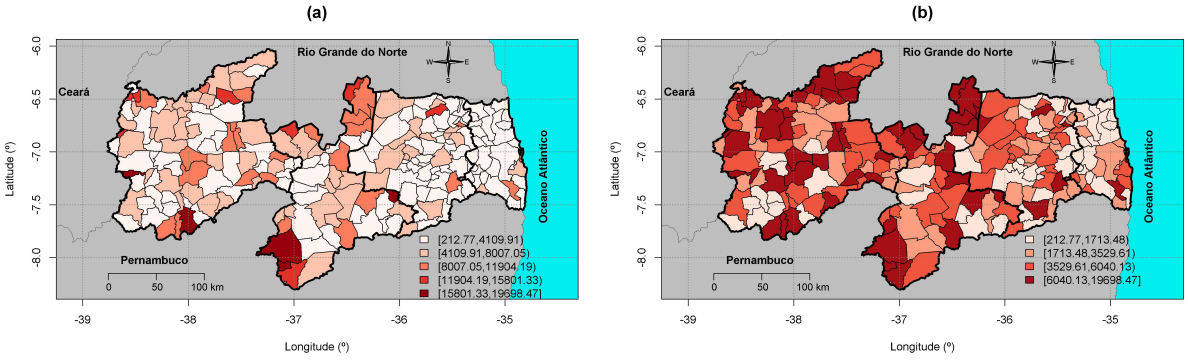


Fonte: Elaboração própria.

Para visualizar o comportamento da taxa de incidência de dengue nos municípios do estado foram construídos os mapas coropléticos apresentados na Figura 5.4. Pode-

se observar que a região da Mata Paraibana concentrou os municípios com as menores taxas de incidência. Em contrapartida, as regiões do Sertão do estado e da Borberema apresentaram os municípios com altas taxas de incidência, com destaque para os municípios de Princesa Isabel e Monte Horebe (Sertão), Monteiro, Zabelê e Caturité (Borborema), representados pelo tom vermelho mais escuro na Figura 5.4(a). Na Figura 5.4(b), pode-se observar que 75% dos municípios apresentam taxa de incidência entre [213; 6.040) casos/100 mil habitantes, representados pelos tons mais claros de vermelho. Os 25% restantes apresentam as maiores taxas de incidência, representados pelo tom vermelho mais escuro.

Figura 5.4 – Mapa de intervalos iguais (Figura 5.4(a)) e mapa de quartis (Figura 5.4(b)) para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018.



Fonte: Elaboração própria.

5.3.2 Análise da Autocorrelação Espacial

A análise espacial sob o enfoque de dados de área é baseada na relação de vizinhança existente entre as áreas da região de estudo. Uma das maneiras de se avaliar essa relação é por meio do cálculo de medidas que captam autocorrelação espacial da variável na região de estudo. Neste sentido, foram realizados testes de significância para o índice I de Moran e a estatística c de Geary, ao nível de 5% de probabilidade, sob a hipótese nula H_0 de completa aleatoriedade dos valores da variável na região. Os resultados para os testes estão expostos na Tabela 5.3.

De acordo com os resultados apresentados na Tabela 5.3, pode-se concluir que há uma autocorrelação espacial positiva para a taxa de incidência de dengue nos municípios do estado da Paraíba, o que implica dizer que há similaridade entre os valores do atributo

Tabela 5.3 – Testes de significância para a autocorrelação espacial para a taxa de incidência de dengue por 100 mil habitantes nos municípios do estado da Paraíba entre os anos 2008-2018.

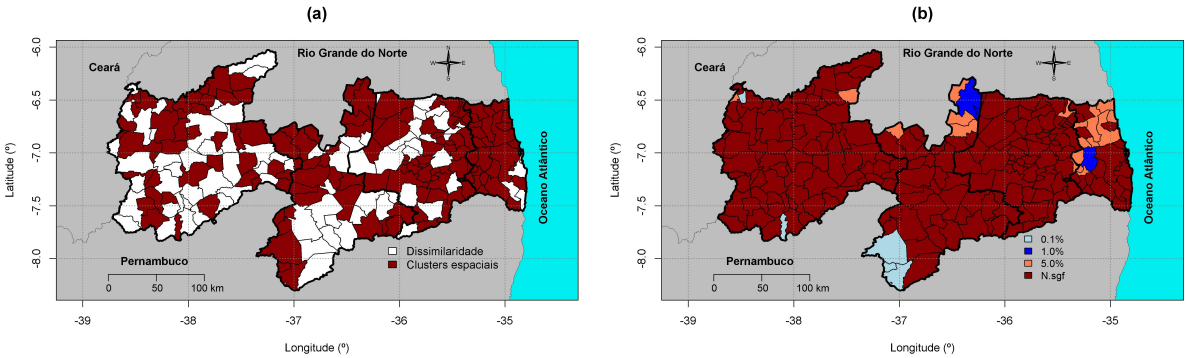
Estatística	**Estimativa**	**Valor p**
I de Moran	0,2184	$< 0{,}001$
c de Geary	0,7476	$< 0{,}001$

Fonte: Elaboração própria.

nos municípios, ou seja, municípios com altas taxas de incidência tendem a estar rodeados de municípios vizinhos que também apresentam altas taxas. O mesmo vale para os municípios que apresentam baixos valores para a taxa de incidência de dengue.

No intuito de identificar a presença de *clusters* espaciais entre os municípios do estado da Paraíba em relação à taxa de incidência de dengue, foram calculados os valores para o índice de Moran local referente a cada município. Estes valores foram dispostos no mapa do estado (Figura 5.5(a)) identificando em quais regiões houve agrupamento de municípios com valores similares e onde houve dissimilaridade dos valores do atributo. Vê-se que, em todo o estado há agrupamentos de municípios com taxas de incidência de dengue similares, com destaque para região leste que concentrou o maior número de municípios agrupados.

Figura 5.5 – Índice I_i de Moran local (Figura 5.5(a)) e LISA *Map* (Figura 5.5(b)) para os municípios do estado da Paíba em relação à taxa de incidência de dengue por 100 mil habitantes.



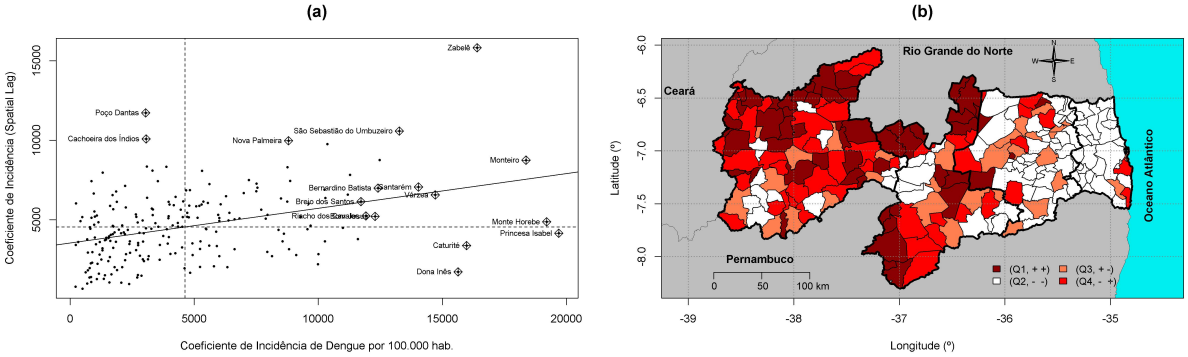
Fonte: Elaboração própria.

No LISA *Map* (Figura 5.5(b)) fica visível que a dependência entre os valores da taxa de incidência de dengue é ainda mais forte nos municípios de Monteiro, São Sebastião do Umbuzeiro e Zabelê na região da Borborema; Santarém e São José de Princesa na região do Sertão Paraibano (destacados na cor azul claro). A seguir na Figura 5.6(b) é possível

observar que estes municípios apresentaram altas taxas de incidência por dengue, cercados por vizinhos que também apresentaram altas taxas.

No diagrama de dispersão de Moran ou Moran *Scatterplot* (Figura 5.6(a)) pode-se perceber que, os municípios localizados no primeiro quadrante apresentam uma situação preocupante em relação à taxa de incidência de dengue no estado, pois além de apresentarem altas taxas, estão rodeados de municípios com mesmo cenário. Estes municípios requerem uma maior atenção do poder público para tomada de decisão de inclusão ou intensificação de medidas para redução dessas taxas.

Figura 5.6 – Moran *Scatterplot* e mapa de *clusters* espaciais para a taxa de incidência de dengue por 100 mil habitantes no estado da Paraíba (2008-2018).



Fonte: Elaboração própria.

Ainda na Figura 5.6(b) é possível observar que há uma claro agrupamento na região da Mata Paraibana de municípios com baixas taxas de incidência por dengue. Enquanto que, em partes das regiões do Sertão e da Borborema observa-se agrupamentos de municípios com altas taxas. Alguns destes municípios podem ser vistos na Figura 5.6(a) localizados no primeiro quadrante $(Q_1, ++)$, dos quais destacam-se os municípios de Monte Horebe, Monteiro e Zabelê com as maiores taxas de incidência neste quadrante.

5.3.3 Ajuste dos Modelos

Uma vez que a taxa de incidência de dengue nos municípios do estado da Paraíba apresentou autocorrelação espacial significativa, é necessário que as análises por intermédio de modelos de regressão considerem essa estrutura de dependência entre as observações. Por meio da função `chooseDist` foram avaliados os modelos plausíveis para o ajuste da distribuição condicional da taxa de incidência de dengue no estado da Paraíba, dado o efeito das covariáveis. Dentre eles, foram escolhidos oito modelos com base nos

melhores valores do GAIC e em critérios de convergência. Na Tabela 5.4 são expostos os resultados.

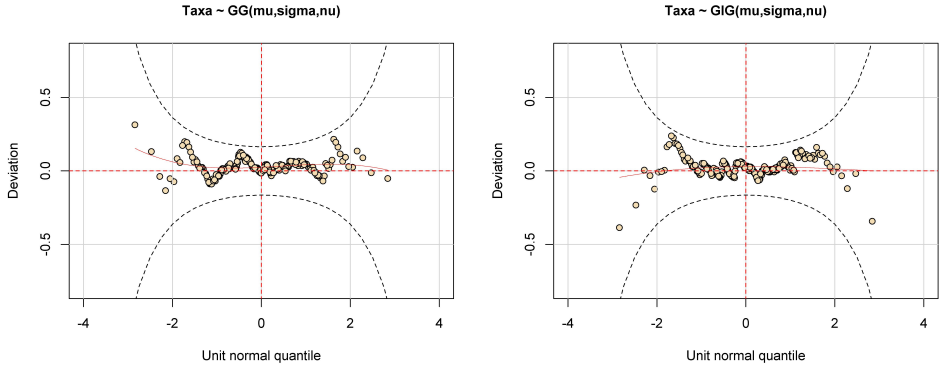
Tabela 5.4 – Possíveis modelos e valores do GAIC para a taxa de incidência de dengue no estado da Paraíba.

Distribuições	GAIC
EXP	4194,59
WEI	4157,59
LO	4154,29
GA	4153,42
GG	**4148,08**
IG	4163,52
GIG	**4139,99**
exGAUS	4157,31

Fonte: Elaboração própria.

Entre os modelos apresentados na Tabela 5.4, os modelos *Generalised Gamma* (GG) e *Generalised Inverse Gaussian* (GIG) parecem apresentar os melhores resultados, baseado no valor do critério de informação de Akaike generalizado. A seguir é exposto na Figura 5.7, o gráfico dos resíduos obtidos para estes modelos.

Figura 5.7 – *Worm plot* para os modelos Gama Generalizado e Inversa Gaussiana Generalizado.



Fonte: Elaboração própria.

Conforme pode ser visto na Figura 5.7, ambos os modelos, Gama Generalizado e Inversa Gaussiana Generalizado apresentam um ajuste satisfatório para a taxa de incidência de dengue no estado da Paraíba. Apesar do modelo GIG apresentar um valor inferior para o GAIC em relação ao modelo GG, optou-se por escolher o modelo GG, devido a um posterior problema de convergência no modelo GIG quando adicionado a componente espacial no modelo.

O modelo Gama Generalizado

A parametrização aqui apresentada para a distribuição Gama Generalizada é definida em Rigby et al. (2019) e foi usada por Lopatatzidis e Green (2000). A função densidade de probabilidade para uma variável aleatória Y com distribuição Gama Generalizada denotada por $Y \sim GG(\mu, \sigma, \nu)$ é então dada por

$$f_Y(y|\mu,\sigma,\nu) = \frac{|\nu|\theta^\theta z^\theta \exp\{-\theta z\}}{\Gamma(\theta) y}, \quad y > 0,$$

em que $z = (y/\mu)^\nu$, $\theta = 1/(\sigma^2\nu^2)$, $0 < \mu < \infty$, $0 < \sigma < \infty$ e $-\infty < \nu < \infty$, para $\nu \neq 0$. Perceba que $Z = (Y/\mu)^\nu$ tem distribuição gama com os seguintes parâmetros $Z \sim GA(1, \sigma\nu)$.

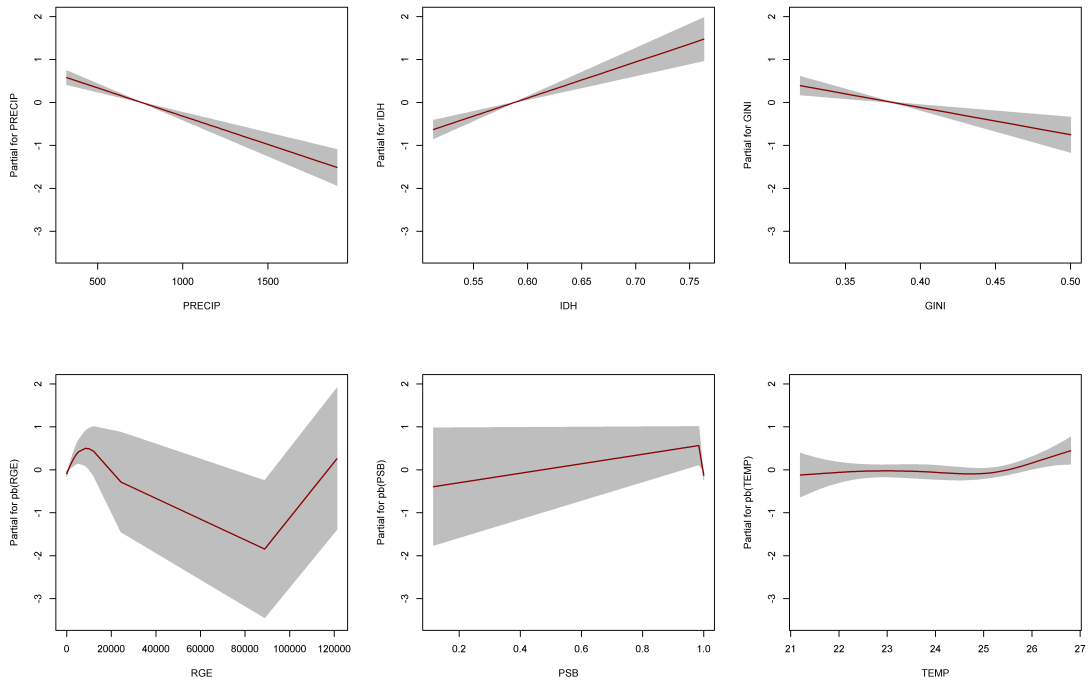
Após a escolha do modelo, foi enfim adicionada a componente espacial $s(Mun)$. Na Equação (5.1) pode-se verificar quais variáveis foram inclusas nos preditores dos parâmetros. Das variáveis investigadas, apenas precipitação, Índice de Desenvolvimento Humano Municipal, Índice de Gini, número de domicílios ligados à rede geral de esgoto, proporção de domicílios com saneamento básico e temperatura apresentaram relação com os preditores do modelo. Sendo que para o preditor do parâmetro μ, todas apresentaram efeito significativo. Para o parâmetro σ apenas precipitação e temperatura apresentaram efeito significativo. Enquanto que para o preditor do parâmetro ν apenas precipitação mostrou efeito significativo.

$$\begin{aligned}
\boldsymbol{Taxa} &\overset{\text{ind}}{\sim} GG(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}) \qquad (5.1)\\
\boldsymbol{\eta}_1 &= \log(\hat{\boldsymbol{\mu}}) = 4,351 - 0,001(PRECIP) + 8,397(IDH) - 6,339(GINI)\\
&\quad + h_{41}(RGE) + h_{51}(PSB) + h_{61}(TEMP) + s(Mun)\\
\boldsymbol{\eta}_2 &= \log(\hat{\boldsymbol{\sigma}}) = 1,963 - 0,105(TEMP) + h_{22}(PRECIP)\\
\boldsymbol{\eta}_3 &= \text{logit}(\hat{\boldsymbol{\nu}}) = 1,619 - 0,001(PRECIP)
\end{aligned}$$

Na Figura 5.8 pode-se ver a relação das variáveis explanatórias com o preditor do parâmetro μ. É possível perceber que há uma diminuição nos valores do parâmetro a medida que há um aumento nos valores de precipitação e do Índice de Gini. Em contrapartida, tem-se um aumento em μ para maiores valores do IDH. Observa-se também um aumento nos valores de μ para o número de domicílios com rede geral de esgoto até

1800 aproximadamente. Após isso, os valores decrescem até 9000, aumentando novamente após este valor. Já para a proporção de domicílios com saneamento básico, observa-se um aumento nos valores de μ em praticamente todo o domínio da variável, com um pequeno decréscimo quando essa proporção aproxima-se de 1. Enquanto que para a temperatura observa-se que os valores do parâmetro permanecem aproximadamente constante, com um aumento mínimo para temperaturas acima de 25ºC.

Figura 5.8 – Relação entre as variáveis explanatórias e o preditor do parâmetro μ.

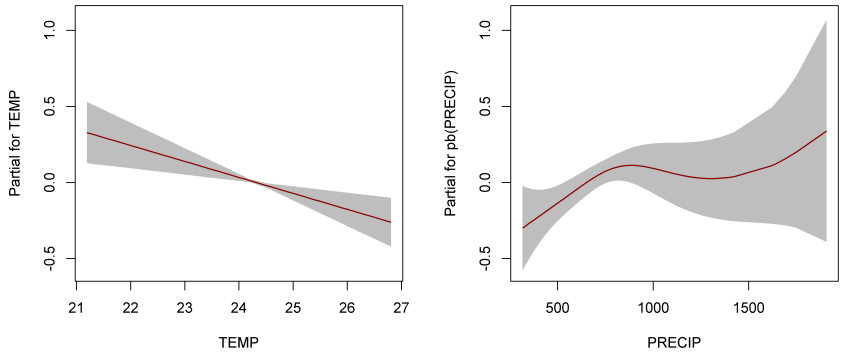


Fonte: Elaboração própria.

A seguir, a Figura 5.9 mostra a relação entre as variáveis explanatórias e o preditor do parâmetro σ, onde pode-se perceber que apenas a variável precipitação apresentou relação não linear com o preditor do parâmetro. Vê-se que os valores do parâmetro aumenta para uma precipitação média de até 800mm, em seguida decresce para valores entre 800mm e 1250mm e apresenta um crescimento novamente para valores acima de 1250mm. A variável temperatura apresenta relação linear com o preditor de σ, em que é possível observar menores valores do parâmetro em temperaturas mais elevadas.

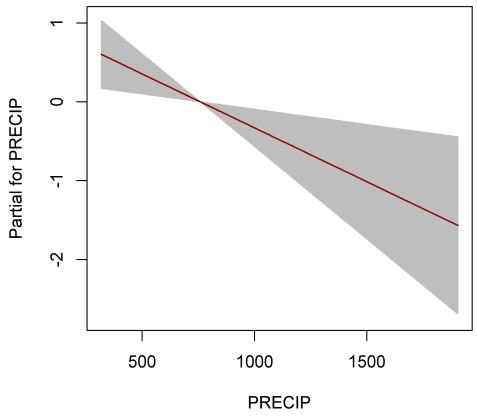
Conforme visto na Equação (5.1), apenas a variável precipitação apresentou relação significativa com o preditor do parâmetro ν. Na Figura 5.10 é possível observar que a variável apresenta uma relação linear inversa com o preditor do parâmetro, obtendo-se menores valores do parâmetro para maiores valores de precipitação.

Figura 5.9 – Relação entre as variáveis explanatórias e o preditor do parâmetro σ.



Fonte: Elaboração própria.

Figura 5.10 – Relação entre a variável precipitação e o preditor do parâmetro ν.

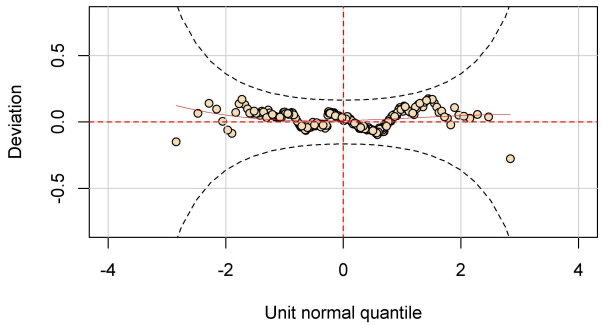


Fonte: Elaboração própria.

A Figura 5.11 mostra os resíduos do modelo final (após a inclusão da componente espacial) dispostos no *worm plot*. Pode-se perceber que o modelo se adequou bem à taxa de incidência de dengue nos municípios do estado da Paraíba e que há uma aparente melhoria no modelo com a inclusão da componente espacial se compararmos ao resultado apresentado na Figura 5.7 para o modelo Gama Generalizado, uma vez que os resíduos estão ainda mais próximos à linha horizontal do gráfico, conforme pode ser visto na Figura 5.11, o que indica que a distribuição dos resíduos se aproxima ainda mais de uma distribuição normal padrão.

Na Figura 5.12 é possível observar que o modelo Gama Generalizado foi bem ajustado à taxa de incidência de dengue nos municípios do estado da Paraíba, em concordância com o que foi observado na Figura 5.11. Vê-se um comportamento aleatório dos resíduos, que pode ser observado nos gráficos de dispersão na linha superior da Figura 5.12. No gráfico da densidade estimada pode-se perceber um comportamento aparentemente nor-

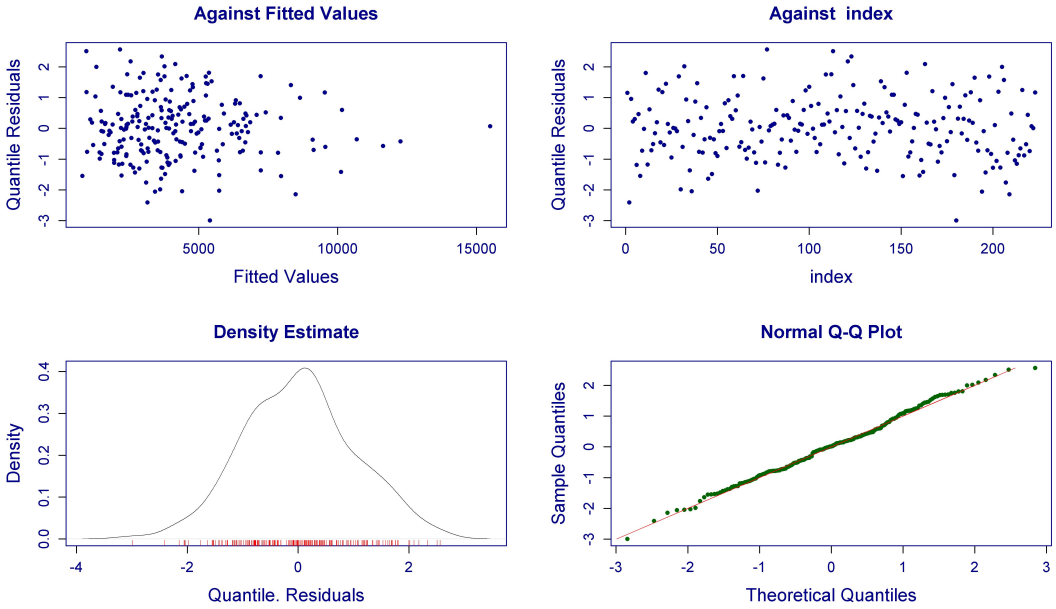
Figura 5.11 – *Worm plot* do modelo Gama Generalizado após a inclusão da componente espacial.



Fonte: Elaboração própria.

mal, o que é reforçado pelo gráfico *Normal Q-Q Plot*, com os resíduos se distribuindo de maneira linear sobre a reta.

Figura 5.12 – Análise de diagnóstico para o modelo GG.



Fonte: Elaboração própria.

5.4 Discussão

Para o período de estudo considerado neste trabalho (2008-2018), verificou-se que 2016 foi o ano com o maior número de casos notificados de dengue no estado da Paraíba, com quase sete mil casos notificados. Quanto aos municípios, Princesa Isabel, Monte Horebe, Monteiro e Zabelê apresentaram as maiores taxas de incidência por dengue no estado.

Foi identificado uma dependência espacial presente na taxa de incidência de dengue nos municípios do estado da Paraíba, tal como visto em Silva et al. (2020), onde a região da Mata Paraibana concentrou os municípios com as menores taxas de incidência. Enquanto que, no Sertão Paraibano e na região da Borborema observou-se as maiores taxas de incidência. Muitos dos municípios localizados nessas regiões não só apresentaram altas taxas de incidência, como também seus vizinhos. Estes municípios requerem uma maior atenção por parte das autoridades competentes, para a inclusão de medidas que visem a redução das taxas de incidência por dengue nessas localidades.

Pode-se verificar que a taxa de incidência de dengue na Paraíba no período de estudo foi influenciada por fatores socioeconômicos e ambientais, como no estudo de González, Beltrán e Guzmán (2017), que evidenciaram um aumento nos casos de dengue ligado a influência de fatores sociais. Já Tannous et al. (2021) verificaram que fatores ambientais como precipitação, temperatura e umidade do ar contribuíram de forma direta para o aumento da incidência da dengue no município de Jataí, sudoeste de Goias, Brasil.

Também Kikuti et al. (2015) apontaram que baixas condições socioeconômicas estiveram associadas ao aumento no risco de ocorrência de dengue na cidade de Salvador, Brasil. Em contrapartida, Telle et al. (2016) identificaram que não só áreas pobres como também áreas com melhores condições socioeconômicas foram afetadas com a presença da dengue, na cidade de Delhi, Índia. Esses resultados estão de acordo com o achado neste trabalho que apontou maiores taxas de incidência em municípios com melhores valores do IDH, menor desigualdade econômica e melhores condições de saneamento básico.

A baixa precipitação contribuiu para o aumento na taxa de incidência no estado. Regiões como o Sertão Paraibano, com baixa precipitação e períodos prolongados de seca, acarretam no abastecimento e armazenamento inadequado de água, que por consequência pode causar um aumento no número de vetores na região. Este resultado corrobora o estudo de Gehrke et al. (2020), que relacionaram o alto número de casos de dengue no ano de 2015 no estado de São Paulo com a crise hídrica sofrida no mesmo ano. Tal fato levou a população a armazenar água das chuvas para suprir suas necessidades, propiciando um ambiente adequado para o aumento no número de criadouros do vetor.

Entender o comportamento da doença em relação a estes fatores pode servir de subsídio não só para auxiliar a tomada de decisão de autoridades competentes, como também para a população, de forma que esta se torne mais alerta diante da presença de

fatores potencializadores da ocorrência de dengue e executem ações de controle do vetor, como tratamento adequado do lixo e não deixar água parada em recipientes abertos.

5.5 Conclusões

Neste estudo fica evidente que a dengue não ocorre de forma aleatória no estado da Paraíba. A taxa de incidência de dengue no estado está fortemente atrelada à geolocalização dos municípios, com regiões apresentando altas taxas de incidência, assim como regiões que agrupam municípios com baixas taxas de incidência. Viu-se ainda que a ocorrência da doença é influenciada pela combinação de uma série de fatores ligados ao contexto social, ambiental e econômico da região de estudo.

Ficou claro, que a taxa de incidência de dengue é impulsionada em regiões com baixas precipitações e altas temperaturas, onde a coleta e armazenamento de água é essencial para a sobrevivência. Viu-se também que altas taxas podem ser encontradas em regiões com bons índices socioeconômicos e que medidas de prevenção, como tratamento adequado do esgoto sanitário pode diminuir a taxa de incidência de dengue no estado da Paraíba.

Este trabalho serve como fonte de consulta para o melhor entendimento do comportamento da taxa de incidência de dengue no estado da Paraíba e fatores potencializadores no aumento dos casos notificados. Podendo dessa forma, contribuir para a tomada de decisão das autoridades responsáveis no que tange à atribuição de medidas mitigadores para diminuição dos casos de dengue no estado.

6 CONSIDERAÇÕES FINAIS

Baseado no que foi exposto no presente trabalho, pode-se concluir que os modelos da família GAMLSS podem servir como uma ferramenta muito poderosa do ponto de vista de modelagem estatística para lhe dar com dados de extrema complexidade. Considerando o fato de que dificilmente seria possível atender aos pressupostos de modelos mais comuns diante da modelagem de dados com características adversas, dificultando a validação desses modelos.

Apesar do pequeno número de variáveis explanatórias obtidas para o estudo dos casos de Tuberculose bovina no estado de Minas Gerais e das relações já esperadas dessas variáveis com a resposta. Os modelos apresentaram bom ajuste, mesmo diante das características apresentas nesta amostra, como alta simetria e curtose, excesso de zeros e superdispersão.

No estudo das ocorrências de casos de dengue nos municípios do estado da Paraíba, a inclusão da componente espacial no modelo tornou-se imprescindível, uma vez que a quebra do pressuposto de independência das obervações comprometem as estimativas do modelo e as relações das variáveis explanatórias com a resposta. Assim, as conclusões retiradas do modelo aqui ajustado podem de fato ser usadas para auxiliar a tomada de decisão de questões relacionadas ao problema expresso nesse capítulo, tendo em vista que o modelo apresentou boa adequabilidade e atende as pressuposições exigidas.

REFERÊNCIAS

ALMEIDA, A. S. d.; MEDRONHO, R. d. A.; VALENCIA, L. I. O. Análise espacial da dengue e o contexto socioeconômico no município do rio de janeiro, rj. **Revista de Saúde Pública**, SciELO Brasil, v. 43, n. 4, p. 666–673, 2009.

ALMEIDA, E. **Econometria Espacial Aplicada**. Campinas, SP: Alínea, 2012.

ANSELIN, L. Local indicators of spatial association—lisa. **Geographical analysis**, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995.

ANSELIN, L.; REY, S. Properties of tests for spatial dependence in linear regression models. **Geographical analysis**, Wiley Online Library, v. 23, n. 2, p. 112–131, 1991.

ASSUNÇÃO, R. M. Estatística espacial com aplicações em epidemiologia, economia e sociologia. **São Carlos: Associação Brasileira de Estatística**, v. 131, 2001.

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. **Hierarchical modeling and analysis for spatial data**. Boca Raton: CRC press, 2014.

BARBIERI, J. d. M. et al. Epidemiological status of bovine tuberculosis in the state of minas gerais, brazil, 2013. **Semina-Ciencias Agrarias**, UNIV ESTADUAL LONDRINA, v. 37, n. 5, p. 3531–3548, 2016.

BARBOSA, I. R.; SILVA, L. P. da. Influência dos determinantes sociais e ambientais na distribuição espacial da dengue no município de natal-rn. **Revista Ciência Plural**, v. 1, n. 3, p. 62–75, 2015.

BASTIDA, A. Z. et al. Spatial analysis of bovine tuberculosis in the state of mexico, mexico. **Veterinaria Italiana**, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise"G. Caporale, 2017.

BELCHIOR, A. P. C. et al. Prevalence and risk factors for bovine tuberculosis in minas gerais state, brazil. **Tropical animal health and production**, Springer, v. 48, n. 2, p. 373–378, 2016.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974.

BESAG, J. Statistical analysis of non-lattice data. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 24, n. 3, p. 179–195, 1975.

BHATT, S. et al. The global distribution and burden of dengue. **Nature**, Nature Publishing Group, v. 496, n. 7446, p. 504–507, 2013.

BIVAND, R.; PIRAS, G. Comparing implementations of estimation methods for spatial econometrics. **Journal of Statistical Software**, v. 63, n. 18, p. 1–36, 2015. Disponível em: <https://www.jstatsoft.org/v63/i18/>.

BIVAND, R.; WONG, D. W. S. Comparing implementations of global and local indicators of spatial association. **TEST**, v. 27, n. 3, p. 716–748, 2018.

BIVAND, R. S.; PEBESMA, E.; GOMEZ-RUBIO, V. **Applied spatial data analysis with R**. 2. ed. New York: Springer, 2013.

BROOK, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. **Biometrika**, JSTOR, v. 51, n. 3/4, p. 481–483, 1964.

CAMPOS, D. I. **Condenação de carcaças bovinas por Tuberculose, Brucelose e Cisticercose em abatedouro-frigorífico de Uberaba - MG e métodos de diagnóstico de Tuberculose em carcaças**. Tese (Doutorado) — Universidade Federal de Uberlândia, 2019.

CAO, Z. et al. Individual and interactive effects of socio-ecological factors on dengue fever at fine spatial scale: A geographical detector-based analysis. **International journal of environmental research and public health**, Multidisciplinary Digital Publishing Institute, v. 14, n. 7, p. 795, 2017.

CASSIDY, A. **Vermin, victims and disease: British debates over bovine tuberculosis and badgers**. Exeter: Springer Nature, 2019.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

CONCEIÇÃO, M. L. da et al. Phylogenomic perspective on a unique mycobacterium bovis clade dominating bovine tuberculosis infections among cattle and buffalos in northern brazil. **Scientific reports**, Nature Publishing Group, v. 10, n. 1, p. 1–13, 2020.

CORDEIRO, G. M.; DEMETRIO, C. G. B. Modelos lineares generalizados e extensões. Unpublished. 2013.

CRESSIE, N. **Statistics for spatial data**. USA: John Wiley & Sons, 1993.

DE BASTIANI, F. et al. Gaussian markov random field spatial models in gamlss. **Journal of Applied Statistics**, Taylor & Francis, v. 45, n. 1, p. 168–186, 2018.

DE BASTIANI, F.; STASINOPOULOS, M. gamlss. spatial: Package to fit spatial data in gamlss. **R package version**, v. 1, 2015.

EDWARDS, D. **Introduction to graphical modelling**. New York: Springer Science & Business Media, 2012.

FISCHER, M. M.; WANG, J. **Spatial data analysis: models, methods and techniques**. Berlin: Springer Science & Business Media, 2011.

GEARY, R. C. The contiguity ratio and statistical mapping. **The incorporated statistician**, JSTOR, v. 5, n. 3, p. 115–146, 1954.

GEHRKE, F. de S. et al. Análise espacial dos casos de dengue e correlação com dados pluviométricos em são paulo no período de 2015 a 2016. **Saúde e meio ambiente: revista interdisciplinar**, v. 9, p. 264–275, 2020.

GOMES, B. S. de M.; BASTOS, S. Q. de A.; NASCIMENTO, B. R. Uma avaliação espacial da incidência da dengue nos municípios de minas gerais, nos anos 2000 e 2010. **Ensaios FEE**, v. 38, n. 1, p. 35–74, 2017.

GONÇALVES, V. S. P. et al. Situação epidemiológica da brucelose bovina no estado de minas gerais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, SciELO Brasil, v. 61, p. 35–45, 2009.

GONZÁLEZ, A. A. M.; BELTRÁN, F. G. O.; GUZMÁN, L. F. S. Modelo bayesiano para el estudio de la enfermedad del dengue en el departamento de atlántico, colombia, años 2010 a 2013. **Perspectiva Geográfica: Revista del Programa de Estudios de Posgrado en Geografía**, Universidad Pedagógica y Tecnológica de Colombia, v. 22, n. 2, p. 85–104, 2017.

HASTIE, T. J.; TIBSHIRANI, R. J. **Generalized additive models**. London: Chapman & Hall, 1990.

HEIJDEN, E. M. van der et al. Mycobacterium bovis prevalence affects the performance of a commercial serological assay for bovine tuberculosis in african buffaloes. **Comparative immunology, microbiology and infectious diseases**, Elsevier, v. 70, p. 101369, 2020.

HONÓRIO, N. A.; OLIVEIRA, R. Lourenço-de. Freqüência de larvas e pupas de aedes aegypti e aedes albopictus em armadilhas, brasil. **Revista de Saúde Pública**, SciELO Public Health, v. 35, p. 385–391, 2001.

KIKUTI, M. et al. Spatial distribution of dengue in a brazilian urban slum setting: role of socioeconomic gradient in disease risk. **PLoS Negl Trop Dis**, Public Library of Science, v. 9, n. 7, p. e0003937, 2015.

LIMA, C. L. et al. Influência de indicadores sociais nos casos positivos de dengue no interior do estado de são paulo. **UNIFUNEC CIÊNCIAS DA SAÚDE E BIOLÓGICAS**, v. 1, n. 2, p. 25–37, 2017.

LIMA, V. H. et al. Achados ultrassonográficos, clínico-laboratoriais e anatomopatológicos em bovinosdiagnosticados com tuberculose – análise de 5 casos. **Revista Agrária Acadêmica**, v. 3, n. 1, 2020.

LOPATATZIDIS, A.; GREEN, P. Nonparametric quantile regression using the gamma distribution. **Private Communication**, 2000.

MAPA. **Sistema de Informação em Saúde Animal**. 2020. Disponível em: <http://antigo.agricultura.gov.br/assuntos/sanidade-animal-e-vegetal/saude-animal/sistema-informacao-saude-animal>.

MARTÍNEZ, H. Z. et al. Spatial epidemiology of bovine tuberculosis in mexico. **Veterinaria Italiana**, v. 43, n. 3, p. 629–634, 2007.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. Chicago: CRC press, 1989. v. 37.

MESSINA, J. P. et al. The current and future global distribution and population at risk of dengue. **Nature microbiology**, Nature Publishing Group, v. 4, n. 9, p. 1508–1515, 2019.

MILNE, G. et al. Spatiotemporal analysis of prolonged and recurrent bovine tuberculosis breakdowns in northern irish cattle herds reveals a new infection hotspot. **Spatial and spatio-temporal epidemiology**, Elsevier, v. 28, p. 33–42, 2019.

MORAN, P. A. Notes on continuous stochastic phenomena. **Biometrika**, JSTOR, v. 37, n. 1/2, p. 17–23, 1950.

NAKAMURA, L. R. et al. A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. **Journal of Applied Statistics**, Taylor & Francis, p. 1–16, 2018.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v. 135, n. 3, p. 370–384, 1972.

OIE. **Bovine tuberculosis: global distribution and implementation of prevention and control measures according to WAHIS data**. [S.l.], 2019. Disponível em: <https://oiebulletin.com>.

PAULA, G. A. Modelos de regressão com apoio computacional. Unpublished. 2013.

PEREZ, A. M. et al. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in argentina. **Preventive veterinary medicine**, Elsevier, v. 56, n. 1, p. 63–74, 2002.

PHANITCHAT, T. et al. Spatial and temporal patterns of dengue incidence in northeastern thailand 2006–2016. **BMC infectious diseases**, Springer, v. 19, n. 1, p. 743, 2019.

PLANT, R. E. **Spatial data analysis in ecology and agriculture using R**. Boca Raton: cRc Press, 2018.

PPM. **Produção da Pecuária Municipal 2018**. [S.l.], 2019. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/84/ppm_2018_v46_br_informativo.pdf>.

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

RIBEIRO, A. F. et al. Associação entre incidência de dengue e variáveis climáticas. **Revista de Saúde Pública**, SciELO Public Health, v. 40, p. 671–676, 2006.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. **Statistics and Computing**, Springer, v. 6, n. 1, p. 57–65, 1996a.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, D. M. Using the box-cox t distribution in gamlss to model skewness and kurtosis. **Statistical Modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 3, p. 209–229, 2006.

RIGBY, R. A.; STASINOPOULOS, D. M. Automatic smoothing parameter selection in gamlss with an application to centile estimation. **Statistical methods in medical research**, Sage Publications Sage UK: London, England, v. 23, n. 4, p. 318–332, 2014.

RIGBY, R. A.; STASINOPOULOS, M. D. Mean and dispersion additive models. In: **Statistical theory and computational aspects of smoothing**. [S.l.]: Springer, 1996b. p. 215–230.

RIGBY, R. A. et al. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: CRC press, 2019.

RUE, H.; HELD, L. **Gaussian Markov random fields: theory and applications**. Boca Raton: CRC press, 2005.

SANTOS, N. et al. Spatial analysis of wildlife tuberculosis based on a serologic survey using dried blood spots, portugal. **Emerging Infectious Diseases**, Centers for Disease Control and Prevention, v. 24, n. 12, p. 2169, 2018.

SILVA, E. T. C. d. et al. Análise espacial da distribuição dos casos de dengue e sua relação com fatores socioambientais no estado da paraíba, brasil, 2007-2016. **Saúde em Debate**, SciELO Public Health, v. 44, p. 465–477, 2020.

SOUZA FILHO, A. F. d. **Diversidade genética de isolados de Mycobacterium bovis em rebanhos bovinos de diferentes localidades do Brasil**. Tese (Doutorado) — Universidade de São Paulo, 2019.

SRINIVASAN, S. et al. Prevalence of bovine tuberculosis in india: A systematic review and meta-analysis. **Transboundary and emerging diseases**, Wiley Online Library, v. 65, n. 6, p. 1627–1640, 2018.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location scale and shape (gamlss) in r. **Journal of Statistical Software**, v. 23, n. 7, p. 1–46, 2007.

STASINOPOULOS, M.; RIGBY, R. **gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape**. [S.l.], 2020. R package version 5.1-6. Disponível em: <https://CRAN.R-project.org/package=gamlss.dist>.

STASINOPOULOS, M. D. et al. **Flexible Regression and Smoothing: Using GAMLSS in R**. Boca Raton: Chapman and Hall/CRC, 2017.

TANNOUS, I. P. et al. Mudanças sazonais no clima, índices pluviométricos e distribuição espacial de casos de dengue em um município do sudoeste de goiás-brasil. **Brazilian Journal of Development**, v. 7, n. 1, p. 6334–6349, 2021.

TELLE, O. et al. The spread of dengue in an endemic urban milieu–the case of delhi, india. **PloS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 1, p. e0146539, 2016.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. New Jersey: John Wiley & Sons, 2004. v. 368.

WOOD, S. N. **Generalized additive models: an introduction with R**. Boca Raton: Chapman and Hall/CRC, 2006.

YIN, C. et al. Effects of urban form on the urban heat island effect based on spatial regression model. **Science of the Total Environment**, Elsevier, v. 634, p. 696–704, 2018.