Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Ajuste de modelos lineares generalizados geoestatísticos em dados de contagem com diferentes abordagens agroclimáticas

Simone Grego

Tese apresentada para obtenção do título de Doutora em Ciências. Área de concentração: Estatística Experimentação Agronômica

Piracicaba
2019

Simone Grego
Bacharel em Estatística

Ajuste de modelos lineares generalizados geoestatísticos em dados de contagem com diferentes abordagens agroclimáticas

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Tese apresentada para obtenção do título de Doutora em Ciências. Área de concentração: Estatística Experimentação Agronômica

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Grego, Simone

Ajuste de modelos lineares generalizados geoestatísticos em dados de contagem com diferentes abordagens agroclimáticas / Simone Grego. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

41 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Geoestatística 2. Modelos lineares generalizados 3. Poisson 4. Efeito aleatório I. Título

DEDICATÓRIA

A meu pai, **Natal** *(in memorian),*
que se faz presente em todos os momentos da minha vida,
com todo meu amor e gratidão.

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me deu o dom da vida e me abençoa todos os dias com seu amor infinito.

A minha mãe Zilda e irmã Célia, que com muito carinho e apoio, não mediram esforços para que eu chegasse até essa etapa da vida. Sou imensamente grata pela paciência e incentivos.

A minha família, em especial, a meu irmão Sérgio, meus três queridos sobrinhos Eduardo, Caio e Henrique, e meus cunhados Mônica e Eduardo, que dignamente me apresentaram a importância da família e ao caminho da honestidade e persistência.

Ao meu namorado Daniel, por toda ajuda e contribuição valiosa para a minha jornada acadêmica. Pessoa que amo partilhar a vida, esteve ao meu lado em todos os momentos, felizes ou não.

A minha orientadora Profa. Sônia, que desde o início nunca me negou ajuda. Obrigada pela paciência, motivação e por toda confiança em mim depositada.

Aos amigos, em especial à Andreza, Douglas, Erasnilson, Everton, Naimara, Talita e Valiana, pelas alegrias e dores compartilhadas. Muito obrigada pela amizade e por me aguentarem.

Ao Prof. José Belasque pela ajuda e por disponibilizar os dados da Citrosuco.

Aos professores e funcionários do Departamento de Ciências Exatas da ESALQ/USP, por toda ajuda, disponibilidade e ensinamento todos esses anos.

A todos os professores do Departamento de Estatística da Faculdade de Ciências e Tecnologia de Presidente Prudente da "Universidade Estadual Paulista Júlio de Mesquita Filho" que contribuiram para meu crescimento profissional.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes pelo auxílio financeiro na forma de bolsa de estudos.

SUMÁRIO

RESUMO 6
ABSTRACT 7
LISTA DE FIGURAS 8
LISTA DE TABELAS 9
1. INTRODUÇÃO 11
REFERÊNCIAS 12
2. ANÁLISE ESPACIAL DE DADOS GEORREFERENCIADOS DE CONTAGEM DE CITROS CONTAMINADOS COM A DOENÇA BACTERIANA DO *GREENING* 15
RESUMO 15
ABSTRACT 15
2.1. INTRODUÇÃO 15
2.2. MATERIAL E MÉTODO 16
2.3. RESULTADOS E DISCUSSÃO 20
2.4. CONCLUSÕES 26
REFERÊNCIAS 27
3. MODELAGEM ESPAÇO-TEMPORAL DA PRECIPITAÇÃO NO ESTADO DO PARANÁ 29
RESUMO 29
ABSTRACT 29
3.1. INTRODUÇÃO 29
3.2. MATERIAL E MÉTODOS 30
3.3. RESULTADOS E DISCUSSÃO 33
3.4. CONCLUSÕES 38
REFERÊNCIAS 38
4 CONSIDERAÇÕES FINAIS 41

RESUMO

Ajuste de modelos lineares generalizados geoestatísticos em dados de contagem em diferentes abordagens agroclimáticas

Geralmente, na análise de dados com contagens, a suposição de equidispersão não é atendida, inviabilizando o uso do modelo clássico de regressão linear generalizado. Além disso, observações medidas ao longo do tempo e correlacionadas espacialmente, dificultam a modelagem visto que essas estruturas causam diferentes fontes de variações. Desse modo, esse trabalho procurou explorar a metodologia dos modelos lineares generalizados geoestatísticos a fim de possibilitar opções em duas áreas do conhecimento. Para tal, utilizou-se dados pluviométricos e relacionados a doença do *grenning*, os quais foram analisados no *software* R. Nos resultados do capítulo 2 destacam-se a tendência e padronização no acometimento da doença nas plantas, principalmente em um determinado local. No capítulo 3, tem-se a flexibilidade e eficiência da metodologia proposta, principalmente ao realizar as análises baseadas em estimativas iniciais oriundas de um modelo considerado bem ajustado.

Palavras-chave: Geoestatística, Modelos lineares generalizados, Poisson, Efeito aleatório

ABSTRACT

Adjustment of geostatistical generalized linear models in counting data in different agroclimatic approaches

Generally, in the analysis of data with counts, the hypothesis of equidistance is not met, making impossible the use of the classic model of generalized linear regression. In addition, observations measured over time and spatially correlated make modeling difficult, since these structures cause different sources of variation. Thus, this work sought to explore the methodology of generalized geostatistical linear models to enable options in two areas of knowledge. The results of chapter 2 highlight the tendency and standardization of disease involvement in plants, especially in a given location. In Chapter 3, we have the flexibility and efficiency of the proposed methodology, especially when performing the analyzes based on the initial estimates of a well considered model.

Keywords: Geostatistics, Generalized linear models, Poisson, Random effects

LISTA DE FIGURAS

Figura 1. A) Mapa da área localizada no município de Anhembi-SP e B) Área de produção de citros em estudo. 17

Figura 2. A) Boxplot da média do número de plantas doentes por ano; B) Histograma da frequência do número de plantas doentes em cada anos.....21

Figura 3. Gráficos das correlações cofenéticas para os anos A) 2008, B) 2009, C) 2010, D) 2013 e E) 2015. 22

Figura 4. Gráfico das distâncias interquatílicas A) 2008, B) 2009, C) 2010, D) 2013 e E) 2015. 22

Figura 5. Dispersão dos valores preditos em relação aos observados do número de plantas doentes para os anos de 2008, 2009, 2010, 2013 e 2015. 24

Figura 6. Mapa de previsão espacial do número de plantas doentes (NDP) para o modelo linear generalizado geoestatístico. 25

Figura 7. Localizações das áreas comerciais e não comerciais próximas à área em estudo. 26

Figura 8. Estações Meteorológicas do Estado do Paraná. 31

Figura 9. Box-Plot da quantidade de dias chuvosos por ano estudado. 34

Figura 10. Covariograma da precipitação pluviométrica. 34

Figura 11. Gráficos do ajuste de uma distribuição Poisson empírica e teórica considerando A) Histograma de densidade; B) Função de densidade cumulativa. 35

Figura 12. Gráfico semi-normal de probabilidade com envelope de simulação para o ajuste do modelo Poisson com efeito aleatório de ano. 36

Figura 13. Gráficos da validação cruzada entre os valores amostrados e preditos (linha azul: reta de regressão ajustada, linha vermelha: reta de referência 0 e 1) para os modelos A) Sem efeito de tempo; e B) Com efeito de tempo. 37

Figura 14. Gráfico dos valores preditos do MLGG com o tempo versus MLGG sem o tempo. 38

LISTA DE TABELAS

Tabela 1. Parâmetros do MLGG estimados via Cadeias de Markov para o número de plantas doentes. 23
Tabela 2. Parâmetros do MLGG estimados via Cadeias de Markov para a quantidade de dias chuvosos 37

1. INTRODUÇÃO

O estudo da distribuição espacial é de grande importância em diversas áreas do conhecimento, como saúde, meio ambiente, geologia, agronomia, entre muitas outras áreas, sendo mais difundida no últimos 30 anos (MONTEIRO et. al, 2004). Devido aos avanços nos Sistemas de Informação Geográfica (GIS) e nos Sistemas de Posicionamento Global (GPS), os quais estão relacionados com sistemas de informação espacial, foi possível facilitar o manuseio e análises de dados com representação espacial e possíveis fenômenos de ocorrência.

De acordo com Vieira (2000), a geoestatística é a metodologia mais apropriada para analisar o comportamento de dados que apresentem variabilidade espacial, uma vez que se diferencia da metodologia clássica pelo fato de considerar que as observações estão de alguma forma relacionadas a uma distribuição espacial. Assim, como em todas as técnicas estatísticas, o método baseia-se em um conceito probabilístico, em que a estimação dos valores não amostrados será realizada por meio de uma função de correlação espacial, de tal forma que se tenha controle do viés e variância mínima. Além disso, a ferramenta utilizada para a estimação desses valores é a krigagem, a qual faz ponderações nas observações mais próximas do ponto a ser estimado, sem que haja tendenciosidade.

Entretanto, toda a análise de dados deve considerar a natureza da variável resposta, seja contínua ou discreta. Assim, assume-se uma distribuição de probabilidade para a variável de interesse, de tal forma que consiga captar diferentes relações entre as observações. Uma alternativa bastante adotada é encontrar uma maneira de transformar a variável resposta, a fim de atender as pressuposições do modelo de regressão Gaussiano. Contudo, essa abordagem produz resultados insatisfatórios, visto que dificulta a interpretação, não contempla a natureza da variável estudada, e ainda quando se tem valores nulos a utilização da transformação logarítmica torna-se um problema (Paula, 2013).

Situações em que a variável de interesse é classificada como contagens, de forma geral, não podem ser analisadas pelo modelo linear normal, pois diante de todos os problemas relatados, tem-se que a transformação nos dados não contempla a relação média-variância, característica presente nessa estrutura de dados, o que poderá produzir estimativas com erros padrão inconsistentes e predições negativas para o número de eventos (King, 1989).

Nesse sentido, Nelder e Wedderburn (1972), propuseram a teoria dos modelos lineares generalizados (MLGs), como alternativa para análise de dados que necessitam de maior flexibilidade quanto à distribuição associada a variável resposta, sendo necessário apenas que essa pertença a família exponencial sob a forma canônica. Tal família contempla as distribuições Poisson, binomial, gama, normal inversa, além da própria distribuição normal (Mccullagh e Nelder, 1989).

Um modelo bastante consolidado na literatura para dados de contagens é a distribuição de Poisson. No entanto, o modelo possui como único parâmetro a taxa de ocorrência dos eventos, a qual representa a média e a variância da distribuição, ou seja, uma relação de identidade na qual média e variância são iguais e que caracteriza a equidispersão, característica particular do modelo (Ross, 2014).

Assim, diversos fenômenos naturais apresentam diferentes dispersões, sendo um dos casos a superdispersão que ocorre quando a variância é muito maior que a especificada pelo modelo atribuído aos dados. O fenômeno pode ocorrer devido à ausência de covariáveis no modelo, excesso de zeros observados, heterogeneidade nas unidades amostrais, correlações entre respostas individuais e agregações (Hinde e Demétrio, 1998). O ajuste de modelos que não levam em consideração a variação extra, pode acarretar em subestimações dos erros padrões das estimativas dos parâmetros, e com isso, em interpretações incorretas acerca das significâncias dos efeitos estimados (Moral, 2013).

Nesses casos, uma abordagem seria os modelos lineares generalizados mistos (MLGMs), os quais foram propostos por Breslow e Clayton (1993). A metodologia consiste em combinar os MLGs com efeitos aleatórios no preditor linear, permitindo acomodar diferentes fontes de variações e estruturas de correlações existentes entre as observações.

Aplicações da metodologia que mistura análise espacial com dados de contagens, não são tão abordadas na literatura. Destaca-se o trabalho apresentado por Diggle e Ribeiro (2007), que introduziu a teoria dos modelos lineares generalizados geoestatísticos (MLGGs). Essa classe de modelos viabiliza acomodar a dependência relacionada as medidas espaciais, e as estruturas complexas referentes às variabilidades inerentes ao estudo observacional.

Diante do exposto, o presente trabalho teve por finalidade apresentar aplicações nas mais diversas áreas do conhecimento relacionadas a metodologia geoestatística com modelos lineares generalizados. Dessa forma, no capítulo 2, tem-se a análise refrente aos dados de *greening,* sendo proposta a distribuição Poisson com variável resposta número de plantas doentes. No capítulo 3, utilizou-se dados pluviométricos com variável resposta quantidade de dias chuvosos, em que foram propostos modelos com e sem feito aleatório de tempo no predito linear, e ajustado o modelo linear generalizado padrão e misto para estimar os valores iniciais, que serão posteriormente incluídos no modelo geoestatístico.

REFERÊNCIAS

BRESLOW, Norman E e CLAYTON, David G. **Approximate inference in generalized linear mixed models**. Journal of the American statistical Association, v. 88, n. 421, p. 9–25, 1993.

DIGGLE, Peter J. e RIBEIRO, Paulo Justiniano. **Model-based Geostatistics (Springer Series in Statistics)**. [S.l: s.n.], 2007.

HINDE, John e DEMÉTRIO, Clarice G B. **Overdispersion: models and estimation**. Computational statistics & data analysis, v. 27, n. 2, p. 151–170, 1998.

KING, Gary. **Variance specification in event count models: From restrictive assumptions to a generalized estimator**. American Journal of Political Science, p. 762–784, 1989.

MCCULLAGH, P; NELDER, J.A. **Generalized linear models**. [S.l.]: Routledge, 1989.

MONTEIRO, Antônio Miguel Vieira e colab. **Análise espacial de dados geográficos**. Planaltina: Embrapa, 2004.

MORAL, Rafael de Andrade. **Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais**. . [S.l.]: Universidade de São Paulo. , 2013

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

PAULA, Gilberto A. **Modelos de Regressão com Apoio Computacional. 2013**. Citado na pág, v. 1, n. 9, p. 10, 2013.

ROSS, Sheldon M. **Introduction to probability models**. [S.l.]: Academic press, 2014.

VIEIRA, SIDNEY ROSE. **Geoestatística em estudos de variabilidade espacial do solo**. Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo, v. 1, p. 1–53, 2000.

2. ANÁLISE ESPACIAL DE DADOS GEORREFERENCIADOS DE CONTAGEM DE CITROS CONTAMINADOS COM A DOENÇA BACTERIANA DO *GREENING*

RESUMO

O *greenning* é uma doença bacteriana que pode atacar os pomares de citros ocasionando redução na produtividade, por isso, é de grande importância para o setor citrícola. Com o objetivo de verificar se a evolução da doença, causada pela bactéria *Candidatus Liberibacter Asiaticus*, segue um padrão espacial, o número de plantas doentes foi monitorado em pomares comerciais ao longo de 10 anos (2007-2016), na propriedade da Citrosuco, localizada em Anhembi no Estado de São Paulo. A modelagem espacial fornece informações relevantes quanto as localidades de maior probabilidade de ocorrência da praga, influenciando por exemplo, na tomada de decisão em relação à intervenção química. Foi realizado o teste de Mantel, pelo qual verificou-se que houve um padrão não aleatório em cinco anos. Por se tratar de uma variável resposta de contagem, utilizou-se a metodologia dos Modelos Lineares Generalizados Geoestatísticos, assumindo-se a distribuição de probabilidade Poisson. Os resultados encontrados confirmam que os padrões espaciais da doença são similares ao longo dos anos.

Palavras-chave: Distribuição de Poisson, Geoestatística, *Grenning*

ABSTRACT

Greening is a bacterial disease that can attack citrus orchards causing a reduction in productivity, so it is of great importance to the citrus sector. In order to verify if the evolution of the disease, caused by the bacterium Candidatus Liberibacter Asiaticus, follows a spatial pattern, the number of diseased plants was monitored in commercial orchards over 10 years (2007-2016), in the property of Citrosuco, located in Anhembi in the State of São Paulo. Spatial modeling provides relevant information on the locations with the highest probability of occurrence of the pest, influencing, for example, decision making in relation to chemical intervention. The Mantel test was performed, in which it was verified that there was a non-random pattern in five years. Because it is a counting response variable, the methodology of Generalized Linear Geostatistical Models was used, assuming the Poisson probability distribution. The results confirm that the spatial patterns of the disease are similar over the years.

Keywords: Poisson distribuition, Geoestatistics, *Grenning*

2.1. Introdução

O Brasil por ser um dos maiores produtores na área de citricultura, necessita de técnicas que viabilizem a oferta por produtos desse setor (LEAL et al., 2010). Nesse sentido, são crescentes os estudos relacionados às doenças que acometem as plantas, e um dos causadores mais prejudiciais nos pomares de citros é a doença do *Huanglongbing* (HLB), também conhecida por *greening*, sendo essa severamente destrutiva e de rápida propagação (GOTTWALD et al., 2007).

A doença que é transmitida pelo inseto *Diaphorina Citri*, foi encontrada inicialmente na China, África e Ásia (COSTA, 1942), e, registrada no Brasil pela primeira vez em 2004, nas regiões Centro e Leste do estado de São Paulo (FILHO et al., 2009). Atualmente, no Brasil está presente em todas as regiões citrícolas de São Paulo, Minas Gerais e Paraná e além de outros países como na Argentina, Paraguai, México, Estados Unidos e quase todos da América Central (DA CITRICULTURA, 2008).

O sintoma inicial da doença é destacado pela coloração amarela na folha, e posteriormente, provoca desfolhas, folhas secas, mortes nos ramos, maturação e deformação irregular nos frutos. Uma das medidas sugeridas para o controle da bactéria é eliminação das plantas que apresentarem incidência da doença, para que não sirvam de fontes de contaminação em pomares vizinhos e outras plantas. Segundo Ribeiro et al. (2014), recomenda-se também, realizar inspeções diárias no campo e, monitorar o inseto causador da doença via armadilhas. No cenário comercial, o impacto da doença foi o declínio da produtividade e no preço do produto final, induzindo baixo retorno financeiro entre os agricultores.

Pesquisas relacionadas aos avanços tecnológicos no controle do *grenning*, teve como estratégia analisar se há, ou não, um padrão de dependência espacial na doença. Segundo Gyenge et al. (1999), o estudo da distribuição espacial da população de insetos é de fundamental importância para estabelecer os melhores critérios para controle do inseto vetor. Uma das técnicas para estudar dependência espacial é a geoestatística, pois considera a posição no espaço (LIEBHOLD et al., 1993).

Na literatura estatística era bastante usual a utilização de dados transformados quando se tinha dados de contagem, o que gerava perda de informação, uma vez que não se usava a amostra no seu estado original. Nelder & Wedderburn (1972), propuseram os Modelos Lineares Generalizados (MLG), os quais são mais flexíveis quanto a distribuição da variável de interesse, exigindo apenas que essa pertença a família exponencial na sua forma canônica. Entretanto, o MLG não é capaz de captar a dependência entre as observações espaciais, nesse caso, uma extensão é o Modelo Linear Generalizado Geoestatístico (MLGG) proposto por Diggle e Ribeiro (2007), em que as medidas da variável resposta estão relacionadas espacialmente.

Diante do exposto, o objetivo desse trabalho foi verificar o comportamento espacial da doença bacteriana *greening* em um pomar de citros no estado de São Paulo, ao longo de 10 anos, no qual foi observado o número de plantas acometidas pela doença. Para a análise dos dados foi verificado se há um comportamento espacial por meio do teste de aleatorização de Mantel e, após a identificação da dependência, o modelo Poisson com efeito aleatório espacial foi proposto a fim de se obter mapas de predição para a infestação do *greening*.

2.2. Material e Método

2.2.1. O Experimento

Os dados de contagem de plantas de citros com a doença bacteriana do *greening*, foram provenientes de um estudo realizado pela empresa Citrosuco na área denominada "Propriedade Referência", localizada em Anhembi-SP com 645 ha de área cultivada. Essa é uma das localidades que adota um manejo de controle rigoroso do HLB, com mais de 4 erradicações nas plantas e mais de 24 pulverizações no inseto, ambos por ano.

Na Propriedade Referência as árvores doentes são contadas e eliminadas dentro de cada um dos 32 lotes, que são áreas delimitadas por ruas, como apresentado na Figura (1). Além disso, têm-se quadras de plantas cítricas, conjunto de 32 lotes de plantas, sob a influência de áreas externas, as quais são mantidas com diferentes manejos do HLB, ou até mesmo, sem manejo (Figura 1).

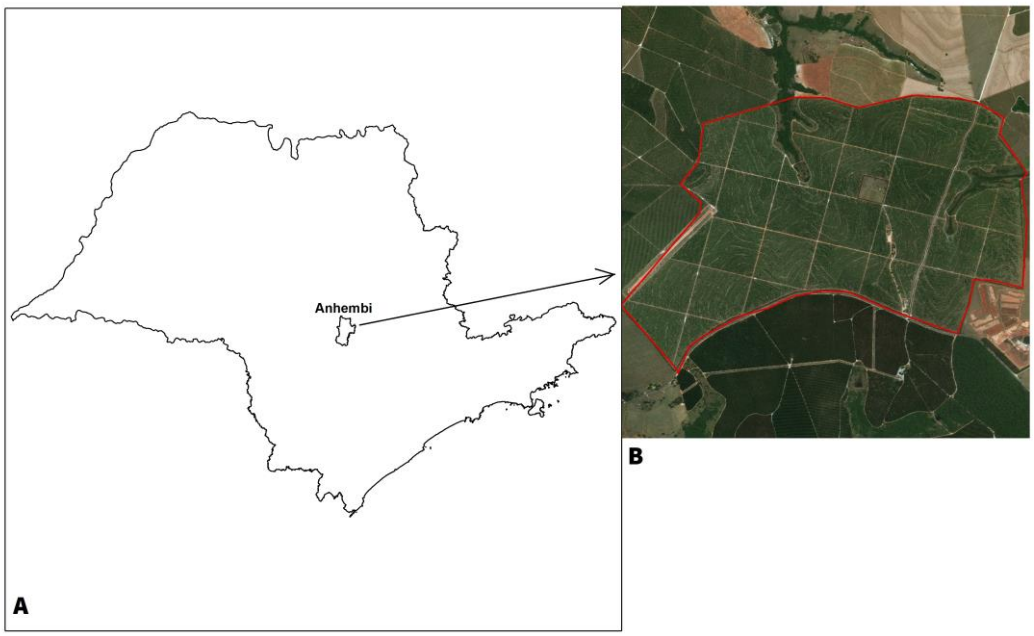

Figura 1. A) Mapa da área localizada no município de Anhembi-SP e B) Área de produção de citros em estudo.

Esse procedimento foi realizado trimestralmente desde o ano de 2007, tendo-se ao final de cada ano o número de plantas doentes removidas por lote. Com o interesse de verificar um padrão espacial da doença HBL na propriedade no decorrer dos anos (2007-2016), o centroide de cada lote foi adotado como coordenada do ponto georreferenciado.

2.2.2. Metodologia Estatística

Geoestatística

A geoestatística trata da modelagem e inferência para acontecimentos espacialmente contínuos ($S(\boldsymbol{x})$), em que as observações $(y_1, y_2, \cdots, y_n)$ são obtidas por amostragem em número finito de locais $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ (RIBEIRO et al., 2003). O modelo geoestatístico é especificado por $(y_i, \boldsymbol{x}_i)$, $com\ i = 1, \ldots, n$ e, $\boldsymbol{x}_i$, as localizações espaciais. Ademais, y_i é o valor medido associado à localização $\boldsymbol{x}_i$, contido em uma região espacial A, que tem processo Gaussiano $\{S(\boldsymbol{x}): \boldsymbol{x} \subset \mathbb{R}^d\}$ com média μ, variância $\sigma^2 = Var(S(\boldsymbol{x}))$ e função de correlação $\rho(\boldsymbol{h}) = Corr(S(\boldsymbol{x}), S(\boldsymbol{x}'))$, em que $\boldsymbol{h}$ é a distância entre duas

observações ($\boldsymbol{h} = ||\boldsymbol{x} - \boldsymbol{x}'||$). Pode-se considerar que a correlação espacial diminui conforme aumenta a distância (DIGGLE et al., 2007).

O modelo geoestatístico capta a dependência espacial por meio do componente de erro aleatório. Assim, supondo-se que $Y(\boldsymbol{x}_i)$ segue uma distribuição Gaussiana, tem-se

$$Y(\boldsymbol{x}) = S(\boldsymbol{x}) + Z_i, \tag{1}$$

sendo Z_i mutuamente independente com distribuição $\mathrm{N}(0, \tau^2)$ e $S(\boldsymbol{x})$ um processo estocástico estacionário, que não é diretamente observado. Vale ressaltar que esse último, em geral, não apresenta valores observados de forma direta. Para validar o modelo (1), é necessário que a correlação $\rho(\boldsymbol{h})$ seja positiva definida (DIGGLE et al, 2007). Com isso, a função de $\rho(\boldsymbol{h})$ é especificada por um modelo, nesse cápitulo, utilizou-se o *Matérn* (MØLLER, 2013)

$$\rho(\boldsymbol{h}) = \{2^{\nu-1}\Gamma(\nu)\}^{-1}(\boldsymbol{h}/\phi)^{\nu}K_{\nu}(\boldsymbol{h}/\phi), \tag{2}$$

em que $\Gamma(.)$ é a função gama, $\nu > 0$ e $\phi > 0$ são parâmetros do modelo, K_{ν} é a equação diferencial de Bessel na ordem modificada ν. Um caso particular da função *Matérn* é quando o valor de ν é $1/2$, em que a função se iguala a função exponencial. Na família *Matérn* o parâmetro kappa ν controla a suavidade do processo estocástico (CRESSIE, 1992).

Teste de Mantel

Em experimentos de campo há necessidade de se identificar a existência, ou não, de um padrão espacial. Isso se justifica pelo fato de entender, por exemplo, se o comportamento da infestação do inseto causador da doença na planta, acontece numa direção específica. Dessa forma, o teste utilizado para detecção e padrão espacial foi proposto por Mantel (1967), o qual considera as seguintes hipóteses,

$$\begin{cases} H_0: A\ dependência\ espacial\ é\ aleatória \\ H_1: A\ dependência\ espacial\ não\ é\ aleatória \end{cases} \Rightarrow \begin{cases} H_0: Não\ há\ padrão\ espacial \\ H_1: Há\ padrão\ espacial. \end{cases}$$

O teste é realizado por meio do coeficiente de correlação de Pearson, entre duas matrizes **A** e **B**, as quais são simétricas com dimensões $n \times n$ (DIGGLE et al, 2007). Portanto, a proposta é analisar os dados que apresentarem p-valor $< \alpha$, em que foi utilizado $\alpha = 0{,}05$.

Para verificar a associação entre a variável resposta em relação as covariáveis, propõe-se a regressão linear dada por

$$b_{ij} = \beta_0 + \beta_1 a_{ij} + \epsilon_{ij}, \tag{3}$$

em que a_{ij} e b_{ij} são os elementos da matriz **A** e **B**, respectivamente; β_0 e β_1 são os parâmetros a serem estimados e, ϵ_{ij} é o erro associado à variável resposta com $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. O modelo dado em (3), pode apresentar formas mais complexas na sua composição (MANTEL, 1967).

Modelos Lineares Generalizados Geoestatísticos (MLGG)

Proposto por Diggle et al. (2007), os MLGGs permitem modelar dados em que a distribuição de probabilidade da variável resposta não é Gaussiana, a qual é uma pressuposição da análise espacial clássica. Nesse contexto, considerando os dados apresentados na seção (2.2.1), o modelo proposto é definido na forma que se segue.

i) Componente aleatório, que é a variável resposta $\boldsymbol{Y}_i$ (número de plantas doentes), que tem uma distribuição condicional $\boldsymbol{Y}_i|\, S(\boldsymbol{x}_i)$, pertencente a família exponencial de distribuições sob a forma canônica, expressa por

$$f_{\boldsymbol{Y}}(y_i|\boldsymbol{S};\theta_i;\phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\}. \tag{4}$$

Para modelagem é utilizado o modelo Poisson, isto é,

$$\boldsymbol{Y}_i|S(\boldsymbol{x}_i) \sim Poisson(\lambda(x_i))$$

$S(\boldsymbol{x}_i)$ um processo estocástico estacionário gaussiano com distribuição $N_n(0, \mathbf{D})$, onde $\mathbf{D} = \sigma^2 C(\boldsymbol{h})$ e $C(\boldsymbol{h})$ é a distância euclidiana $\boldsymbol{h}$ entre duas localizações.

ii) Componente sistemático, que é a estrutura linear do modelo, representado por

$$\boldsymbol{\eta}_i = \boldsymbol{\beta} + S(\boldsymbol{x}_i), \qquad i = 1, \cdots, n, \tag{5}$$

em que $S(\boldsymbol{x}_i)$ é o parâmetro que capta a dependência espacial e $\boldsymbol{\beta}$ é o parâmetro de efeito fixo do modelo.

iii) Uma função de ligação que permita relacionar a média com o preditor linear $\boldsymbol{\eta}_i$, neste caso, é considerada a função de ligação logarítmica, por se propor o modelo Poisson, isto é

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i \tag{6}$$

$$log\big(\lambda(\boldsymbol{x}_i)\big) = \boldsymbol{\beta} + S(\boldsymbol{x}_i). \tag{7}$$

Nesse contexto, é importante ressaltar que a metodologia proposta por Diggle et al. (2007) além de modelar a estrutura espacial contida nos dados, também capta a variabilidade extra, como o fenômeno da superdispersão, a qual está presente em dados de contagens. Dessa forma, seja o modelo Poisson com a variável resposta $\boldsymbol{Y}$ o número de plantas doentes e, média $\alpha_i exp\big(S(\boldsymbol{x}_i)\big)$, sendo α_i o parâmetro relacionado ao número de plantas doentes, que captura a variabilidade extra do modelo, tem-se a densidade na forma

$$P[Y(\boldsymbol{x})|S(\boldsymbol{x}) = y(\boldsymbol{x}_i)] = \frac{e^{-\alpha_i e^{S(\boldsymbol{x}_i)}}\left(\alpha_i e^{s(\boldsymbol{x}_i)}\right)^{y(\boldsymbol{x}_i)}}{y(\boldsymbol{x}_i)!}. \tag{8}$$

Escrevendo-se a função de verossimilhança, descrita pela integral e, $f(\boldsymbol{S}, \boldsymbol{\phi})$ definida pela distribuição gaussiana conjunta de $\boldsymbol{S}$ com o parâmetro $\boldsymbol{\phi}$, segue-se

$$L(\boldsymbol{Y}_i|\boldsymbol{S},\boldsymbol{\phi}) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{S},\boldsymbol{\phi})\, f(\boldsymbol{S},\boldsymbol{\phi})\ ds\,, \tag{9}$$

$$L(\boldsymbol{Y}_i|\boldsymbol{S},\boldsymbol{\phi}) = \int \frac{e^{-\alpha_i e^{S(x)}}\left(\alpha_i e^{S(x)}\right)^{y(x_i)}}{y(\boldsymbol{x}_i)!} \cdot \frac{1}{\sqrt{2\pi\sigma^2\,|\boldsymbol{R}|}}\, e^{\frac{-1}{2\sigma^2}(S(\boldsymbol{x}_i)-\mu_i)^T \boldsymbol{R}(S(\boldsymbol{x}_i)-\mu_i)}\ ds, \tag{10}$$

em que $\boldsymbol{R}$ é a matriz de correlação amostral e a dimensionalidade $(n \times n)$ é igual ao número de observações.

Contudo, a expressão dada em (10) não apresenta uma solução analítica, assim, é necessário a utilização de métodos numéricos para a obtenção das estimativas dos parâmetros. Nesse sentido, uma possível abordagem é o algoritmo Cadeias de Markov via Monte Carlo (MCMC), o qual constitui em uma classe de algoritmos de simulação sequencial, baseado em valores iniciais dos parâmetros, gerando assim uma cadeia com valores simulados de forma sequencial, em que cada valor depende do anterior (BONAT et al., 2012). Tal algoritmo está implementado no pacote geoRglm (CHRISTENSEN e RIBEIRO, 2007), o qual foi utilizado nesse trabalho, considerando uma cadeia de 500.000 interações, ciclo de pré-convergência (*burn in*) de 10.000 e armazenando uma amostra a cada 100 amostras geradas, no total fazendo 5.000 amostras.

A fim de validar o modelo proposto foi utilizado o semivariograma, uma vez que esse é indicado para análise de diagnóstico (RIBEIRO JR et al., 2009). Assim, comparam-se os variogramas experimental e ajustado, os quais foram obtidos por meio dos resíduos expressos por

$$2\gamma_Y(\boldsymbol{h}) = Var[Y(\boldsymbol{x}+\boldsymbol{h}) - Y(\boldsymbol{x})] \tag{11}$$

$$2\gamma_{Y(\boldsymbol{h})} = Var[Y(\boldsymbol{x}+\boldsymbol{h})] + Var[Y(\boldsymbol{x})] - cov[Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})], \tag{12}$$

$$\gamma_{Y(\boldsymbol{h})} = exp\left(\beta + \frac{\sigma^2}{2}\right) + exp(2\beta+\sigma^2)\{exp(\sigma^2) - exp\left(\sigma^2\rho(\boldsymbol{h})\right)\}, \tag{13}$$

em que $\boldsymbol{h}$ é a distância euclidiana (DIGGLE et al., 1998).

Após definir o modelo, foi construído o mapa de predição a fim de descrever o comportamento da variável de interesse em uma determinada localidade. Todas as análises foram implementadas no *Software* R (TEAM, 2013), que permite o uso de bibliotecas externas de comandos, que manipulam arquivos cartográficos, admitindo gerar e configurar visualização cartográfica consequente dos métodos empregados.

2.3. Resultados e Discussão

Inicialmente foram realizadas análises descritivas em que foi observado que existe um comportamento tendencioso, crescente no tempo, em relação à quantidade de plantas acometidas pela

doença (Figura 2A e 2B). Além disso, foi detectado uma maior dispersão no decorrer dos anos, sendo destacado maiores ocorrências o ano de 2016.

Uma possível explicação do aumento na quantidade de doenças é devido à falta de conhecimento prévio, quanto ao deslocamento da *Diaphorina Citri,* e ausência de rigor no que tange ao controle químico nas localidades vizinhas da área do plantio em estudo.

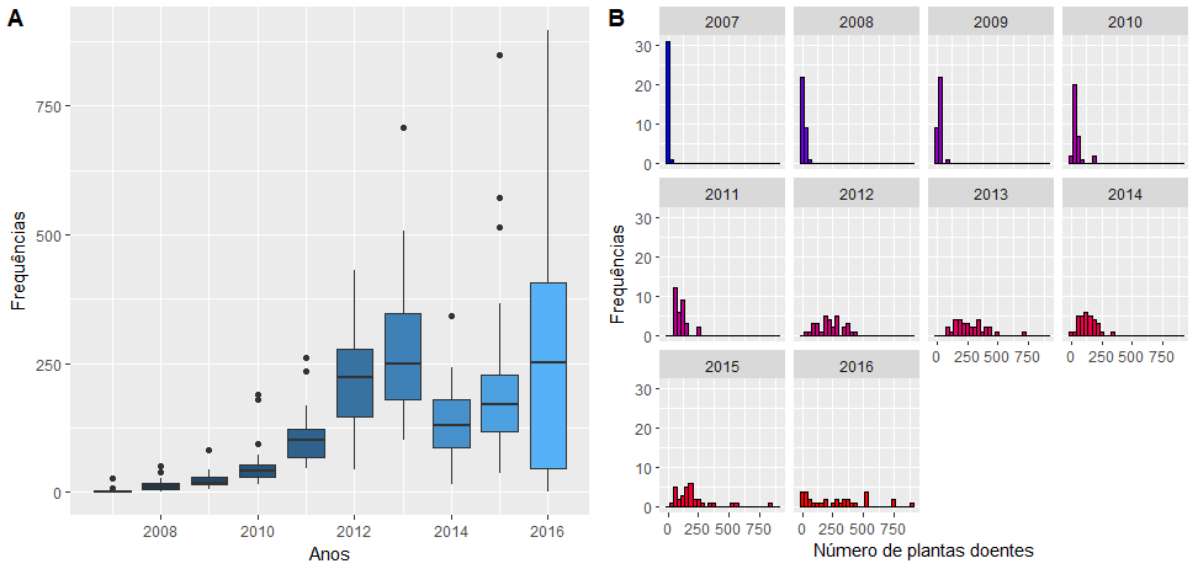

Figura 2. A) Boxplot da média do número de plantas doentes por ano; B) Histograma da frequência do número de plantas doentes em cada anos.

A geoestatística tem como objetivo principal analisar o comportamento da distribuição espacial nas observações, determinando se há, ou não, um padrão sistemático, aleatório ou em grupos. Nesse sentido, foi aplicado o teste de Mantel, que analisou se havia um comportamento padrão nas observações ao longo dos 10 anos de estudo. Assim, os p-valores obtidos para o número de plantas doentes nos anos 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015 e 2016 foram respectivamente 0,75; **0,04; 0,02**; **0,02**; 0,1; 0,29; **0,04**; 0,23; **0,03** e 0,52.

Observa-se que em 2008, 2009, 2010, 2013 e 2015 houve indícios significativos, considerando 95% de confiança, de que existiu um padrão espacial referente ao número de plantas doentes. Adicionalmente, na Figura (3) têm-se os gráficos das correlações cofenéticas, os quais mostraram associações positivas, de intensidade fraca, indicando baixa dependência entre as distâncias dos pontos observados.

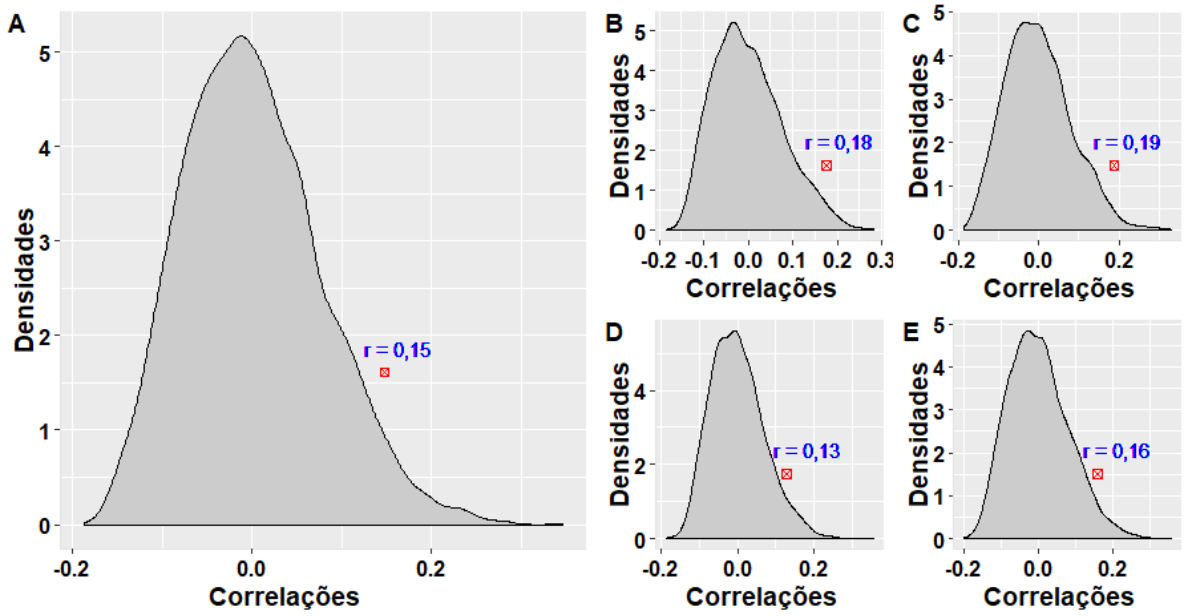

Figura 3. Gráficos das correlações cofenéticas para os anos A) 2008, B) 2009, C) 2010, D) 2013 e E) 2015.

A Figura 4 é representada pelo gráfico dos quartis de uma distribuição empírica dos valores medidos, e pretendeu-se explorar as observações de tal modo que, ao analisar o comportamento em relação às coordenadas geográficas, pode-se verificar se há sinal de tendência quanto à quantidade de plantas doentes e possíveis influências dos pontos observados. No ano de 2013 notou-se maior incidência da doença principalmente nas bordas ao leste. Além disso, os quartis indicaram a existência de um padrão espacial na variável de interesse, uma vez que existem conglomerados das categorias.

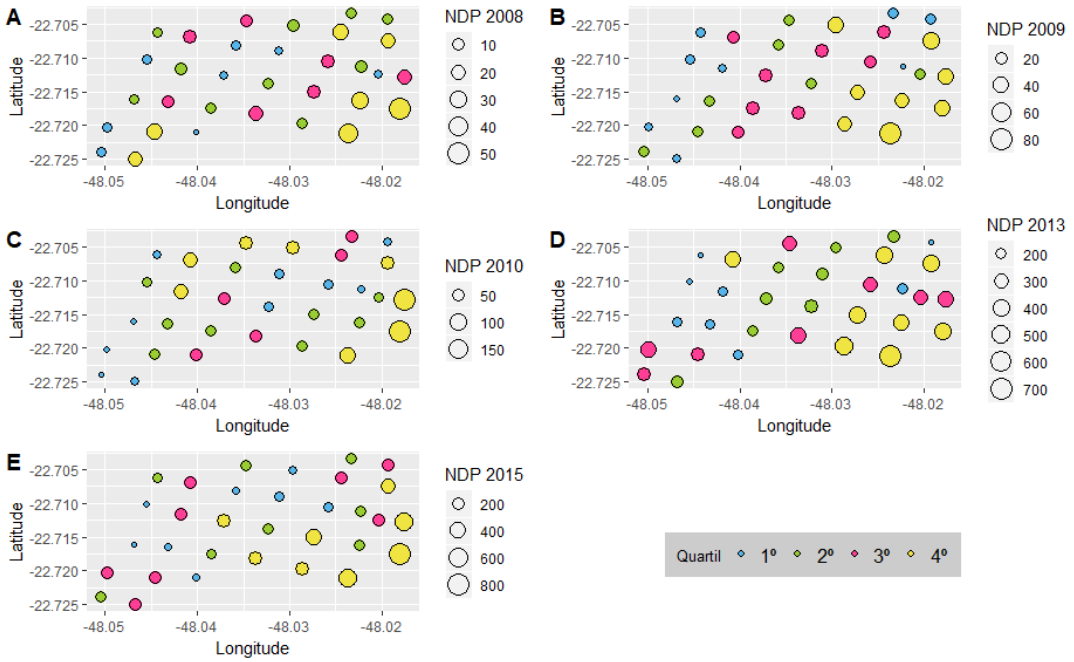

Figura 4. Gráfico das distâncias interquartílicas A) 2008, B) 2009, C) 2010, D) 2013 e E) 2015.

Considerando o modelo (8) obteve-se as estimativas dos parâmetros e a função do logaritmo da verossimilhança conforme mostra a Tabela 1. Ao ajustar o modelo considerando a pressuposição da média constante, tem-se o parâmetro da média geral associada a função de ligação ($\hat{\beta}$); a variância do processo espacial, que define a estrutura de covariância das observações ($\hat{\sigma}^2$); o erro de predição ou efeito pepita ($\hat{\tau}^2$), sendo esses três últimos associados à superfície de S(**x**); e o parâmetro da função de correlação ($\hat{\phi}$).

Fixando o valor *kappa* = 0,5 observou-se nas estimativas de $\hat{\sigma}^2$ e $\hat{\phi}$ indicativos que houve uma dependência espacial. De acordo com BONAT et al., (2011), valores elevados da proporção da variância (ω), proveniente de $\hat{\sigma}^2$ em relação a $(\hat{\sigma}^2 + \hat{\tau}^2)$, indicam dependência espacial quando os valores se aproximarem de 1. Assim, pode-se dizer que houve maiores dependências espaciais em todos os anos, e, consequentemente, se terão melhores predições (BONAT et al., 2011).

Vale ressaltar que o patamar $(\hat{\sigma}^2 + \hat{\tau}^2)$ representa o valor de estabilização do semivariograma, ou seja, quando a variância dos dados se torna constante devido ao aumento da distância entre as amostras. Nesse caso, a distância limite entre a dependência e independência entre as amostras foi maior no ano de 2008. Considerando-se o modelo *Matèrn*, assintoticamente, a correlação aproxima-se de zero e o alcance prático (AP), em que $\hat{\phi}$ é multiplicado por uma constante 3. Definido como a distância em que a correlação é 0,05 (DIGGLE e RIBEIRO, 2007), apresentou valores baixos, isto é, alcançou a independência de forma rápida.

Tabela 1. Parâmetros do MLGG estimados via Cadeias de Markov para o número de plantas doentes.

Ano	Estimativas						AP
	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$(\hat{\sigma}^2 + \hat{\tau}^2)$	$\hat{\omega}$	
2008	2,38	2,06	1029,00	0,02	2,08	0,99	3087,00
2009	3,01	0,65	980,30	0,05	0,70	0,93	2940,90
2010	3,73	0,54	821,10	0,03	0,57	0,95	2463,30
2013	5,51	0,23	637,20	0,01	0,24	0,96	1911,60
2015	5,24	0,52	784,20	0,01	0,53	0,98	2352,60

Ao ajustar o modelo com distribuição Poisson a esses dados, foi possível comparar as predições feitas para cada ano (Figura 5), de modo geral, percebe-se boa estimativa para todos os anos, apresentando os pontos bem distribuídos em torno da reta, indicando que as estimativas do número de plantas doentes foram praticamente iguais as observações. Confirmando que o modelo em estudo apresentou predições adequadas para a variável estudada, como mostrado em SCALON et al., (2012), o modelo geoestatístico com distribuição Poisson apresentou melhores estimativas, e, consequentemente, melhores predições.

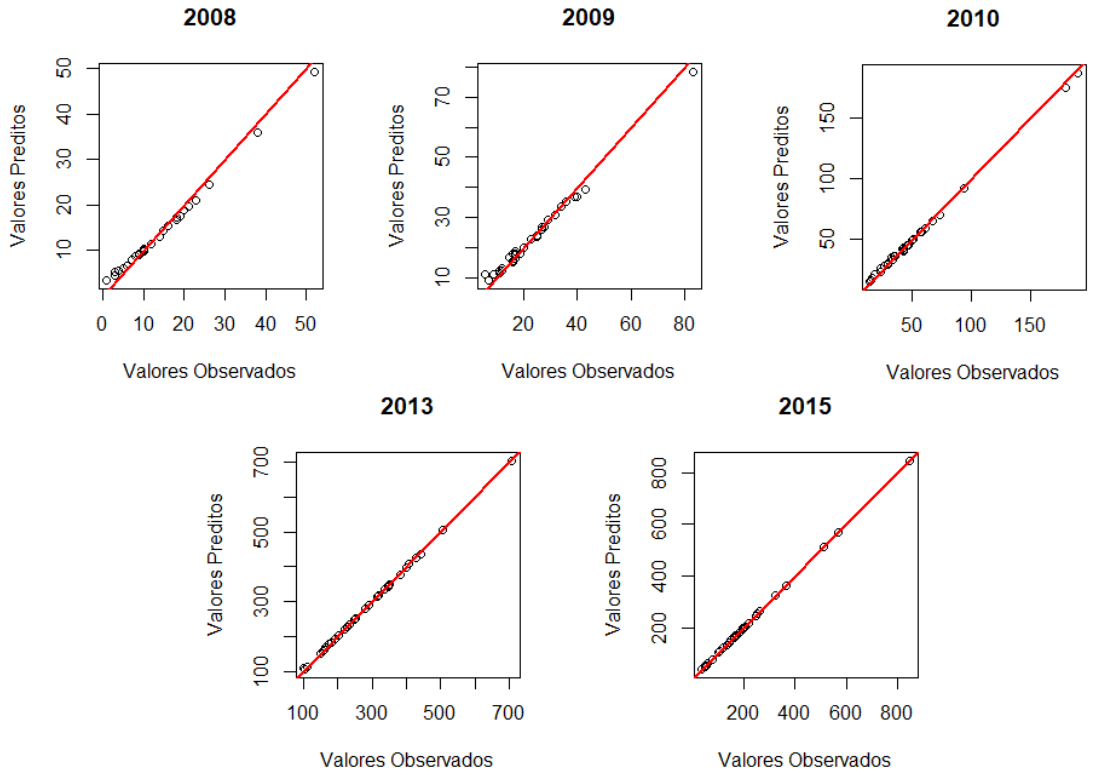

Figura 5. Dispersão dos valores preditos em relação aos observados do número de plantas doentes para os anos de 2008, 2009, 2010, 2013 e 2015.

A Figura 6 apresenta os mapas de predição de plantas doentes com *greening* obtidos por krigagem e gerados pelo MLGG proposto. Independentemente dos anos, observaram-se zonas mais diferenciadas do número de plantas doentes indicando maiores incidências nas bordas ao leste da propriedade. Uma explicação para a tendência da quantidade de números de plantas doentes no determinado local apresentado no mapa, é a presença de um plantio comercial que até 2013 não realizava nenhum controle de manejo contra a doença e também de cultivos não comerciais (Figura 7), aumentando a chance das plantas apresentarem sintomas da doença. As menores incidências ocorreram no lado oeste da divisa, onde existe propriedades comerciais de citros que realizam pulverizações de inseticidas para o controle de *Diaphorina Citri* periodicamente, isso também pode ser percebido em Arruda (2017) em que utilizou metodologia de modelos monomoleculares.

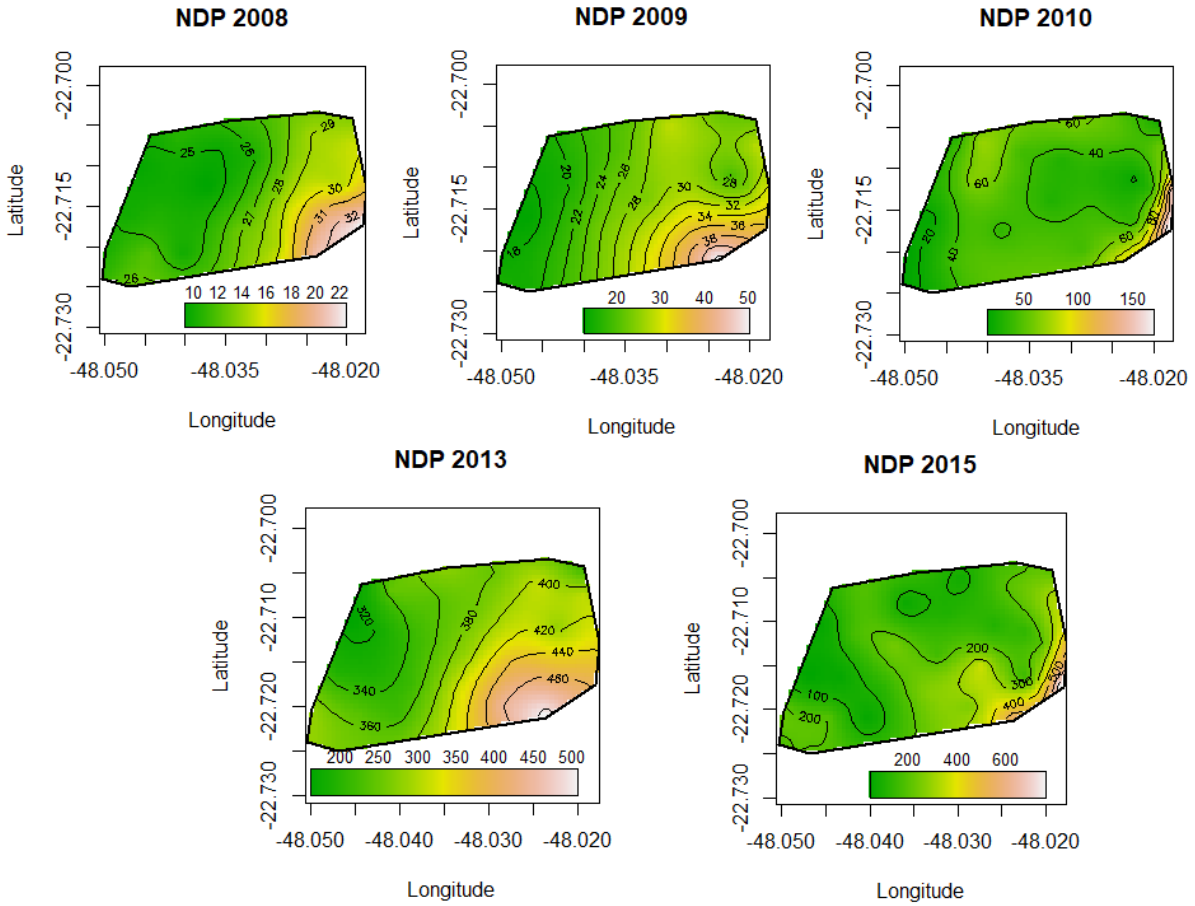

Figura 6. Mapa de previsão espacial do número de plantas doentes (NDP) para o modelo linear generalizado geoestatístico.

Nesse sentido, baseado no mapa, há indícios de alta incidência em um determinado local ao lado de plantios não comerciais. Dessa forma, é sugerido para o controle da bactéria, formas de manejo que contornem a contaminação da área identificada e nas proximidades com maior ocorrência, uma vez que a doença se espalha com extrema rapidez.

Fonte: Google Maps.

Figura 7. Localizações das áreas comerciais e não comerciais próximas à área em estudo.

2.4. Conclusões

O modelo linear generalizado geoestatístico mostrou-se adequado para a modelagem espacial da ocorrência de *Huanglongbing*, uma vez que a distribuição Poisson com efeito aleatório espacial conseguiu acomodar a variabilidade extra presente no conjunto de dados.

No contexto metodológico, as técnicas apresentadas e utilizadas nesse trabalho possibilitaram a detecção de padrões espaciais quanto ao número de plantas doentes, bem como a indicação de estratégias que viabilizam o controle da doença em condições de campo.

A inclusão das coordenadas da área e o parâmetro de suavidade para a função de correlação (*kappa*=0,5) permitiram aumentar a acurácia das estimativas. Entretanto, o padrão espacial não foi detectado em alguns anos. Uma possível explicação a baixa presença de plantas doentes em determinados períodos é a alta variabilidade nas observações. Ademais, a ausência de covariáveis no modelo MLGG pode ter ocasionado um aumento nos resíduos, levando a falta de significância no padrão espacial e diminuindo a precisão do ajuste.

Uma estratégia de controle do *Grenning* que dê uma atenção especial às bordas do pomar é uma abordagem interessante no combate à praga. São mais importantes fontes de contagio da doença as áreas não comerciais (residenciais, matas, pastagens ou canaviais) do que em plantios comerciais em que há controle químico do vetor.

Recomenda-se para estudos futuros a inclusão das estações do ano, verificando-se assim, se a doença muda o comportamento conforme a temperatura, regime pluviométrico e demais alterações causadas pelo ciclo intra-anual.

Referências

ARRUDA, Josicléa Hüffner. Disseminação primária e influência de fontes de inóculo sobre o progresso temporal do Huanglongbing em pomares com manejo da doença e seu vetor (Diaphorina citri). . [S.l.]: Universidade de São Paulo. , 2017

BONAT, Wagner Hugo e colab. Métodos Computacionais em Inferencia Estatística. Curitiba: Associação Brasileira de Estatística-ABE, SINAPE, 2012.

BONAT, Wagner Hugo e JUNIOR, Paulo Justiniano Ribeiro e ZEVIANI, Walmes Marques. Comparando predições por modelos geoestatístico e aditivo generalizado para reconstituição de superfícies contínuas gaussianas. Energia na Agricultura, v. 26, n. 2, 2011.

CâMARA, Gilberto; et al. Análise espacial de áreas. Análise espacial de dados geográficos. Brasília: EMBRAPA, p. 157–182, 2004.

CHRISTENSEN, Ole F. e RIBEIRO, Paulo Jr. The geoRglm Package. 2007.

COSTA, LIMA. AM da. Homópteros. COSTA LIMA, AM da. Insetos do Brasil. Rio de Janeiro: Escola Nacional de Agronomia, v. 3, 1942.

CRESSIE, Noel. Statistics for spatial data. Terra Nova, v. 4, n. 5, p. 613–617, 1992.

DA CITRICULTURA, FUNDECITRUS-Fundo de Desfesa. Disponível em: http://www. fundecitrus. com. br\doencas\greening. Acesso em, v. 7, 2008.

DIGGLE, Peter J. e RIBEIRO, Paulo Justiniano. Model-based Geostatistics (Springer Series in Statistics). [S.l: s.n.], 2007.

DIGGLE, Peter J e TAWN, Jonathan A e MOYEED, R A. Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics), v. 47, n. 3, p. 299–350, 1998.

FILHO, Hermes Peixoto Santos e BARBOSA, Francisco Ferraz Laranjeira e NASCIMENTO, Antonio Souza Do. Greening, a mais Grave e Destrutiva Doença dos Citros: Nova Ameaça à Citricultura. Embrapa Mandioca e Fruticultura - Cominicado Técnico (INFOTECA-E), n. 75, p. 1–2, 2009. Disponível em: <http://www.cnpmf.embrapa.br/publicacoes/produto_em_foco/Citros_31.pdf>.

GOTTWALD, Tim R e GRAÇA, John V Da e BASSANEZI, Renato B. Citrus huanglongbing: the pathogen and its impact. Plant Health Progress, v. 8, n. 1, p. 31, 2007.

GYENGE, Javier E e TRUMPER, Eduardo V e EDELSTEIN, Julio D. Diseño de planes de muestro con niveles fijos de precisión del pulgón manchador de la alfalfa, Therioaphis trifolii Monell (Homoptera: Aphididae) en Alfalfa (Medicago sativa L.). An. Soc. Entomol. Brasil, n. 28, 1999.

JR., Paulo J. Ribeiro e CHRISTENSEN, Ole F. e DIGGLE, Peter J. Geostatistical software-geoR and geoRglm. Proceedings of DSC, p. 15, 2003. Disponível em: <http://leg.ufpr.br/~paulojus/pub/geoRDSC.pdf>.

LEAL, Renata Moreira; et al. Distribuição espacial de Huanglongbing (Greening) em citros utilizando a geoestatística. Revista Brasileira de Fruticultura, v. 32, n. 3, p. 808–818, 2010.

LIEBHOLD, Andrew M e ROSSI, Richard E e KEMP, William P. Geostatistics and geographic information systems in applied insect ecology. Annual review of entomology, v. 38, n. 1, p. 303–327, 1993.

MANTEL, Nathan. The detection of disease clustering and a generalized regression approach. Cancer research, v. 27, n. 2 Part 1, p. 209–220, 1967.

MATHERON, Georges. Principles of geostatistics. Economic geology, v. 58, n. 8, p. 1246–1266, 1963.

MØLLER, Jesper. Spatial statistics and computational methods. [S.l.]: Springer Science & Business Media, 2013. v. 173.

MONTEIRO, Antônio Miguel Vieira e colab. Análise espacial de dados geográficos. Planaltina: Embrapa, 2004.

NELDER J.A. AND WEDDERBURN R.W.M. Generalized Linear Models. Journal of the Royal Statistical Society, v. 135, n. 3, p. 370–384, 1972.

RIBEIRO JR, Paulo Justiniano; et al. Spatial pattern detection modeling of thrips (Thrips tabaci) on onion fields. Scientia Agricola, v. 66, n. 1, p. 90–99, 2009.

RIBEIRO, A M de A e LUNARDI, D M C. A precipitação mensal provável para Londrina-PR, através da função gama. Energia na agricultura, v. 12, n. 4, p. 37–44, 1997.

RIBEIRO, Patricia Pedroso Estevam e DE PAIVA, Maria Stela Veludo e JORGE, L A de C. Application of texture analysis for differentiation of the greening from others pests. 2014, [S.l.]: In: WORKSHOP DE VISÃO COMPUTACIONAL-WVC, 10., 2014, Uberlândia. Anais …, 2014. p. 202–207.

SCALON, João Domingos; et al. ANÁLISE DO PADRÃO ESPACIAL DA MORTALIDADE INFANTIL UTILIZANDO GEOESTATÍSTICA. Revista Univap, v. 18, n. 32, p. 149–160, 2012.

TEAM, R Core. R: A language and environment for statistical computing. 2013.

3. MODELAGEM ESPAÇO-TEMPORAL DA PRECIPITAÇÃO NO ESTADO DO PARANÁ

RESUMO

A modelagem espacial em dados pluviométricos permite a identificação espacial da probabilidade da ocorrência de chuva, uma informação importante em várias áreas da economia, especialmente na agricultura. Nesse contexto, é necessário a identificação de técnicas que permitam analisar dados cuja variável resposta apresente uma estrutura mais complexa, por exemplo, o fenômeno da superdispersão e estruturas de dependência espacial e temporal entre observações. Assim, foi proposto nesse trabalho comparar dois modelos, um sem a variável tempo e outro com o a inclusão de um efeito aleatório do tempo no preditor linear de dados para precipitação em 30 anos no estado do Paraná (mês de dezembro). O modelo que apresentou melhor ajuste quanto à precipitação foi o aquele que considerou o do efeito de tempo.

Palavras-chave: Distribuição de Poisson, Geoestatística, Superdispersão

ABSTRACT

Spatial modeling in rainfall data allows the spatial identification of the probability of rainfall, important information in several areas of the economy, especially in agriculture. In this context, it is necessary to identify techniques that allow analyzing data whose variable response presents a more complex structure, for example, the phenomenon of superdispersion and structures of spatial and temporal dependence between observations. Thus, it was proposed in this work to compare two models, one without the time variable and the other with the inclusion of a random effect of time in the linear predictor of data for precipitation in 30 years in the state of Paraná (December). The model that presented the best fit for the precipitation was the one that considered that of the time effect.

Keywords: Poisson distribuition, Geoestatistics, Overdispersion

3.1. Introdução

Na agricultura, a informação antecipada das condições locais referentes ao solo, radiação solar, precipitação pluvial e sua variação ao longo de um ciclo de cultivo, são importantes para obter bons resultados, visto que esses fatores são determinantes para o sucesso dos cultivos.

Os efeitos causados pela variabilidade pluviométrica provocam alterações quanto ao desenvolvimento da economia, desfavorecendo por exemplo setores como agricultura, pecuária e saneamento básico no que tange o abastecimento de água potável (Ribeiro e Lunardi, 1997).

O Paraná é um dos principais estados agrícolas da cultura de soja do Brasil em razão do grande volume IBGE (2017). Entretanto, o fato de estar localizado em uma região de transição climática, isto é, apresentar características de dois ou mais domínios morfoclimáticos (SIMÕES, 1954), provoca uma variabilidade em relação a produção ao longo dos anos, o que dificulta o planejamento da exploração agrícola.

A investigação dessa variação climática nesse estado pode ser realizada mediante estudos de amostragem georreferenciada, pois estas informações de correlação ou dependência espacial entre as localizações. O conhecimento correto acerca da distribuição espacial vinculada à precipitação é relevante

no planejamento agrícola, já que a instalação e desenvolvimento de culturas é altamente dependente dessa (CARAMORI et al.,1998).

Assim, estudar a relação existente entre o espaço e o tempo associado a incidência de chuva, pode auxiliar na obtenção de estimativas mais precisas dessa variável. Além do mais, é possível estabelecer se uma determinada espécie vegetal, terá condições satisfatórias, tanto econômica quanto fitotécnicas, para o seu cultivo. Nesse contexto, a análise da precipitação pluviométrica relacionada a quantidade de dias chuvosos em um determinado período, em diferentes estações, é de grande importância para o melhor entendimento climático e ambiental no estado. Para tal finalidade, o método utilizado para descrever a distribuição espacial é a geoestatística, o qual associa os graus de dependência com as medidas de distância e direção, entre os pontos amostrados, considerando que as observações foram provenientes de uma população com distribuição Gaussiana (DIGGLE e RIBEIRO, 2007).

Na prática, a terminologia geoestatística se caracteriza pela prática de krigagem, que por sua vez é um método de interpolação, o qual emprega a autocorrelação espacial entre observações vizinhas para obter previsões em locais não-amostrados (GELFAND et al., 2010). Uma escolha popular para estimar a dependência estatística é a família de funções de covariância Matérn que inclui, por exemplo, as funções de covariância exponencial e gaussiana como casos particulares (CRESSIE, 1992).

Como alternativa aos métodos que necessitam de transformação de variável, Nelder & Wedderburn (1972) propuseram os modelos lineares generalizados (MLGs), usualmente utilizados para dados de contagem por serem mais flexíveis em relação a distribuição da variável de interesse, sendo necessário que essa pertença a família exponencial sob a forma canônica. Porém, uma vez que os MLGs não consideram em sua estrutura a dependência espacial, é essencial a utilização dos modelos lineares generalizados geoestatísticos (MLGGs), visto que essa classe consegue captar a variação das medidas correlacionadas espacialmente (DIGGLE e RIBEIRO, 2007).

Nesse contexto, o presente trabalho tem por motivação de modelar o padrão espacial em estudos relacionados a dados pluviométricos de 30 anos amostrais para o estado do Paraná, mensurados na forma de contagens, que tem distribuição Poisson. Além disso, comparar dois modelos, um sem a covariável tempo e outro que incluí o efeito de tempo no preditor linear, contribuindo para a compreensão dos fatores que podem influenciar nas condições de plantios e manejo.

3.2. Material e Métodos

3.2.1. Dados Pluviométricos

Para esse estudo foram utilizados os dados de precipitação pluviométrica referente ao Estado do Paraná, obtidos por meio do site da Agência Nacional de Águas (ANA). A Figura 8 mostra as 78 estações meteorológicas, em que para cada um dos pontos têm-se as coordenadas geográficas (latitude e longitude).

A variável de interesse, classificada como resposta, é a quantidade de dias chuvosos no mês de dezembro, sendo avaliados os anos de 1985 a 2015, totalizando 30 anos amostrais.

O mês selecionado é considerado importante na cultura agrícola do referido estado, uma vez que para a plantação de soja se tem um período limite para o plantio, sendo recomendado pela Agência de Defesa Agropecuária do Paraná (ADAPAR) até o dia 31/12. O prazo regulamentado pelo órgão é devido à diminuição do risco de proliferação de ferrugem asiática, e também evitar uma possível falta de eficácia dos agrotóxicos usados para controlar a proliferação do fungo (NETO, 2018).

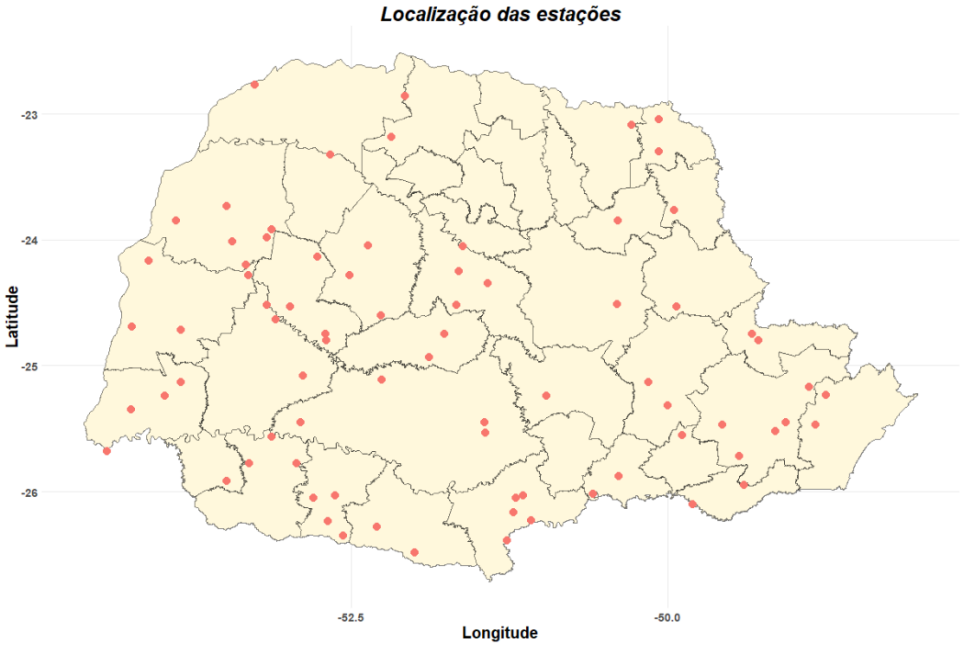

Figura 8. Estações Meteorológicas do Estado do Paraná.

3.2.1. Metodologia Estatística

Modelos Lineares Generalizados Mistos (MLGMs)

Proposto por Breslow e Clayton (1993), os MLGMs permitem uma maior flexibilidade em relação a variável resposta, e além disso, captam diversas fontes de variação e estruturas de correlações presentes no conjunto de dados, como por exemplo, o fenômeno da superdispersão, naturalmente encontrado em dados de contagens e que é definido por ser uma variabilidade muito maior do que a esperada pelo modelo especificado.

Assim, seja $\mathbf{Y}|\mathbf{u}$ uma variável aleatória com distribuição pertencente a família exponencial na forma canônica, em que Y é o vetor com *n* observações independentes, tem-se o preditor linear dado por

$$\boldsymbol{\eta_i = x_i^T \beta + z_i^T u} \quad \textbf{(14)}$$

em que $\boldsymbol{x_i^T}$ é a *i*-ésima linha da matriz de delineamento para os efeitos fixos, $\boldsymbol{\beta}$ é o vetor de efeitos fixos desconhecidos, $\boldsymbol{z_i^T}$ é a *i*-ésima linha da matriz de delineamento para os efeitos aleatórios, e $\boldsymbol{u}$ é o vetor de efeitos aleatórios, com $\boldsymbol{u \sim N(0, \Sigma)}$.

Considerando que os dados utilizados nesse trabalho estão na forma de contagem, é apropriado assumir que a variável resposta, quantidade de dias chuvosos, provém de uma distribuição Poisson, com função de ligação logarítmica expressa pela equação 15

$$\boldsymbol{Y_{ij} | \, u_i \sim Poisson\left(\lambda_{ij}\right)}$$

$$\boldsymbol{\log\left(\lambda_{ij}\right) = \exp\left(\beta_0 + z_i^T u\right)} \quad (15)$$

$$\boldsymbol{u_i \sim N\,(0, \Sigma)},$$

em que $\boldsymbol{z_i^T}$ é o efeito do *i*-ésimo tempo, e o modelo é caracterizado como Poisson-normal (Hinde e Demétrio, 1998). Para o modelo expresso em (15), foi proposto o preditor reduzido, o qual não contempla o efeito aleatório de tempo, ou seja, caracterizando o modelo de média.

A estimação do parâmetro de efeito fixo, e predição do efeito aleatório é por meio da máxima verossimilhança. No entanto, por não existir forma fechada, é necessário a utilização de métodos numéricos, sendo mais usual, o algoritmo da quadratura adaptativa de Gauss-Hermite, que está implementado no pacote lme4 (Bates et al., 2015).

Modelos Lineares Generalizados Geoestatísticos (MLGG)

Propostos por Ribeiro Jr e Diggle (2004), são classificados como uma extensão dos modelos mistos de Breslow e Clayton (1993), em que as medidas da variável resposta estão relacionadas espacialmente. Nesse contexto, considerando os dados apresentados na seção (3.2.1), tem-se o preditor linear definido pela equação (16)

$$Y_i | S(\boldsymbol{x}_i) \sim Poisson(\lambda(x_i))$$

$$\log\left(\lambda(x_i)\right) = \boldsymbol{Z}_{ij}^{T} \boldsymbol{u}_i + S(\boldsymbol{x}_i), \qquad i = 1, \cdots, n, \quad (16)$$

$$\boldsymbol{u_i \sim N\,(0, \Sigma)},$$

sendo $S(\boldsymbol{x}_i)$ um processo estocástico estacionário gaussiano com distribuição $N_n(0, \mathbf{D})$, onde $\mathbf{D} = \sigma^2 C(\boldsymbol{h})$ e $C(\boldsymbol{h})$ é a distância euclidiana $\boldsymbol{h}$ entre duas localizações (Diggle et al.,1998) . Além disso, $\boldsymbol{Z}_{ij}^T$ é a matriz de incidência, $\boldsymbol{u}_i$ é a matriz dos parâmetros desconhecidos de efeito aleatório, tal que $\boldsymbol{u}_i \sim N(0, \Sigma)$, que neste caso, é considerado o tempo por se ter medidas repetidas. Para o modelo expresso em (16), foi proposto o preditor reduzido, o qual não contempla o efeito aleatório de tempo, ou seja, caracterizando o modelo de média.

Assim, seja o modelo Poisson com a variável resposta $\boldsymbol{Y}$ a quantidade de dias chuvosos, com média $\alpha_i exp\big(S(\boldsymbol{x}_i)\big)$, sendo α_i o parâmetro relacionado aos dias de chuva, que captura a variabilidade extra do modelo tem-se a densidade na forma

$$P[Y(\boldsymbol{x})|S(\boldsymbol{x}) = y(\boldsymbol{x}_i)] = \frac{e^{-\alpha_i e^{\boldsymbol{Zu}+S(x)}} \left(\alpha_i e^{\boldsymbol{Zu}+S(\boldsymbol{x})}\right)^{y(\boldsymbol{x}_i)}}{y(\boldsymbol{x}_i)!}, \tag{17}$$

e a função de verossimilhança descrita pela integral e $f(S, \phi)$ conjunta de S com o parâmetro ϕ é

$$L(Y_i|\mathrm{S}, \phi) = \int \frac{e^{-\alpha_i e^{\boldsymbol{Zu}+S(x)}} \left(\alpha_i e^{\boldsymbol{Zu}+S(\boldsymbol{x})}\right)^{y(x_i)}}{y(\boldsymbol{x}_i)!} \cdot \frac{1}{\sqrt{2\pi\sigma^2}\,|\boldsymbol{R}|}\, e^{\frac{-1}{2\sigma^2}(S(\boldsymbol{x}_i)-\boldsymbol{\mu}_i)^T \boldsymbol{R}(S(\boldsymbol{x}_i)-\boldsymbol{\mu}_i)}\, ds, \tag{18}$$

sendo $\boldsymbol{R}$ a matriz da correlação amostral com dimensionalidade igual ao número de observações. Contudo, a expressão dada em (18), por não ter solução analítica utiliza-se métodos numéricos para se obter os coeficientes estimados. Uma possível abordagem é o algoritmo da Cadeia de Markov via Monte Carlo (MCMC), o qual está implementado no pacote geoRglm (CHRISTENSEN e RIBEIRO, 2007).

Para validação do modelo escolhido foi utilizado o semivariograma (DIGGLE et al., 2007) e para a comparação dos ajustes foram construídos gráficos dos valores observados *versus* estimados, e posteriormente, obtidos os coeficientes de determinação. Todas as análises foram implementadas no *Software* R (TEAM, 2013).

3.3. Resultados e Discussão

Inicialmente, ao verificar a análise descritiva por meio do Box-Plot referente a quantidade de dias chuvosos ao longo dos anos, observou-se que houve um comportamento e variabilidade entre os períodos. Assim, o grau de dispersão dos valores em relação ao valor mediano, nos anos de maior quantidade de ocorrência de chuva, apresentou um desvio padrão relativamente mais disperso. Além disso, em 2015 algumas localizações tiveram quase 30 dias de chuva no mês de dezembro, apresentando valores maiores que o esperado, quando comparados aos demais (Figura 9).

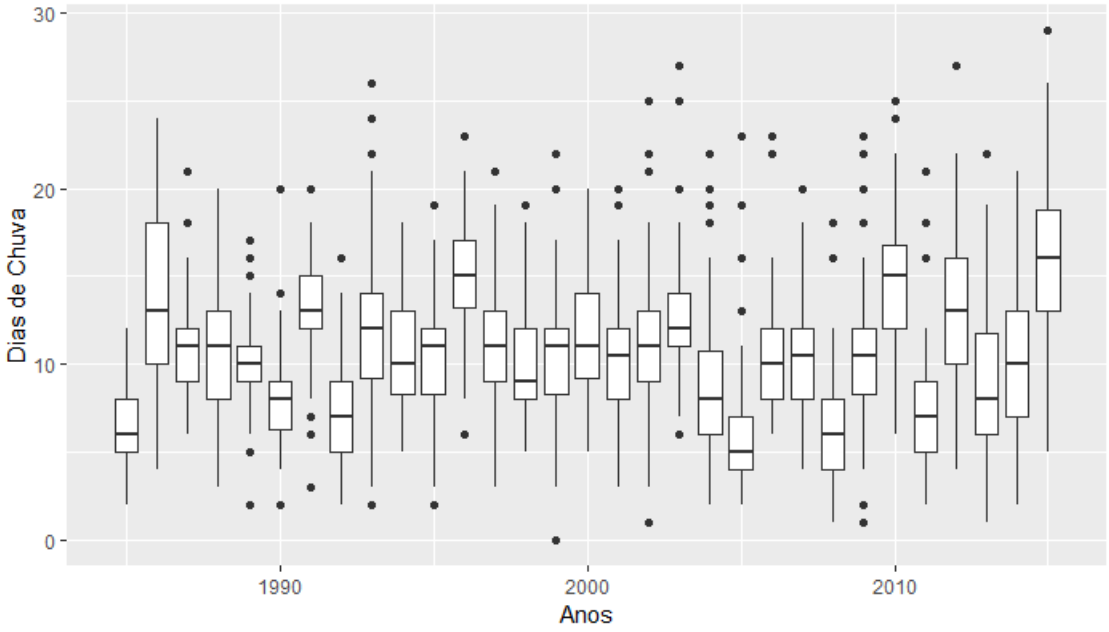

Figura 9. Box-Plot da quantidade de dias chuvosos por ano estudado.

A fim de verificar a existência, ou não, de um padrão aleatório espacial, foi realizado o teste de Mantel, no qual obteve-se p-valor < 0,05, ou seja, pode-se dizer ao nível de 5% de significância, que existem evidências de um padrão entre os pontos observados. Variogramas empíricos para todos os anos foram obtidos para representar a variância das diferenças entre os pares de pontos pelas distâncias correspondentes de cada par. Na Figura 10, alguns pontos estão fora do envelope simulado, computados de 500 simulações, rejeitando-se a hipótese de independência espacial entre as observações, o que corrobora o teste estatístico.

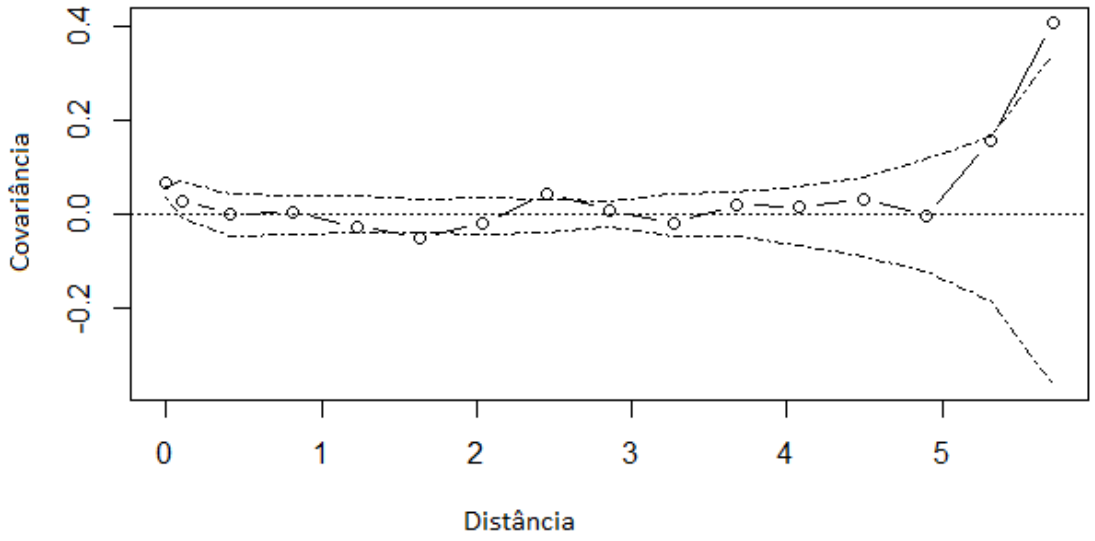

Figura 10. Covariograma da precipitação pluviométrica.

De acordo com Isaaks e Srivastava (1989), não é necessário que os dados sejam provenientes de uma distribuição normal para que seja utilizado a geoestatística, e sim que os mesmos apresentem uma distribuição com caudas não muito longas, visto que no processo de krigagem as estimativas serão influenciadas pelos valores médios. Dessa forma, têm-se os gráficos de densidade dos valores empíricos e da distribuição Poisson, para a variável resposta, os quais sugeriram a distribuição Poisson como a mais adequada, visto que quando comparando a distribuição empírica e teórica em relação a densidade, e a função de densidade acumulada, obteve-se um comportamento similar ao real (Figura 11).

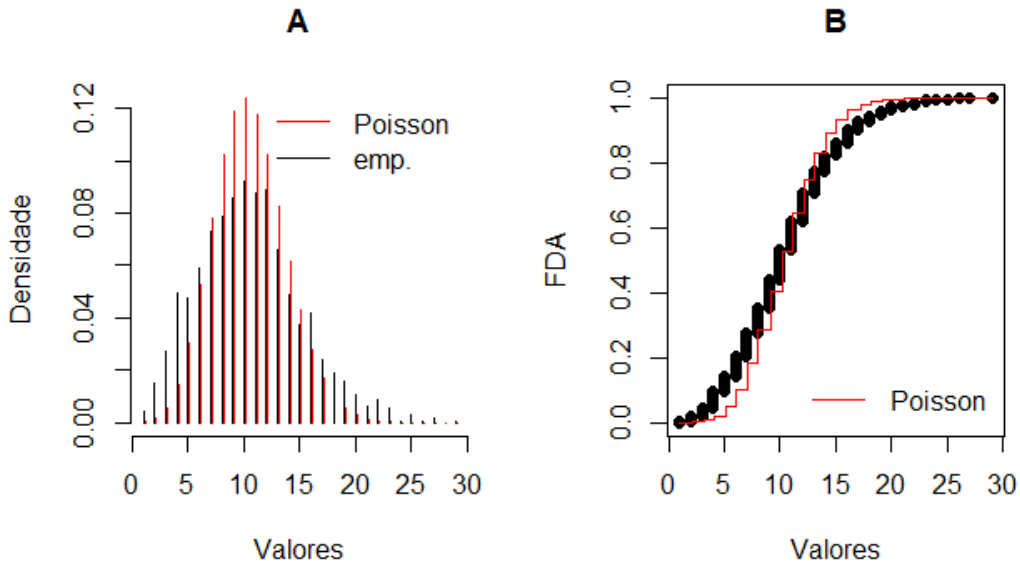

Figura 11. Gráficos do ajuste de uma distribuição Poisson empírica e teórica considerando A) Histograma de densidade; B) Função de densidade cumulativa.

De acordo com Diggle e Ribeiro (2007), a utilização do MLG padrão será suficiente para posterior modelagem espacial com dados de contagens. Assim, para estimar os parâmetros do modelo geoestatístico, foi dado o valor inicial em função de uma distribuição Poisson sem efeito aleatório de tempo ($\beta_0 = 10{,}5529$). Contudo, para o ajuste com efeito aleatório de tempo no preditor linear optou-se pelo uso do modelo linear generalizado misto para tentar ajustar o fenômeno da superdispersão, e conseguir captar as correlações entre as observações. Dessa forma, como descrito na subseção 3.2.1, ajustou-se o modelo expresso em (15), em que a estimativa de efeito fixo ($\beta_0 = 2{,}3290$) e o componente de variância predito ($\sigma_i^2 = 0{,}0572$) para o ajuste, são consistentes e confiáveis, visto que foram obtidos de um modelo bem ajustado (Figura 12).

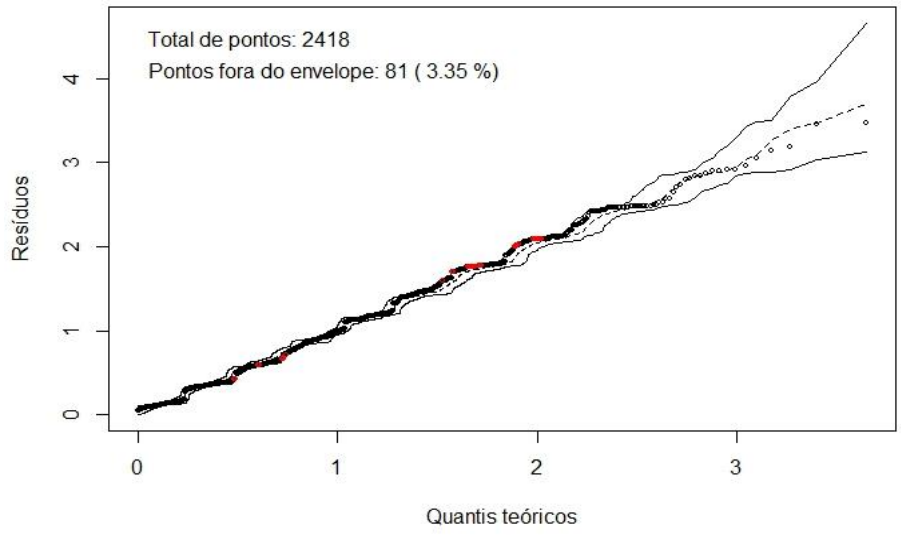

Figura 12. Gráfico semi-normal de probabilidade com envelope de simulação para o ajuste do modelo Poisson com efeito aleatório de ano.

Considerando os modelos com e sem efeito aleatório, obteve-se as estimativas dos parâmetros e a função do logaritmo da verossimilhança conforme mostra a Tabela 2. Ao ajustar o modelo considerando a pressuposição da média constante, tem-se o parâmetro da média geral associada a função de ligação ($\hat{\beta}$); a variância do processo espacial, que define a estrutura de covariância das observações ($\hat{\sigma}^2$); o erro de predição ou efeito pepita ($\hat{\tau}^2$), sendo esses três últimos associados à superfície de S(**x**); e o parâmetro da função de correlação ($\hat{\phi}$).

Fixando o valor *kappa* = 0,5 observou-se nas estimativas de $\hat{\sigma}^2$ e $\hat{\phi}$ indicativos que houve uma dependência espacial. De acordo com BONAT et al., (2011), valores elevados da proporção da variância (ω), proveniente de $\hat{\sigma}^2$ em relação a $(\hat{\sigma}^2 + \hat{\tau}^2)$, indicam dependência espacial quando os valores se aproximarem de 1. Assim, pode-se dizer que houve maiores dependências espaciais em ambos os modelos, e que consequentemente, apresentaram melhores predições (BONAT et al., 2011). Entretanto, ao verificar os valores obtidos da log-verossimilhança, pode-se dizer que o modelo de efeito aleatório, ajustado por um MLGM, se mostrou superior, uma vez que propiciou menores valores, a ponto de ser aconselhável a adição da covariável tempo. Vale ressaltar que o patamar $(\hat{\sigma}^2 + \hat{\tau}^2)$ representa o valor de estabilização do semivariograma. Nesse sentido, a distância limite entre a dependência e independência entre as amostras foi maior no modelo com efeito aleatório.

Tabela 2. Parâmetros do MLGG estimados via Cadeias de Markov para a quantidade de dias chuvosos.

Modelo	Estimativas						LV
	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$(\hat{\sigma}^2 + \hat{\tau}^2)$	$\hat{\omega}$	
Sem tempo	2,3424	0,0514	0,0555	0,0022	0,0536	0,9589	57,25
Comtempo	2,3306	0,0485	0,0663	0,0015	0,05	0,9700	47,93

Com a finalidade de verificar a eficiência da interpolação, foi feita análise de validação cruzada (Figura 13). Nota-se que no modelo apresentado na Figura 13B os pontos estão mais próximos da reta de regressão ajustada, ou seja, o ajuste com a inclusão de efeito aleatório no preditor linear, conseguiu equilibrar os dois lados da diagonal da reta, sendo este o mais indicado para a validação cruzada.

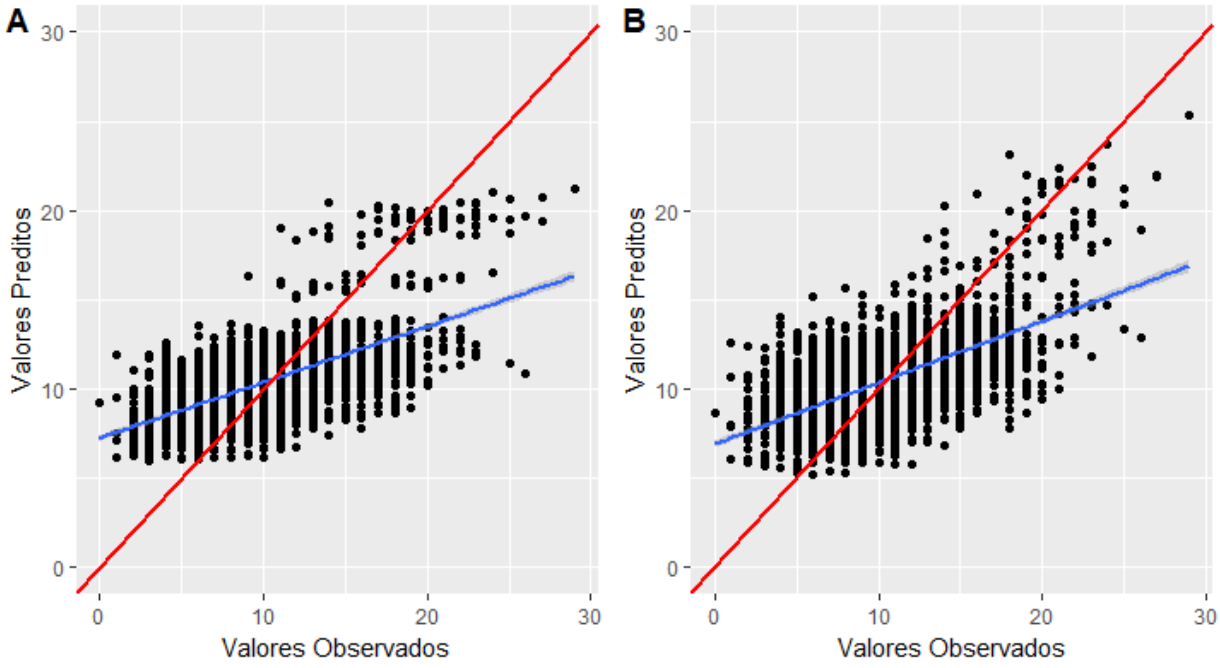

Figura 13. Gráficos da validação cruzada entre os valores amostrados e preditos (linha azul: reta de regressão ajustada, linha vermelha: reta de referência 0 e 1) para os modelos A) Sem efeito de tempo; e B) Com efeito de tempo.

Observa-se na da Figura 14 que os modelos apresentaram uma consistência quanto aos ajustes dos dados tanto para o modelo com efeito aleatório de tempo quanto sem esse efeito. Dessa forma, pode-se dizer que os valores preditos são semelhantes, havendo uma tendência no modelo sem o tempo superestimar os valores preditos em relação ao modelo com o tempo, conforme eles aumentam.

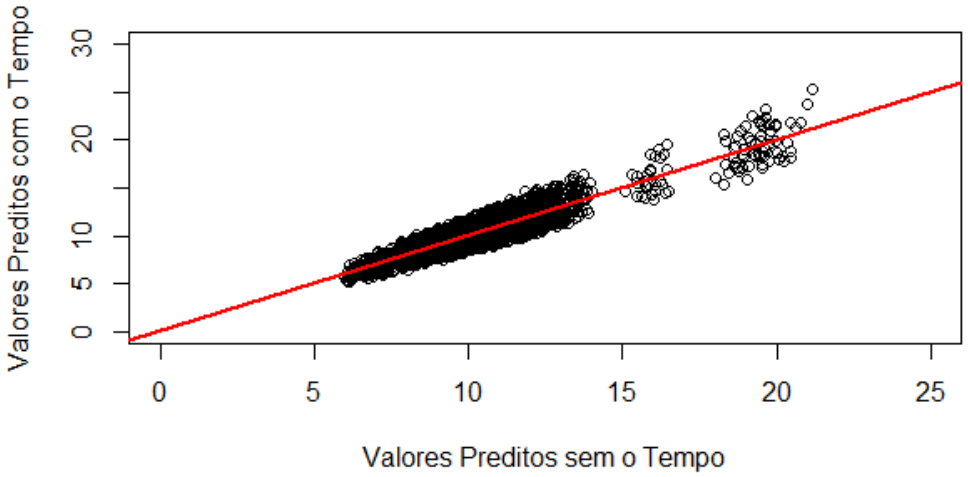

Figura 14. Gráfico dos valores preditos do MLGG com o tempo versus MLGG sem o tempo.

3.4. Conclusões

Foi constatado a robustez da krigagem em relação aos modelos propostos para dados de precipitação no estado do Paraná, considerando mês de dezembro durante 30 anos. As alterações no preditor linear evidenciaram diferenças nos resultados da validação. Desse modo, dentre os modelos propostos nesse trabalho, o que se ajustou de forma mais adequada aos dados de precipitação foi o que considerou efeito aleatório no preditor linear, uma vez que o ajuste conseguiu captar tanto a variação extra Poisson quanto a estrutura de correlação existente. Esse fato deve-se ao uso de valores iniciais mais confiáveis, haja visto que se utilizou o modelo linear generalizado misto.

A utilização do MLGG com a inclusão do tempo permite melhor acurácia na modelagem e consequentemente na interpolação de dados de precipitação no estado do Paraná, sendo assim a utilização deste modelo é benéfica a setores como a agricultura, seguros, hidrelétricas e outros que dependam da precipitação para o andamento de suas atividades. Considerando-se ainda que o efeito do tempo não é nulo, futuros estudos são necessários para mensurar o impacto deste, tendo em vista o cenário de mudanças climáticas em escala global.

Referências

BATES, Douglas; et al. **Package 'lme4'**. Convergence, v. 12, n. 1, 2015.

BONAT, Wagner Hugo e JUNIOR, Paulo Justiniano Ribeiro e ZEVIANI, Walmes Marques. **Comparando predições por modelos geoestatístico e aditivo generalizado para reconstituição de superfícies contínuas gaussianas**. Energia na Agricultura, v. 26, n. 2, 2011.

BRESLOW, Norman E e CLAYTON, David G. **Approximate inference in generalized linear mixed models**. Journal of the American statistical Association, v. 88, n. 421, p. 9–25, 1993.

CARAMORI, Paulo Henrique; et al. **Zoneamento da cultura do feijao no estado do Paraná.** . [S.l.]: Instituto Agronómico do Paraná, Paraná (Brasil)., 1998.

CHRISTENSEN, Ole F. e RIBEIRO, Paulo Jr. **The geoRglm Package**. 2007.

CRESSIE, Noel. **Statistics for spatial data**. Terra Nova, v. 4, n. 5, p. 613–617, 1992.

DIGGLE, Peter J. e RIBEIRO, Paulo Justiniano. **Model-based Geostatistics (Springer Series in Statistics)**. [S.l: s.n.], 2007.

DIGGLE, Peter J e TAWN, Jonathan A e MOYEED, R A. **Model-based geostatistics**. Journal of the Royal Statistical Society: Series C (Applied Statistics), v. 47, n. 3, p. 299–350, 1998.

GELFAND, Alan E; et al. **Handbook of spatial statistics**. [S.l.]: CRC press, 2010.

HINDE, John e DEMÉTRIO, Clarice G B. **Overdispersion: models and estimation**. Computational statistics & data analysis, v. 27, n. 2, p. 151–170, 1998.

IBGE. **Levantamento sistemático da produção agrícola–março 2017**. . [S.l.]: IBGE Rio de Janeiro. , 2017

ISAAKS, Edward H e SRIVASTAVA, R Mohan. **An introduction to applied geostatistics. 1989**. New York, USA: Oxford University Press. Jones DR, A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, v. 23, p. 345–383, 1989.

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

NETO, NELSON NIERO. No Title. Disponível em: <https://revistagloborural.globo.com/Noticias/Agricultura/Soja/noticia/2018/10/limite-do-plantio-da-soja-no-parana-volta-discussao-e-gera-polemica.html>. Acesso em: 30 jan 1BC.

SIMÕES, Ruth Mattos Almeida. **Notas sobre o clima do Estado do Paraná**. Revista Brasileira de Geografia. Rio de Janeiro, v. 16, n. 1, p. 126–132, 1954.

TEAM, R Core. **R: A language and environment for statistical computing**. 2013.

4 CONSIDERAÇÕES FINAIS

Modelos que exigem estruturas mais complexas devido a natureza dos dados, ainda são os entraves para a área de geoestatística, principalmente no que tange os procedimentos de inferência. O trabalho apresentado realizou aplicações considerando observações mensuradas na forma de contagens, sendo utilizada na ánalise a distribuição Poisson por ser a mais indicada para essa estrutura. Assim, a utilização dos MLGs para a obtenção das estimativas para os valores iniciais no processo de inferência geoestatístico, foi de grande valia, pois o ajuste considerando a relação média-variância da distribuição propicia melhores resultados, visto que são valores basais mais precisos.

Os resultados obtidos no capítulo 2, indicaram que as plantas cítricas presentes em áreas não comerciais, são fontes de propagação do *greening,* ao contrário dos pomares onde há controle químico da doença. Assim, ressalta-se a necessidade de atuação diversificada nas áreas não-comerciais, com o intuito de propor uma estratégia para diminuir a contaminação primária do *greening.*

No capítulo 3 foram propostas comparações dos preditores lineares, os quais consideraram a covariável de efeito aleatório tempo e sem esse efeito para dados de precipitações pluviométricas. Para o modelo cujo preditor linear não foi adicionado o efeito aleatório, fez o uso do MLG padrão. Por outro lado, no ajuste com o preditor linear que teve a adição da covariável utilizou-se a estimativa e o componente de variância como valores iniciais, oriundos do MLGM, que considera a superdispersão e estrutura de correlação. Dessa forma, verificou-se que a inclusão do efeito aleatório de tempo proporcionou melhores resultados quanto ao ajuste dos dados pluviométricos, sendo este modelo o mais indicado.

Recomenda-se para estudos futuros considerar outras distribuições para a variável de interesse nos MLGGs, haja visto que só está implementado no pacote utilizado (geoRglm), as distribuições Binomial e Poisson. Além disso, estruturas de variâncias e covariâncias poderão ser adicionadas para entender possíveis dependências espaço temporal. Pretende-se também, a utilização de métodos numéricos mais robustos, principalmente ao se ter uma grande quantidade de dados, pois apesar de ter utilizado um conjunto amostral pouco expressivo em relação a quantidade, têve-se problemas computacionais, principalmente, relacionadas a otimização do tempo de processamento.