

UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

**MODELOS LINEARES GENERALIZADOS E PROCESSOS PONTUAIS EM
ANÁLISE ESPACIAL DE DADOS AGRÍCOLAS**

DANIELA TRENTIN NAVA

CASCADEL - PR
FEVEREIRO - 2018

DANIELA TRENTIN NAVA

**MODELOS LINEARES GENERALIZADOS E PROCESSOS PONTUAIS EM ANÁLISE
ESPACIAL DE DADOS AGRÍCOLAS**

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola da Universidade Estadual do Oeste do Paraná, em cumprimento parcial aos requisitos para a obtenção do título de doutor em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais.

Orientador: Prof. Dr. Miguel Angel
Uribe-Opazo

Co-orientador: Profa. Dra. Fernanda De
Bastiani

Profa. Dra. Orietta Nicolis

CASCADEL - PR
FEVEREIRO - 2018

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Sistema de Bibliotecas – UNIOESTE)

N236m Nava, Daniela Trentin.
Modelos lineares generalizados e processos pontuais em análise espacial de dados agrícolas / Daniela Trentin Nava. --- Cascavel (PR), 2018.
ix, 73 f.: il.

Orientador: Prof. Dr. Miguel Angel Uribe-Opazo
Coorientadora: Profa. Dra. Fernanda De Bastiani
Coorientadora: Profa. Dra. Orietta Nicolis
Tese (Doutorado em Engenharia Agrícola) – Universidade Estadual do Oeste do Paraná, Campus de Cascavel, 2018. Centro de Ciências Exatas e Tecnológicas, Programa de Pós-Graduação em Engenharia Agrícola.
Inclui Bibliografia

1. Análise espacial (Estatística). 2. Distribuição binomial. 3. Função K direcional. I. Uribe-Opazo, Miguel Angel. II. De Bastiani, Fernanda. III. Nicolis, Orietta. IV. Universidade Estadual do Oeste do Paraná. V. Título.

CDD 20.ed. 526.1

Rosângela A. A. Silva – CRB 9º/1810

DANIELA TRENTIN NAVA

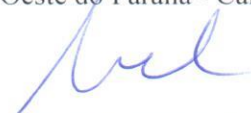
Modelos lineares generalizados e processos pontuais em análise espacial de dados agrícolas

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutora em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais, linha de pesquisa Geoprocessamento, Estatística Espacial e Agricultura de Precisão, APROVADO(A) pela seguinte banca examinadora:



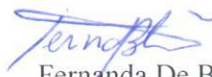
Orientador(a) - Miguel Angel Uribe Opazo

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)



Manuel Jesus Galea Rojas

Pontificia Universidade Católica do Chile (PUC-Chile)



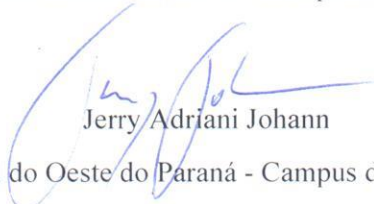
Fernanda De Bastiani

Universidade Federal de Pernambuco (UFPE)



Luciana Pagliosa Carvalho Guedes

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)



Jerry Adriani Johann

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Cascavel, 2 de fevereiro de 2018

Dedico a meus pais, com todo o meu amor.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, minha fortaleza e fonte de toda inspiração;

Ao meus pais, que dos bastidores fizeram este doutorado comigo. Agradeço por todas as vezes que deixaram sua casa para cuidar de mim, pelas palavras de incentivo que em todos os telefonemas me deram, amo demais vocês!

Ao meu orientador, professor Dr. Miguel Angel Uribe-Opazo, pela confiança e conhecimento que sempre me ofereceu; As minhas queridas co-orientadoras professoras Dra. Fernanda De Bastiani e Dra. Orietta Nicolis, por terem me ajudado a construir o conhecimento necessário, pela paciência com minhas inúmeras perguntas, e pela amável amizade. Ao professor Dr. Manuel Galea por toda a ajuda. Foi uma grande honra trabalhar com essa equipe;

Aos meus colegas da Universidade Tecnológica Federal do Paraná - campus Toledo, por terem permitido minha licença e por todo o suporte que me proporcionaram;

Ao PGEAGRI, por acreditar que eu pudesse finalizar com sucesso este doutorado; Ao CNPq, pela bolsa de financiamento para cursar doutorado sanduíche no Chile durante quatro meses; Ao senhor Agassis Linhares Neto, pela parceria na implantação do experimento em sua propriedade;

Agradeço a minha amada família, que soube entender minha ausência durante todos esses anos, em especial a minha sobrinha que me deu muito mais amor que ela pode imaginar!

Aos meus queridos amigos do LEE, que se tornaram parte de minha família: Regiane, Denise, Elisabeth, Fabiana, Gustavo, Rosângela e Marcio. Nós compartilhamos importantes momentos que sempre terei comigo;

Finalmente, mas não menos importante, a todos os amigos que sempre estiveram me lembrando que algum dia toda a dedicação e esforço valeriam a pena. Este dia está chegando, eu já posso senti-lo!

RESUMO

MODELOS LINEARES GENERALIZADOS E PROCESSOS PONTUAIS EM ANÁLISE ESPACIAL DE DADOS AGRÍCOLAS

O objetivo deste trabalho foi discutir distribuições discretas espaciais utilizando pontos de vista distintos, a saber, processos pontuais espaciais e distribuição binomial para dados espacialmente correlacionados. Os dados utilizados são provenientes de um experimento agrícola implantado em uma área comercial agrícola no município de Cascavel, estado do Paraná, cultivada com a cultura do milho. Subdividiu-se a área experimental em 40 parcelas georeferenciadas e observou-se o número de plantas atacadas pela lagarta do cartucho, do total de plantas de cada parcela. Para tal, assumiu-se que os dados possuem distribuição binomial. Propôs-se um estudo de análise de influência local de primeira ordem com o interesse em verificar possíveis pontos influentes. Os resultados obtidos sugerem que a presença de observações influentes nos dados modificam a inferência estatística, os valores preditos e os respectivos mapas. Em um segundo estudo, que teve como interesse a distribuição espacial da lagarta do cartucho na área experimental, utilizou-se de ferramentas de estatística espacial pontual. Para tal, cada planta infectada pelo inseto dentro da área experimental foi considerada como um evento de interesse. Realizou-se um estudo de anisotropia a partir de diferentes técnicas de processos pontuais, como K direcional e teste de ondaletas. Os resultados mostraram que a distribuição espacial da lagarta segue um processo pontual de Poisson agrupado com evidente anisotropia principalmente devido à forma da área experimental.

Palavras—chave: análise de influência; distribuição binomial para dados espacialmente correlacionados; função K direcional; processos de Poisson; teste de ondaletas.

ABSTRACT

GENERALIZED LINEAR MODELS AND POINT PROCESSES IN SPATIAL ANALYSIS OF AGRICULTURAL DATA

This thesis aimed at studying spatial discrete distributions based on two different points of view, that are, spatial point processes and spatial correlated binomial distribution. The data set came from an experiment setted in an agricultural commercial area in Cascavel city Paraná State, cropped with corn. The experimental area was subdivided into 40 georeferenced patch of land and the number of plants infected by *Spodoptera frugiperda* was observed within each patch of land. Thus, it is assumed that the data set have a binomial distribution. A study of first order local influence was proposed in order to verify possible influential points. The results suggest that the presence of influential observations in the data set have changed the statistical inference, the predicted values and the respective maps. In a second study, our interest was the spatial distribution of the fall armyworm in the experimental area. In order to do that, we used spatial point processes, where each plant infected by the insect within the experimental area was considered as an event of interest. An anisotropy study was carried out using different point process techniques, such as K directional function and wavelet test. The results show that the spatial distribution of the fall armyworm follow a Poisson cluster process with an evident anisotropy, mainly due to the shape of the experimental area.

Keywords: Influence analysis; K directional; Poisson processes; spatial correlated binomial distribution; wavelet test.

SUMÁRIO

LISTA DE TABELAS	x
LISTA DE FIGURAS	xi
1 INTRODUÇÃO	1
2 OBJETIVOS	3
2.1 Objetivo Geral	3
2.2 Objetivos específicos	3
3 REVISÃO BIBLIOGRÁFICA	4
3.1 Dados de superfícies aleatórias	4
3.1.1 Variável regionalizada	4
3.1.2 O semivariograma	5
3.2 Modelos lineares generalizados para dados dependentes	7
3.3 Dados de processos pontuais	9
3.3.1 Pressupostos para análise de processos pontuais espaciais	10
3.3.1.1 Completa aleatoriedade espacial	10
3.3.1.2 Estacionariedade	10
3.3.1.3 Isotropia	11
3.3.1.4 Propriedade de primeira ordem	12
3.3.1.5 Propriedade de segunda ordem	12
3.4 Processos de Poisson	12
3.4.1 Processo de Poisson Homogêneo	13
3.4.2 Estimção de primeira ordem	14
3.4.2.1 Estimção	14
3.4.3 Estimção de segunda ordem	15
3.4.3.1 A função K homogênea	16
3.4.4 Teste para CAE	18
3.5 Referências	19
4 ARTIGOS	23
4.1 ARTIGO 1: Influência local para dados binomiais espacialmente correlacionados: Uma aplicação da infestação de <i>Spodoptera frugiperda</i> no milho	23
4.1.1 Introdução	23

4.1.2	Modelos binomiais com correlação espacial	25
4.1.2.1	Básico em Geoestatística	26
4.1.2.2	Equações de estimação de quase-verossimilhança	27
4.1.2.3	Critério de seleção para a estrutura de correlação de trabalho	29
4.1.3	Influência local para observações em predição espacial	30
4.1.3.1	Predição	31
4.1.4	Aplicação ao conjunto de dados reais e resultados	32
4.1.5	Simulações	38
4.1.6	Conclusão	39
4.1.7	Apêndice A	40
4.1.8	Apêndice B	41
4.1.9	Referências	42
4.2	ARTIGO 2: Métodos estatísticos para identificar anisotropia na distribuição espacial da <i>Spodoptera frugiperda</i>	45
4.2.1	Introdução	45
4.2.2	Materiais e métodos	47
4.2.3	Resultados	53
4.2.4	Discussões	59
4.2.5	Agradecimentos	60
4.2.6	Referências	60
5	CONSIDERAÇÕES FINAIS	63
6	ANEXOS	64
6.1	Normas do Journal of Agricultural, Biological and Environmental Statistics	64
6.2	Normas do Spanish Journal of Agricultural Research	67

LISTA DE TABELAS

Tabela 1	Estatísticas descritivas para os dados de <i>Spodoptera frugiperda</i>	33
Tabela 2	Estimativas para o vetor de parâmetros α considerando-se todos os dados.	34
Tabela 3	Parâmetros estimados $\hat{\beta}$ e respectivos desvios padrões em parêntesis para o modelo considerando-se x_{i1} , x_{i2} and x_{i3} como covariáveis.	34
Tabela 4	Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} and x_{i3} como covariáveis.	35
Tabela 5	Estimativas para o vetor de parâmetros α sem as observações #1, #2, #27 e #32.	36
Tabela 6	Parâmetros estimados $\hat{\beta}$ e, em parênteses, respectivos desvios padrões para o modelo considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem os pontos #1, #2, #27 e #32.	36
Tabela 7	Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem os pontos #1, #2, #27 e #32.	36
Tabela 8	Estimativas do vetor de parâmetros α sem o ponto #1.	37
Tabela 9	Parâmetros estimados $\hat{\beta}$ e, em parênteses, respectivos desvios padrões para o modelo considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem o ponto #1.	37
Tabela 10	Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem o ponto #1.	37
Tabela 11	Percentual de vezes que a correta observação influente foi identificada sobre 1000 réplicas para o grid dos dados e para o grade regular.	39
Tabela 12	Dados da <i>Spodoptera frugiperda</i>	41
Tabela 13	Resultados do teste de contagem de quadrados, considerando-se diferentes quantidades de quadrados	59

LISTA DE FIGURAS

Figura 1	Simulação dos três tipos de padrões pontuais CAE (a), cluster (b) e regular (c), respectivamente.	15
Figura 2	Localização da área de estudo.	32
Figura 3	Semivariograma experimental.	33
Figura 4	Gráficos de influência B_i (a) e $ l_{max} $ (b), do esquema de perturbação para o modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$ e estrutura de correlação Matérn com $\kappa = 2, 0$	35
Figura 5	Mapas preditos para todo o conjunto de dados (a), removendo-se os pontos #1, #2, #27 e #32 (b), e removendo-se o ponto #1 (c).	38
Figura 6	Área experimental em Cascavel, estado do Paraná (sul do Brasil).	47
Figura 7	Imagem de satélite da área experimental com o polígono da área de estudo (a), localizações das plantas de milho (b), e localização das plantas infectas pela <i>Spodoptera frugiperda</i> (c).	48
Figura 8	Simulações de processos pontuais espaciais de Poisson aleatório (a), regular (b) e cluster (c). Função de intensidade estimada para os padrões simulados aleatório (d), regular (e) e cluster (f).	49
Figura 9	Função K de Ripley para padrão pontual aleatório (a), regular (b) e cluster (c). A linha pontilhada vermelha representa o caso teórico CAE, $K(r) = \pi r^2$ e a região cinza um envelope empírico de 95%. A linha contínua preta representa os dados observados.	51
Figura 10	Função de intensidade estimada usando kernel quártico, parâmetro de suavização $h = 2$ e uma grade regular de 400×400 pixels.	54
Figura 11	Histograma para os ângulos mais próximos para o primeiro vizinho (a) e terceiro vizinho (b). Neste gráfico, 0 corresponde a 0^0 ; 0,79 a 45^0 ; 1,57 a 90^0 ; 2,36 a 135^0 ; 5,50 a 315^0 e 6,28 a 360^0	54
Figura 12	Padrão pontual original (a), padrão pontual rotacionado 45^0 no sentido horário (b) e recorte da região central do padrão rotacionado (c).	55
Figura 13	Função K direcional para o padrão original (a), rotacionado (b) e recortado (c) com envelopes empíricos simulados de 95% sob a hipótese nula de CAE para cada um. A distância r no eixo x é dada em metros.	56
Figura 14	Escalograma direcional para o padrão original (a), padrão rotacionado (b) e padrão recortado (c) usando a ondaletas de Morlet anisotrópica . As transformações das ondaletas foram avaliadas em uma grade de dimensão 64×64 pixels para cada direção $\theta \in [0, 180]$ (eixo x), e $a = 1 \dots 30$ escalas (eixo y).	57
Figura 15	Teste baseado em ondaletas direcional. Valores da estatística $T(\theta_i), i = 1, \dots, 180$ para os dados de <i>Spodoptera frugiperda</i> (linha contínua) e intervalo	

empírico de 95% de confiança (linhas vermelhas) obtido pela simulação de 1000 padrões pontuais sob a hipótese de CAE, com $\lambda = 1303$ para o conjunto de dados originais (a) e rotacionados (b), e $\lambda = 557$ para o conjunto de dados recortado (c). 58

1 INTRODUÇÃO

Estatística espacial é um ramo da estatística que estuda métodos científicos para coleta, descrição, apresentação e análise de dados que possam ser modelados como processos estocásticos, em que o espaço índice é um conjunto de dimensão maior que um. A estatística espacial difere da clássica por incluir a localização espacial dos pontos no espaço e verificar propriedades e relacionamentos dos dados geográficos. Assim, é possível estudar o fenômeno de interesse ao longo do espaço com informações mais acuradas para a tomada de decisão.

As três formas mais conhecidas de dados espaciais são: dados de superfícies aleatórias, onde o que é considerado aleatório é a superfície do fenômeno de interesse; dados de área, geralmente visualizados em mapas, onde o espaço é particionado em áreas e cada área é diferenciada da outra por uso de polígonos; e dados de processos pontuais, em que o interesse está nas coordenadas geográficas, ou seja, a própria localização espacial do evento. Assim, o mecanismo gerador dos dados aleatórios é o que interessa neste tipo de dados. Para cada uma dessas classificações, a diferença é o tipo de dado aleatório, e, portanto, cada uma possui diferentes metodologias para seu estudo.

O conhecimento das características da distribuição espacial de um fenômeno natural ocorrido é importante no sentido de solucionar e evitar possíveis problemas futuros. Tais estudos têm recebido grandes avanços com o desenvolvimento dos sistemas de informações geográficas desenvolvidos atualmente. A ênfase da análise espacial é mensurar propriedades e relacionamentos por meio de modelagem, levando-se em consideração a localização espacial do fenômeno sob estudo. Para a realização deste trabalho de tese foram considerados dados de superfície aleatória e dados de processos pontuais.

A análise do dados de superfície aleatória foi obtida pelos ferramentais da geoestatística, os quais são capazes de realizar previsões do fenômeno sob estudo levando-se em consideração a continuidade espacial desse. Além disso, permitem incorporar possíveis direções privilegiadas de propagação do fenômeno na análise. Constitui-se portanto, uma alternativa de grande relevância para as mais diversas áreas do conhecimento. Em um estudo de estatística espacial pontual, o objetivo é investigar a distribuição espacial dos eventos ocorridos em uma região limitada do espaço, testando hipóteses sobre o padrão observado, ou seja, se esse está disposto ao longo da região de estudo de forma aleatória, regular ou agrupada. Casos de doenças em uma lavoura, tipos de ervas daninhas, distribuição de plantas, ataques de pragas são exemplos deste tipo de dados espaciais.

A cultura do milho é atacada por várias espécies de insetos desde a semeadura à colheita. Estes podem atacar raízes, colmos, folhas e espigas. Dentre as mais prejudiciais à cultura está a *Spodoptera frugiperda*, ou lagarta-do-cartucho, que ataca as plantas tanto na

fase vegetativa quanto na fase reprodutiva. No Brasil, pode causar prejuízos anuais estimados em mais de U\$ 400 milhões. A redução de produtividade causada pela praga pode atingir 60%, dependendo da cultivar e da época de ataque. O inseto pode ocorrer em todas as regiões produtoras, tanto nos cultivos de verão como nos de segunda safra. Especialmente na cultura de milho, os prejuízos contabilizados não estão relacionados à falta de tratamento fitossanitário, pois o número de aplicações de produtos químicos tem também aumentado e, em algumas regiões é comum uma média acima de cinco aplicações na safra. Mas, a grande preocupação no momento é o desenvolvimento de populações resistentes a produtos químicos, já verificado em algumas regiões, e a diminuição sensível da diversidade de agentes de controle biológico, em consequência do mau uso dos agrotóxicos. Portanto, é importante a adoção de um programa de manejo integrado, principalmente para que se restabeleça o equilíbrio ecológico dentro do sistema de produção.

Nas últimas décadas, com o avanço da tecnologia, muitos estudos foram realizados e contribuíram para o desenvolvimento, crescimento e aperfeiçoamento de novas técnicas para descrições e modelagem de dados de superfície aleatória de distribuições discretas e processos pontuais. Estudos de estatística espacial foram especialmente aplicados e aprimorados para as áreas de ecologia, problemas de cunho social, como criminologia e saúde pública bem como em biologia. Embora encontrem-se diversos exemplos de aplicação, pouco se tem visto para a área de Engenharia Agrícola, e esta foi a motivação para o desenvolvimento deste trabalho de tese.

Este trabalho de tese está organizado em forma de artigos como segue. Na seção 2, apresentam-se os objetivos deste trabalho de tese. Na seção 3, é realizada uma revisão bibliográfica dos principais assuntos utilizados na elaboração dos artigos. A seção 4 consta de dois artigos, resultados da tese. As considerações finais são apresentadas na seção 5 e a seção de anexos 6 contém as regras das revistas utilizadas para a escrita dos artigos.

2 OBJETIVOS

2.1 Objetivo Geral

Estudar modelos lineares generalizados para distribuição discreta espacial e processos pontuais considerando-se os dados agrícolas em estudo.

2.2 Objetivos específicos

- (i) Desenvolver metodologia para análise de influência local de primeira ordem para dados binomiais espacialmente correlacionados, utilizando modelos lineares generalizados e geoestatística;
- (ii) Aplicar a metodologia desenvolvida a um conjunto de dados reais agrícola;
- (iii) Realizar simulações para avaliar a metodologia desenvolvida;
- (iv) Estudar características de processos pontuais espaciais aplicados aos dados experimentais coletados da lagarta do cartucho na cultura do milho;
- (v) Realizar estudo de anisotropia aos dados da lagarta do cartucho usando teste baseado em ondaletas e função K direcional;

3 REVISÃO BIBLIOGRÁFICA

A estatística espacial preocupa-se com dados observados no espaço \mathbb{R}^d , $d \geq 2$ em que as observações podem ser medidas quantitativas coletadas em várias localizações ou simplesmente, um fenômeno de interesse nas próprias localizações espaciais. Nota-se, portanto, duas subdivisões para processos espaciais: uma que corresponde aos processos espaciais quantitativos, que utilizam observações espaciais quantitativas e os processos pontuais espaciais em que a ocorrência de cada evento é o interesse de estudo.

3.1 Dados de superfícies aleatórias

A Geoestatística é um braço da estatística espacial que se ocupa do estudo de um conjunto finito de dados espaciais relacionados a um fenômeno de interesse (DIGGLE & RIBEIRO JR, 2007), cujo início foi nas minas de carvão da África do Sul a partir do trabalho do engenheiro de minas Daniel G. Krige e do estatístico H.S. Sichel e posteriormente formalizado por G. Matheron (MATHERON, 1963). Neste enfoque, a superfície $Y(\mathbf{s})$, $\mathbf{s} \in \mathbb{R}^2$ do fenômeno sob estudo é considerada aleatória.

3.1.1 Variável regionalizada

A maioria dos fenômenos naturais apresenta continuidade espacial, uma vez que dois pares de pontos próximos tendem a ter valores mais parecidos do que dois que estejam distantes um do outro (ISAAKS & SRIVASTAVA, 1989). A variável de interesse está associada ao espaço físico que ocupa. Assim, seja a variável regionalizada definida pela função $f : \mathbb{R}^n \rightarrow \mathbb{R}$, se o ponto $P \in \mathbb{R}^n$ se aproxima de $P_0 \in \mathbb{R}^n$, então $f(P)$ se aproxima de $f(P_0)$.

Dada a impossibilidade de se coletar em todos os valores da superfície aleatória ao longo de todo o espaço de interesse, a geostatística permite que se caracterize a variabilidade a partir de um subconjunto discreto desta. Para tal, definem-se n locais fixos de coleta $\mathbf{s}_1, \dots, \mathbf{s}_n$, em que $\mathbf{s}_i = (x_i, y_i)$ é a coordenada geográfica da posição \mathbf{s}_i . Em cada posição \mathbf{s}_i observa-se o valor da variável aleatória $Y(\mathbf{s}_i)$. Assim, a distribuição conjunta do vetor de variáveis aleatórias $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^T$ é induzida pela superfície aleatória $Y(\mathbf{s})$.

Para cada $Y(\mathbf{s}_i)$ define-se a média $\mu_i = E[Y(\mathbf{s}_i)]$, a variância $Var[Y(\mathbf{s}_i)] = E[Y(\mathbf{s}_i)^2] - E^2[Y(\mathbf{s}_i)]$, a covariância $C[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = E[Y(\mathbf{s}_i)Y(\mathbf{s}_j)] - E[Y(\mathbf{s}_i)]E[Y(\mathbf{s}_j)]$, a correlação $\rho[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = C[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] / (\sqrt{Var[Y(\mathbf{s}_i)]} \sqrt{Var[Y(\mathbf{s}_j)]})$ e a semivariância dada por

$$\gamma[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = \frac{1}{2} E[(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2],$$

para $i, j = 1, \dots, n$.

Usualmente é conhecida apenas uma única realização de cada variável aleatória que compõe o vetor n -variado \mathbf{Y} e para que seja possível realizar inferências, a teoria admite considerar que $\{Y(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^n\}$ é um processo estocástico estacionário (MARDIA & MARSHALL, 1984; DIGGLE & RIBEIRO JR, 2007) com as propriedades:

1. A média de todas as variáveis aleatórias é constante, ou seja

$$E[Y(\mathbf{s}_1)] = \dots = E[Y(\mathbf{s}_n)] = \mu.$$

2. A covariância entre duas variáveis aleatórias depende somente da distância espacial que as separa e é independente de sua localização,

$$C[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = C[Y(\mathbf{s}_i), Y(\mathbf{s}_i + h_{ij})] = C(h_{ij}),$$

em que h_{ij} é a distância euclidiana entre \mathbf{s}_i e \mathbf{s}_j . Logo, tem-se

$$\rho[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = \frac{C(h_{ij})}{C(0)} = \rho(h_{ij}),$$

que ao assumir a hipótese de estacionariedade, resulta em

$$\gamma(h_{ij}) = C(0) - C(h_{ij}) \quad \text{e} \quad \rho(h_{ij}) = 1 - \frac{\gamma(h_{ij})}{C(0)}.$$

3.1.2 O semivariograma

A quantificação da variabilidade espacial em geoestatística é feita pelo do semivariograma $\gamma(h_{ij})$. A literatura apresenta diversos estimadores para $\gamma(h_{ij})$ conforme pode ser visto em Matheron (1963), Cressie & Hawkins (1980) e Srivastava & Parker (1989), por exemplo.

O estimador clássico de Matheron (MATHERON, 1963) é calculado como

$$\hat{\gamma}(h_{ij}) = \frac{1}{2N(h_{ij})} \sum_{N(h_{ij})} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2,$$

em que $N(h_{ij})$ é o número de pares de valores medidos $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$, separados pela distância h_{ij} , onde $h_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ é a distância euclidiana entre as localizações \mathbf{s}_i e \mathbf{s}_j , e $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$ são os valores da i -ésima e j -ésima observações da variável regionalizada amostrada nas localizações \mathbf{s}_i e \mathbf{s}_j , $i, j = 1, \dots, n$.

No semivariograma, os parâmetros que definem a estrutura de dependência espacial são o efeito pepita, o alcance e o patamar. O efeito pepita, denotado por $\phi_1 \geq 0$, indica a descontinuidade do semivariograma para distâncias menores que a menor distância entre as amostras. Pode ser atribuído a erros de medição ou fontes de variação espacial a distâncias inferiores ao intervalo amostral ou ambos (CRESSIE, 1993). O alcance, denotado como $a = c(\phi_3)$, é uma função do parâmetro ϕ_3 que delimita a distância a partir da qual o valor da variável em uma localização não tem mais influência sobre a localização vizinha. O patamar, obtido como $\phi_1 + \phi_2$, representa o limite do variograma quando as distâncias desse tendem ao infinito.

Maiores detalhes veja Diggle & Ribeiro Jr (2007) e Uribe-Opazo et al. (2012).

Uma análise do semivariograma pode indicar presença de anisotropia, ou seja, comportamentos distintos da variável aleatória quando consideradas diferentes direções. Maiores informações sobre anisotropia podem ser verificadas em Isaaks & Srivastava (1989), Diggle & Ribeiro Jr (2007) e Guedes et al. (2013).

Uma vez obtido o semivariograma experimental, é necessário ajustar algum modelo que seja capaz de representar os parâmetros da estrutura de dependência espacial, e mais ainda, quantificar a continuidade espacial de $Y(\mathbf{s})$. Assume-se que $Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n)$ são dados pelo modelo

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \dots, n, \quad (3.1)$$

em que o termo determinístico $\mu(\mathbf{s}_i)$ e o termo estocástico $\epsilon(\mathbf{s}_i)$ dependem da localização espacial onde $Y(\mathbf{s}_i)$ foi obtida, sendo $\mu(\mathbf{s}_i) = \sum_{j=0}^p \beta_j x_j(\mathbf{s}_i)$, em que $x_j(\mathbf{s}_i)$ são funções conhecidas de \mathbf{s}_i e β_j parâmetros desconhecidos a serem estimados, e $E[\epsilon(\mathbf{s}_i)] = 0$. Isto implica que $E[Y(\mathbf{s}_i)] = \mu(\mathbf{s}_i)$. Em forma matricial, o modelo em (3.1) é tal que

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

em que $\mathbf{Y}(\mathbf{s}) = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^T$ e $\boldsymbol{\epsilon}(\mathbf{s}) = (\epsilon(\mathbf{s}_1), \dots, \epsilon(\mathbf{s}_n))^T$, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ com \mathbf{X} uma matriz $n \times p$ composta pelo vetor de uns e pelas p covariáveis e, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ o vetor de parâmetros desconhecidos a serem estimados.

Admite-se ainda que a dependência espacial é determinada por uma matriz escala $\boldsymbol{\Sigma}$ de ordem $n \times n$ simétrica, não singular e positiva definida proporcional a covariância do processo, cuja forma paramétrica é

$$\boldsymbol{\Sigma} = \phi_1 \mathbf{I} + \phi_2 \mathbf{R}(\phi_3),$$

em que ϕ_1, ϕ_2 e ϕ_3 são definidos anteriormente e $\mathbf{R}(\cdot)$ é matriz simétrica $n \times n$ com elementos na diagonal principal igual a 1 e que depende do modelo teórico de variância, adotado na modelagem do semivariograma experimental.

A literatura oferece duas classificações de modelos teóricos para explicar a variância espacial, os modelos transitivos e modelos não transitivos. Os modelos não transitivos são aqueles que não possuem patamar, ou seja, o crescimento de $\gamma(h_{ij})$ é contínuo em relação à distância h e satisfaz apenas a hipótese intrínseca, não apresentando variância finita. Em contrapartida, os modelos transitivos possuem patamar e indicam a ocorrência de estacionariedade de segunda ordem. Além disso, esses modelos consideram que a variável em estudo é isotrópica (MARDIA & MARSHALL, 1984; CRESSIE, 1993). Neste trabalho de tese considerou-se modelos transitivos apenas. Diversos são os modelos teóricos apresentados na literatura, ver por exemplo Isaaks & Srivastava (1989), Soares (2006), Diggle & Ribeiro Jr. (2007) e Webster & Oliver (2007).

Um modelo teórico transitivo flexível é o modelo da família Matérn (MATÉRN, 1986).

Define-se a função semivariância para a família Matérn como

$$\gamma(h_{ij}) = \begin{cases} 0, & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[\{1 - 2^{\kappa-1} \Gamma(\kappa)\}^{-1} \left(\frac{h_{ij}}{\phi_3}\right)^\kappa K_\kappa\left(\frac{h_{ij}}{\phi_3}\right) \right], & \text{if } h_{ij} > 0 \end{cases},$$

em que $\phi_j \geq 0$, $j = 1, 2, 3$ e o parâmetro de ordem $\kappa > 0$; $\Gamma(\kappa)$ é a função gama, e $K_\kappa(\cdot)$ é a função Bessel modificada de terceiro tipo de ordem κ . Se $\kappa = 0,5$, tem-se o modelo exponencial

$$\gamma(h_{ij}) = \begin{cases} 0 & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[1 - \exp\left(-\frac{h_{ij}}{\phi_3}\right) \right], & \text{if } 0 < h_{ij} \leq \phi_3 \\ \phi_1 + \phi_2, & \text{if } h_{ij} > \phi_3, \end{cases}$$

e se $\kappa \rightarrow \infty$, tem-se o modelo Gaussiano

$$\gamma(h_{ij}) = \begin{cases} 0, & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[1 - \exp\left[-\left(\frac{h_{ij}}{\phi_3}\right)^2\right] \right], & \text{if } 0 < h_{ij} \leq \phi_3 \\ \phi_1 + \phi_2, & \text{if } h_{ij} > \phi_3. \end{cases}$$

Para maiores detalhes, veja Diggle & Ribeiro Jr. (2007) e Webster & Oliver (2007).

3.2 Modelos lineares generalizados para dados dependentes

Os ferramentais geoestatísticos são muito úteis para análise e modelagem de dados espaciais e, aliados ao crescente desenvolvimento de softwares estatísticos, possibilitam que pesquisadores das mais diversas áreas do conhecimento como hidrologia, geologia, ciência do solo, agronomia e ecologia dentre outros, a utilizarem tais métodos (GBUR et al., 2012). Eles podem permitir maior compreensão do processo espacial em estudo, mas também podem apresentar situações em que a metodologia geoestatística atual é insuficiente para fornecer um conjunto abrangente de modelos e métodos. Um exemplo desta carência está na falta de estudos em análise e predição de dados espaciais discretos categóricos (GOTWAY & STROUP, 1997).

Uribe-Opazo et al. (2012) usaram técnicas de diagnóstico para dados com distribuição Gaussiana e para avaliar a sensibilidade dos estimadores de máxima verossimilhança, funções de covariância e preditores lineares quando existem pequenas perturbações nos dados e/ou nas suposições do modelo espacial linear Gaussiano. Para a distribuição t de Student, Assumpção et al. (2014) descreveram dois métodos para análise de diagnóstico de influência local para dados espacialmente correlacionados ao utilizarem modelos lineares espaciais n variados com distribuição t de Student com graus de liberdade fixos. De Bastiani et al. (2015) consideraram a família de distribuição elíptica para estimar a dependência espacial de dados georeferenciados e avaliaram a sensibilidade dos estimadores de máxima verossimilhança a partir de técnicas de influência local. Garcia-Papani et al. (2017) apresentaram modelagem e diagnósticos para dados assimétricos espaciais através da distribuição espacial Birnbaum-Saunders com aplicações a dados agrícolas. Em seus trabalhos, Prentice (1988)

e Albert & McShane (1995) utilizam equações de estimação generalizadas (EEGs) para dados binomiais espacialmente correlacionados. Gotway & Stroup (1997) propuseram uma metodologia para análise, estimação e predição de dados de variáveis espaciais categóricas e discretas.

Nelder & Wedderburn (1972) apresentam uma extensão do modelo normal linear, denominados modelos lineares generalizados (MLGs), os quais podem ser utilizados quando os dados são provenientes da família de distribuições exponenciais, da qual a distribuição normal é um caso particular (GOTWAY & SROUP, 1997). Além disso, permite uma relação não linear da média da variável resposta com as variáveis explicativas, por intermédio da função de ligação.

Um MLG consiste de três componentes: (1) um componente estocástico que define a distribuição de probabilidades ou a função de quase verossimilhança da variável resposta; (2) um preditor linear que é o componente sistemático que descreve o modelo linear definido pelas variáveis exploratórias; e (3) uma função de ligação que relaciona a média da variável resposta com uma combinação linear das variáveis explicativas (GBUR et al., 2012). Formalmente, sejam Y_1, \dots, Y_n variáveis aleatórias, não necessariamente identicamente distribuídas, cada uma com distribuição na família exponencial, e sejam y_1, \dots, y_n os correspondentes valores. A contribuição de algum y_i para o logaritmo da função verossimilhança pode ser escrita como

$$L(\theta_i, \phi_i; y_i) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i),$$

em que θ_i é o parâmetro natural e é uma função de $\mu_i = E(Y_i)$ e ϕ_i é parâmetro de escala. A variância de Y_i é $var(Y_i) = v(\mu_i)a(\phi_i)$, onde $v(\mu_i)$, a função variância, é o componente de $var(Y_i)$ que depende de μ_i .

O logaritmo da verossimilhança conjunta é

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{y}) = \mathbf{y}^\top \mathbf{A}^{-1} \boldsymbol{\theta} - (\mathbf{b}_\theta^{1/2})^\top \mathbf{A}^{-1} (\mathbf{b}_\theta^{1/2}) + \mathbf{1}^\top \mathbf{c},$$

em que \mathbf{y} é o vetor $n \times 1$ de dados observados, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^\top$ é o vetor de parâmetros naturais, $\mathbf{b}_\theta^{1/2} = (\sqrt{b(\theta_1)}, \dots, \sqrt{b(\theta_n)})^\top$, $\mathbf{A} = \text{diag}[a(\phi_i)]$ é uma matriz $n \times n$, \mathbf{c} é um vetor $n \times 1$ cujo i -ésimo elemento é $c(y_i; \phi_i)$, e $\mathbf{1}$ é vetor $n \times 1$ de uns.

Modelos lineares generalizados supõem que as observações não são necessariamente relacionadas linearmente com μ_i , um MLG é definido como

$$E(\mathbf{Y}) = \boldsymbol{\mu}$$

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

em que $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros, \mathbf{X} é uma matriz $n \times p$ de constantes conhecidas, e $var(\mathbf{Y}) = \mathbf{V}$, $g(\cdot)$ é a função monótona e diferenciável denominada função de ligação (PAULA, 2013). Em particular, se a função de ligação é igual à função que define o parâmetro canônico, esta é denominada função de ligação canônica. Ainda, se a variável resposta for binomial, tem-se $v(\mu_i) = \mu_i(1 - \mu_i)$ e $g(\mu_i) = \log(\mu_i/(1 - \mu_i))$.

Nelder e Wedderburn (1972) provaram que a estimativa de máxima verossimilhança de β pode ser obtida pelo método de scoring de Fisher como uma solução iterativa do sistema de equações

$$\mathbf{X}^\top \mathbf{W} \mathbf{X} \beta = \mathbf{X}^\top \mathbf{W} \mathbf{y}^*,$$

em que $\mathbf{W} = \mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D}$, $\mathbf{D} = \text{diag}[\partial \mu_i / \partial \eta_i]$ é uma matriz $n \times n$, $\mathbf{y}^* = \boldsymbol{\eta} + \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, e \mathbf{V} é a matriz covariância de trabalho de ordem $n \times n$ de \mathbf{Y} ,

$$\mathbf{V} = v_{\boldsymbol{\mu}}^{1/2} \mathbf{R}(\phi_3) v_{\boldsymbol{\mu}}^{1/2}$$

em que $\mathbf{R}(\phi_3)$ é a matriz de correlação de trabalho de ordem $n \times n$ que é completamente especificada pelo vetor de parâmetros de correlação α de ordem $a \times 1$ (ZEGGER & LIANG, 1986; GOTWAY & STROUP, 1997; PARDO & ALONSO, 2017). Existem diversos critérios para a escolha da melhor estrutura de correlação de trabalho propostos na literatura. Veja por exemplo, Pan (2001), Cui (2007), Gosho et al. (2011) e Pardo & Alonso (2017).

Resumidamente, MLG é uma extensão sistemática de teoria de modelos lineares para dados além da distribuição normal. Originalmente, os MLGs utilizam o método de máxima verossimilhança. Entretanto, Wedderburn (1974) e posteriormente extensivamente estudado por McCullagh (1983), conseguiram estender via método de quase verossimilhança, MLGs para dados em que é possível obter o primeiro e o segundo momento, mas a representação da distribuição conjunta não pode ser escrita explicitamente. Porém, a maioria dos métodos de quase verossimilhança baseiam-se na suposição de que as observações são independentes. Para dados espacialmente correlacionados, Liang & Zeger (1986) e Zeger & Liang (1986) apresentam EEGs baseadas em aproximação de quase verossimilhança. Para dados dependentes, a função de quase verossimilhança pode ser definida como

$$\frac{\partial Q(\boldsymbol{\mu}; \mathbf{y})}{\partial \boldsymbol{\mu}} = \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

em que $Q(\cdot)$ é a quase verossimilhança. Para maiores detalhes sobre a função $Q(\cdot)$ veja Nava et al. (2017). Diferenciando Q com respeito a β , a função escore pode ser escrita como

$$\mathbf{U} = \boldsymbol{\Delta}^\top \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

em que $\boldsymbol{\Delta}$ é uma matriz $n \times p$ com elementos $\partial \mu_i / \partial \beta_j$.

3.3 Dados de processos pontuais

Acredita-se que a origem dos processos pontuais surgiu com o desenvolvimento de problemas de contagem, que trata de números de eventos em intervalos ou regiões de vários tipos (GUAN, 2003). Dentre os processos pontuais, o processo de Poisson, apresentado em Poisson (1837), é um exemplo bastante simples mas de extrema importância. Com o avanço da tecnologia, os processos pontuais apresentaram grande desenvolvimento nas mais diversas áreas do conhecimento, e algumas delas podem ser observadas em Diggle (2003) e

Daley & Vere-Jones (2002).

Um processo pontual $N \subset \mathbb{R}^2$ é um processo estocástico em que cada realização consiste de um conjunto finito ou contável de pontos no plano. Esse conjunto de pontos é denominado padrão pontual e um ponto em particular é denominado evento. Isto é, são conjuntos de dados em que o principal interesse está no conjunto de coordenadas geográficas que representam as exatas localizações dos eventos, como por exemplo, o endereço de pessoas doentes, local de roubos de carros, localização de árvores em uma região, locais de ocorrência de ervas daninhas (CRESSIE, 1993; CÂMARA & CARVALHO, 2005).

Em um padrão pontual, cada evento representa a existência do fenômeno do qual se está interessado em sua localização particular em uma dada região. Em outras palavras, um padrão pontual é uma coleção de eventos em alguma área ou conjunto de um fenômeno de interesse, e é tipicamente interpretado como uma amostra ou uma realização de um processo pontual (ILLIAN et al., 2008).

Denota-se um padrão pontual de $N(\mathbf{s})$ eventos como o conjunto de localizações $\mathbf{s}_1, \dots, \mathbf{s}_n$, em que cada evento/ponto é tal que $\mathbf{s}_i = (x_i, y_i)$. O padrão pontual que ocorre na região de estudo A possui uma área $|A|$.

3.3.1 Pressupostos para análise de processos pontuais espaciais

3.3.1.1 Completa aleatoriedade espacial

Assim como na estatística clássica, em que a distribuição normal de probabilidades representa a distribuição teórica de referência, em processos pontuais também existe a necessidade de uma distribuição pontual que possa servir como referência para o desenvolvimento de métodos formais capazes de averiguar a significância dos resultados exploratórios (DIGGLE, 2003).

O modelo teórico mais simples, conhecido como completa aleatoriedade espacial (CAE), corresponde ao caso em que é igualmente provável que um evento aconteça em qualquer local dentro da área de estudo. No modelo CAE, os eventos em cada sub-região são não correlacionados e estão associados à mesma distribuição de probabilidade de Poisson. Em outras palavras, a posição de cada evento é independente e os eventos têm igual probabilidade de ocorrência em toda a região A .

3.3.1.2 Estacionariedade

Dizer que um processo pontual N é estacionário, significa que todas as propriedades do processo são invariantes sob translação (DIGGLE, 2003). Suponha que $N(A)$ sejam variáveis aleatórias que representam o número de eventos em uma região $A \in \mathbb{R}^2$,

$$N(A) = \#(\mathbf{s}_i \in A),$$

em que $\#$ significa 'o número de'. O processo é dito ser estacionário se, para qualquer inteiro k e regiões disjuntas $A_i, i = 1, \dots, k$, a distribuição conjunta de $N(A_1), \dots, N(A_k)$ é invariante

à translação por uma quantidade arbitrária \mathbf{v} ,

$$N \stackrel{d}{=} N_{\mathbf{v}}, \quad (3.2)$$

em que $\stackrel{d}{=}$ significa tem a mesma distribuição, e $N_{\mathbf{v}}$ é o processo pontual resultando de uma mudança de todos os pontos de N pelo mesmo vetor \mathbf{v} . Se $N = \{\mathbf{s}_1, \mathbf{s}_2, \dots\}$, então $N_{\mathbf{v}} = \{\mathbf{s}_1 + \mathbf{v}, \mathbf{s}_2 + \mathbf{v}, \dots\}$. A expressão ‘tem a mesma distribuição’, significa

$$\begin{aligned} P[N(A_1) = n_1, \dots, N(A_k) = n_k] &= P[N_{\mathbf{v}}(A_1) = n_1, \dots, N_{\mathbf{v}}(A_k) = n_k] \\ &= P[N(A_1 - \mathbf{v}) = n_1, \dots, N(A_k - \mathbf{v}) = n_k], \end{aligned}$$

em que a igualdade da segunda linha se deve ao fato de que $N_{\mathbf{v}}(A) = N(A - \mathbf{v})$ para todo A e \mathbf{v} , onde $A - \mathbf{v} = \{\mathbf{s} - \mathbf{v} : \mathbf{s} \in A\}$ é o conjunto A modificado pelo vetor $-\mathbf{v}$, pois o número de pontos resultantes de um conjunto de observação fixo A e o número de pontos do processo deslocado $N_{\mathbf{v}}$, é igual ao número de pontos obtidos quando fixado o processo pontual e deslocando o conjunto de observação na direção oposta. A Equação (3.2) implica

$$N(A) \stackrel{d}{=} N(A_{\mathbf{v}}), \quad \text{para todo } A \text{ e } \mathbf{v},$$

isto é, o número de pontos na região A e no conjunto deslocado $A_{\mathbf{v}}$ tem a mesma distribuição. Isto produz importantes consequências para as características de distribuição dos processos pontuais (ILLIAN et al., 2008; BADDELEY et al., 2015).

O termo estacionariedade é geralmente usado para se referir a resultados de alguns processos que possuem propriedades similares em todas as localizações na região de estudo. Em outras palavras, as propriedades estatísticas, como por exemplo média e variância, da variável ou variáveis, não se alteram ao longo da área de interesse. Assim, um modelo estacionário possui todos os parâmetros iguais em todas as localizações (ILLIAN et al., 2008).

3.3.1.3 Isotropia

Um processo é dito ser isotrópico se ele é invariante em relação à rotação em torno da origem (ILLIAN et al., 2008). No caso do plano, uma rotação é descrita por um ângulo θ que varia de 0° à 360° . Se \mathbf{s} é um ponto do \mathbb{R}^2 , então o ponto rotacionado $R_\theta \mathbf{s}$ tem como coordenadas,

$$x_\theta = x \cos \theta + y \sin \theta \quad \text{e} \quad y_\theta = -x \sin \theta + y \cos \theta.$$

O processo pontual N é dito ser isotrópico se

$$N = \{\mathbf{s}_1, \mathbf{s}_2, \dots\} \quad \text{e} \quad R_\theta N = \{R_\theta \mathbf{s}_1, R_\theta \mathbf{s}_2, \dots\},$$

possuem a mesma distribuição para todo θ .

3.3.1.4 Propriedade de primeira ordem

O momento de primeira ordem ou propriedade de primeira ordem é representado por uma função $\lambda(\mathbf{x})$, que é a função intensidade de primeira ordem. Ela indica a intensidade do processo na localização \mathbf{x} . Segundo Câmara et al. (2000), os efeitos de primeira ordem, considerados globais ou de larga escala, correspondem a variações no valor médio do processo no espaço.

A função intensidade de primeira ordem de um processo pontual é dada por

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}| \rightarrow 0} \left\{ \frac{E[N(d\mathbf{s})]}{|d\mathbf{s}|} \right\},$$

em que $E[\cdot]$ denota o valor esperado, $d\mathbf{s}$ é uma pequena região ao redor do ponto \mathbf{s} e $|d\mathbf{s}|$ é a área desta região.

3.3.1.5 Propriedade de segunda ordem

A função de intensidade de segunda ordem que descreve a estrutura de correlação espacial de um processo pontual, que é uma medida de estrutura de dependência entre as localizações \mathbf{s}_1 e \mathbf{s}_2 , é dada por

$$\lambda_2(\mathbf{s}_1, \mathbf{s}_2) = \lim_{|d\mathbf{s}_1|, |d\mathbf{s}_2| \rightarrow 0} \frac{E[N(d\mathbf{s}_1)N(d\mathbf{s}_2)]}{|d\mathbf{s}_1||d\mathbf{s}_2|},$$

em que $E[\cdot]$ denota o valor esperado, $d\mathbf{s}_1$ e $d\mathbf{s}_2$ são uma pequena região ao redor dos pontos \mathbf{s}_1 e \mathbf{s}_2 respectivamente, sendo $|d\mathbf{s}_1|$ e $|d\mathbf{s}_2|$ a área de cada região (DIGGLE, 2003). Esta função representa os efeitos de segunda ordem, denominados locais ou de pequena escala.

Se o processo é estacionário, tem-se $\lambda(\mathbf{s}) = \lambda$, e se também for isotrópico, $\lambda(\mathbf{s}_1, \mathbf{s}_2)$ se reduz à $\lambda(|h|)$, sendo $|h|$ a distância entre os dois pontos. Em caso de processo não estacionário, ou seja, quando a intensidade média varia na região A , a modelagem da estrutura de dependência $\lambda(\mathbf{s}_1, \mathbf{s}_2)$ deve incorporar a variação de $\lambda(\mathbf{s})$ (DIGGLE, 2003; ILLIAN et al., 2008).

3.4 Processos de Poisson

Seja $\{N(t), t \geq 0\}$ um processo estocástico de valor inteiro, onde $N(t)$ é uma variável aleatória que indica o número de ocorrências no intervalo $(0, t]$, diz que é um processo de contagem, se cumpre as seguintes propriedades:

1. $N(t) \geq 0$;
2. Se $t_1 < t_2$, então $N(t_1) \leq N(t_2)$;
3. Para $t_1 < t_2$, $N(t_2) - N(t_1)$ é um número de sucessos (eventos) que ocorrem ou ocorrências no intervalo $(t_1, t_2]$.

Um processo de contagem tem incrementos independentes se o número de sucessos que ocorrem em intervalos disjuntos são independentes, isto quer dizer que o número de

sucessos que ocorre no intervalo $(t_1, t_2]$: $N(t_2) - N(t_1)$ é independente do número de sucessos que ocorre no intervalo $(t_3, t_4]$: $N(t_4) - N(t_3)$, para $t_1 < t_2 < t_3 < t_4$. Diz-se que um processo de contagem tem incrementos estacionários se a distribuição de probabilidade do número de sucessos que ocorrem em um intervalo depende apenas da amplitude do intervalo, isto é, seja $(t_1, t_1 + t_2]$ o intervalo de ocorrência, a probabilidade

$$P[N(t_1 + t_2) - N(t_1)] = P[N(t_2) - N(0)].$$

Um processo de Poisson é um processo de contagem que conta o número de ocorrências de algum evento específico através do tempo. O processo de Poisson é uma ferramenta natural de modelagem de inúmeros problemas de probabilidade aplicados. Ele não modela apenas muitos fenômenos do mundo real, mas também análises matemáticas passíveis de serem obtidas.

Um processo de contagem $\{N(t), t \geq 0\}$ se diz que é um processo de Poisson, com taxa $\lambda > 0$, se e só se

1. $N(0) = 0$;
2. $\{N(t), t \geq 0\}$ é um processo com incrementos independentes e estacionários;
3. $P[N(t) = 1] = \lambda t, \quad t > 0$;

A probabilidade de que em um intervalo $[0, t]$ exista a ocorrência de um sucesso é diretamente proporcional à amplitude t , já que λ é uma taxa positiva do processo.

4. $P[N(t) \geq 2] = O(t)$, ou seja, a probabilidade que exista mais de um sucesso em um intervalo pequeno tende a zero. $O(t)$ é uma função tal que

$$\lim_{t \rightarrow 0} \frac{O(t)}{t} = 0.$$

Se $\{N(t), t \geq 0\}$ é um processo de Poisson com taxa $\lambda > 0$, então a probabilidade da ocorrência de n sucessos no intervalo $(0, t]$ é

$$P_n(t) = P[N(t) - N(0) = n] = P[N(t) = n] = \frac{e^{-\lambda t} (\lambda t)^n}{n!}, \quad n = 0, 1, \dots$$

O processo de Poisson possui um papel fundamental em análise de processos pontuais. Além de ser utilizado quando os pontos apresentam dispersão aleatória sobre a região em estudo, também serve como referência ou modelo nulo para a elaboração de demais modelos (ILLIAN et al., 2008).

3.4.1 Processo de Poisson Homogêneo

O processo de Poisson homogêneo (PPH) é o mais simples e mais importante modelo de processos pontuais infinitos (ILLIAN et al., 2008). Um PPH N é caracterizado pelas seguintes propriedades:

- Homogeneidade e distribuição de Poisson: o número de eventos de N em qualquer região limitada A , segue uma distribuição de Poisson com média $\lambda|A|$ para alguma constante λ ;
- Independência: o número de eventos de $N(A \cap A_j)$ em m conjuntos disjuntos A_1, \dots, A_m , formam m variáveis aleatórias independentes, para m arbitrário.

λ é o parâmetro característico, denominado de intensidade ou densidade pontual do PPH. Tem por definição descrever o número médio de pontos em uma unidade de área.

O PPH é um processo pontual espacial que produz um padrão pontual com intensidade constante dentro do conjunto aleatório A_i , ou seja, quaisquer dois subconjuntos disjuntos de A possuem intensidade igual, e o número de eventos dentro dos subconjuntos são independentes e com distribuição de Poisson. Assim, eventos em um PPH são distribuídos uniformemente e independentemente dentro de A . Por esta razão, as propriedades de um PPH são coletivamente chamadas de CAE (ILLIAN et al., 2008).

Se os eventos de um padrão pontual estão dispostos aleatoriamente sobre A , esse é denominado de CAE. Um padrão pontual pode exibir interação positiva entre eventos e possuir aglomerados com maior densidade de ocorrências, ou ainda pode exibir regularidade. A Figura 1 representa simulações de processos de Poisson para os três tipos de distribuição espacial de eventos aleatório, cluster e regular, respectivamente.

3.4.2 Estimação de primeira ordem

Uma questão fundamental sobre a intensidade de um processo pontual é se ela é homogênea ou não. A partir dela, é possível observar preferências, evasões e aglomerações. Quando a intensidade é espacialmente variante, ela é efetivamente uma função da localização espacial, e por conseguinte, métodos de estatística podem ser utilizados para estimar a função a partir dos dados (BADDELEY & TURNER, 2006).

A estimação da intensidade λ de um processo de Poisson homogêneo é a questão estatística fundamental (ILLIAN et al., 2008) uma vez que é o único parâmetro de um modelo espacial pontual. Se todos os pontos dentro da região A puderem ser contados, o número de pontos $N(A)$ dividido pela área de A é o estimador natural

$$\hat{\lambda} = \frac{N(A)}{|A|},$$

que é o estimador de máxima verossimilhança não viesado. A investigação da intensidade de um padrão pontual é uma das primeiras e mais importantes etapas em análise de dados pontuais (BADDELEY & TURNER, 2006) por ser a característica descritiva básica de um processo pontual.

3.4.2.1 Estimação

A estimação da propriedade de primeira ordem, ou intensidade do padrão pontual, é usualmente realizada pelas funções kernel, que conseguem obter uma estimativa de uma

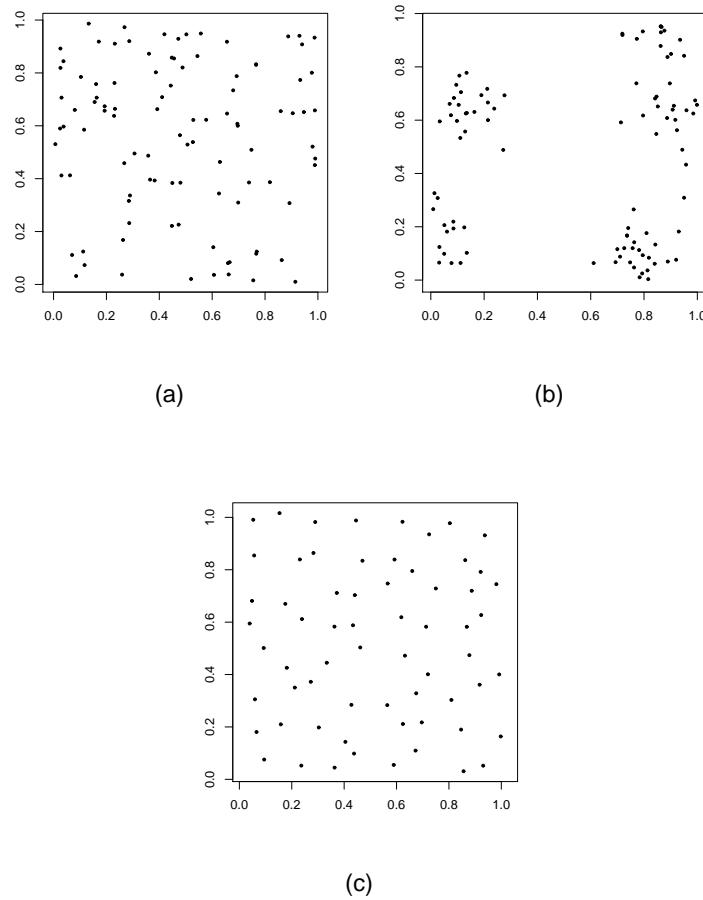


Figura 1 Simulação dos três tipos de padrões pontuais CAE (a), cluster (b) e regular (c), respectivamente.

densidade de probabilidade univariada ou bivariada a partir de uma amostra observada,

$$f(\mathbf{s}) = \frac{1}{N(A)} \sum_{i=1}^{N(A)} k\left(\frac{\mathbf{s} - \mathbf{s}_i}{q}\right),$$

em que $k(\cdot)$ é a função kernel com $\int k(\mathbf{s})d\mathbf{s} = 1$, e $q > 0$ é o parâmetro de suavização (BADDELEY & TURNER, 2006; BADDELEY et al., 2015). A estimativa da intensidade no ponto \mathbf{s} depende somente da relação espacial entre \mathbf{s} e os elementos da amostra \mathbf{s}_i ($i = 1, \dots, N(A)$), quantificada pela função kernel. Para maiores detalhes sobre estimação kernel, veja Li & Racine (2007) e Baddeley et al. (2000).

3.4.3 Estimação de segunda ordem

A propriedade de segunda ordem de um processo pontual está intimamente relacionada com a contagem de pares de pontos ou a contribuição de cada par de ponto no processo e descreve o relacionamento entre os eventos. As distâncias entre as localizações dos eventos possibilitam estudar características do processo de interesse por descrições sumárias e averiguação da hipótese de completa aleatoriedade espacial (CAE), por exemplo. Uma técnica muito utilizada para estimar a correlação de um padrão pontual é a função K de

Ripley (RIPLEY, 1976; RIPLEY, 1977).

3.4.3.1 A função K homogênea

A função $K(r)$ de Ripley tem sua origem em Bartlett (1964), porém foi Ripley (1976) que mostrou sua importância como resumo de dependência espacial sobre uma ampla gama de escalas. Como ela permite quantificar a dependência espacial entre diferentes regiões do processo, é também conhecida como análise de segunda ordem reduzida.

Esta função tem sido muito utilizada em processos pontuais bidimensionais. Apresenta as vantagens de permitir a detecção do padrão pontual em diferentes escalas de distâncias simultaneamente e o padrão pontual observado pode ser comparado aos modelos de processos estocásticos conhecidos para diferentes configurações pontuais (DIGGLE, 2003).

Diz-se que um processo é ordenado se eventos coincidentes múltiplos não podem ocorrer, ou mais precisamente, $P[N(d\mathbf{s}) > 1]$ é de ordem de magnitude menor que $|d\mathbf{s}|$. Isto significa que $E[N(d\mathbf{s})] \sim P[N(d\mathbf{s}) = 1]$, no sentido de que a razão destas duas quantidades tende a 1 quando $|d\mathbf{x}| \rightarrow 0$ (DIGGLE, 2003).

Uma caracterização alternativa das propriedades de segunda ordem de um processo estacionário, isotrópico e ordenado é dado por

$$\lambda K(r) = E[N_0(r)],$$

em que $N_0(r)$ é o número de eventos dentro da distância r de um evento arbitrário. A noção de um evento arbitrário de um processo envolve o limite conceitual de amostragem aleatória simples de uma população finita. Intuitivamente, prevê-se um grande, mas finito número N de eventos em alguma região finita do plano. Um evento selecionado ao acaso desta população é, por definição, um evento arbitrário. Similarmente, na prática, um ponto arbitrário significará um ponto distribuído uniformemente sobre alguma região finita (DIGGLE, 2003).

Em outras palavras, $\lambda K(r)$ denota o número médio de pontos em um disco de raio r , centrado em um evento aleatório (que não é contado). Sejam $N(A) = n$ e $n_i(r) = N[b(\mathbf{s}_i, r) \setminus \{\mathbf{s}_i\}]$ o número de pontos do padrão pontual N com distância r do ponto \mathbf{x}_i , excluindo \mathbf{s}_i . Aqui, $b(\mathbf{s}_i, r)$ é a bola de centro \mathbf{s}_i e raio r . Então $\bar{n}(r)$, a média do número de pontos que distam r de \mathbf{s}_i , é um estimador natural para $E(N_0(r))$, dado como

$$\bar{n}(r) = \frac{\sum_{i=1}^n n_i(r)}{n}. \quad (3.3)$$

Os pontos $\mathbf{s}_1, \dots, \mathbf{s}_n$ são tomados sucessivamente e N é mudado de tal forma que o evento aleatório fique na origem o , isto é, $N - \mathbf{s}_i$ é considerada para $i = 1, \dots, n$. Assim, o número $n_i(r)$ em $N - \mathbf{s}_i$ na bola $b(o, r)$ pode ser obtido. A sua média produz uma estimativa do número médio de pontos em uma bola de raio r centrada em um ponto do padrão pontual, em que, para todos os pontos \mathbf{s}_i ele nunca é contado. Logo,

$$K(r) = \frac{E_o[N(b(o, r)) \setminus \{o\}]}{\lambda}, \quad r \geq 0,$$

a função $K(r)$ depende do raio r , com o aumento de r , o número de pontos na bola $b(o, r)$ aumenta, e como resultado, $K(r)$ também aumenta. A estimação de K está baseada na variação do número médio de outros pontos em torno de um ponto qualquer, à medida que se varia o raio considerado (DIGGLE, 2000).

Denota-se por $E(r) = E[N_0(r)]$ o número esperado de outros eventos dentro de distância r de um evento arbitrário. Assim, pode-se construir um estimador para $E(r)$. Para tal, seja $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$, distância euclideana entre dois eventos \mathbf{s}_i e \mathbf{s}_j . A Equação (3.3) pode ser reescrita como uma soma dupla, dado que o número de pontos na bola $b(\mathbf{s}_i, r)$ é igual ao número de pontos em N com distância igual ou menor que r de \mathbf{s}_i ,

$$\tilde{E}(r) = n^{-1} \sum_{i=1}^n \sum_{j \neq i} I_r(d_{ij}), \quad (3.4)$$

em que $I(\cdot)$ denota a função indicadora que será igual a 1 se $d_{ij} \leq r$ e 0 caso contrário.

A forma do estimador $\tilde{E}(r)$ da Equação (3.4) sugere, convenientemente, que a função K é intimamente conectada com a distribuição das distâncias inter-eventos. Entretanto, $\tilde{E}(r)$ é negativamente viesada para $E(r)$ por causa dos efeitos de borda (ILLIAN et al., 2008). Para um evento referência dentro da distância r da fronteira de A , a contagem observada de outros eventos dentro da distância r necessariamente exclui quaisquer eventos que podem ter ocorrido dentro da distância r mas fora de A (DIGGLE, 2003). Métodos têm sido propostos para corrigir esta fonte de viés e o descrito a seguir foi desenvolvido por Ripley (1976).

Seja $w(x, s)$ a proporção da área da circunferência com centro \mathbf{s} e raio u que se encontra em A . Faz-se w_{ij} para denotar $w(\mathbf{s}_i, u_{ij})$. Então, para um processo isotrópico e estacionário, w_{ij} é a probabilidade condicional que um evento é observado, dado que é apenas uma distância do i -ésimo evento \mathbf{x}_i . Em geral $w_{ij} \neq w_{ji}$. Então um estimador não-viesado para $E(r)$ é

$$\hat{E}(r) = n^{-1} \sum_{i=1}^n \sum_{j \neq i} w_{ij}^{-1} I_r(u_{ij}).$$

E, ao substituí-lo da intensidade desconhecida λ estimada por $\hat{\lambda} = (n - 1)/|A|$, obtêm-se o estimador para $K(r)$

$$\hat{K}(r) = \frac{|A|}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n w_{ij}^{-1} I_r(d_{ij}),$$

em que n é o número de pontos do padrão pontual com área $|A|$ e w_{ij} é um fator de correção que representa a proporção da circunferência ao redor do ponto \mathbf{x}_i , passando sobre o ponto \mathbf{s}_j que está dentro da área $|W|$.

Esta estimativa de $K(r)$ baseia-se na variação do número médio de pontos em torno de um ponto qualquer à medida que se varia o raio r , o que garante correção dos efeitos de bordas. A distância r não possui regra de determinação, pois depende do objeto de interesse e, geralmente, é escolhida pelo pesquisador.

Outros autores também estudaram formas de estimação da função $K(r)$, como por

exemplo Veen & Schoerberg (2006) que obtiveram uma função ponderada análoga à função K de Ripley. Osher (1983) e Osher & Stoyan (1981) apresentaram métodos de correção de borda para a função de Ripley, distintos dos apresentados em Ripley (1976). A função K também pode ser utilizada para avaliar anisotropia do padrão pontual a partir da análise de setores, na qual os pares de pontos são contados apenas quando o vetor que une o par acontece de se encontrar em uma determinada faixa de ângulos. Outra abordagem não paramétrica para estudar anisotropia em padrões de pontos espaciais por meio desta função foi proposta em Rajala et al. (2016).

Segundo Illian et al. (2008), a forma de $K(r)$ em relação ao processo de Poisson fornece uma importante informação acerca da distribuição pontual do processo. Casos típicos de processos que não são de Poisson são processos agrupados ou regulares, e estes podem ser observados a partir da análise do gráfico da função $K(r)$. Em particular, para um processo de Poisson homogêneo no \mathbb{R}^2 , $K(r) = \pi r^2$.

Para processos pontuais agrupados, por característica, os pontos possuem mais vizinhos do que o esperado sob a hipótese nula de CAE, portanto, as estimativas para $K(r)$ serão maiores que πr^2 . Assim, em processos agrupados, cada ponto estará provavelmente mais próximo do entorno de indivíduos do mesmo grupo, logo para valores pequenos de r , $K(r)$ será relativamente grande. Situação contrária deve ocorrer em processos regulares, em que os pontos têm, em média, menor número de vizinhos do que o esperado sob CAE. As estimativas de $K(r)$ serão menores que πr^2 , o que significa que eventos regulares implicam espaços vazios entre os pontos. Logo, para pequenos valores de r , $K(r)$ será relativamente pequena (DIGGLE, 2000). Na prática, padrões pontuais geralmente exibem uma combinação destes dois casos extremos (ILLIAN et al., 2008).

3.4.4 Teste para CAE

A suposição de que um determinado fenômeno ocorre de forma aleatória nem sempre é verdadeira. Uma forma de se averiguar esta afirmação é por testes de hipóteses. Um teste clássico, cuja hipótese nula é tal que a distribuição dos eventos é CAE, ou equivalentemente que o processo é PPH, e a hipótese alternativa é a negação, é o teste χ^2 . Este consiste em subdividir a área $|A|$ em m quadrados de igual área $|A_1|, \dots, |A_m|$, conta-se o número de pontos em cada quadrado $n_j = N(A \cap A_j)$ para $j = 1, \dots, m$. Sob a hipótese nula, n_j são variáveis aleatórias independentes e identicamente distribuídas com distribuição de probabilidade de Poisson com mesmo valor esperado. A estatística de teste é

$$\chi^2 = \frac{\sum_j (n_j - n/m)^2}{n/m},$$

em que $n = \sum_j n_j$. No R, este teste é realizado pelo comando `quadrat.test()`.

O poder do teste χ^2 depende do tamanho dos quadrados. Um outro teste é o Kolmogorov-Smirnov em que se comparam as distribuições dos valores observados e esperados de alguma função. Deve-se especificar uma função $T(\mathbf{s}_1, \mathbf{s}_2)$ definida para todas

as localizações ($\mathbf{s}_1, \mathbf{s}_2$) em A . Avalia-se a função em cada evento de A e compara-se esta distribuição empírica dos valores de T com a distribuição predita dos valores de T sob CAE a partir do teste de Kolmogorov-Smirnov clássico.

3.5 Referências

ALBERT, P.S.; McSHANE, L. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data, **Biometrics**, v.51, n.2, p.627-638, 1995.

ASSUMPÇÃO, R.A.B., URIBE-OPAZO, M.A.; GALEA, M. Analysis of local influence in geostatistics using Student-t distribution, **Journal of Applied Statistics**, v.41, n.1, p.2323-2341, 2014.

BADDELEY, A.; MOLLER, J.; WAAGEPETERSEN, R. Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. **Statistica Neerlandica**, v.54, n.3, p.329-350, 2000.

BADDELEY, A.; RUBAK, E.; TURNER, R. **Spatial point patterns: Methodology and Applications with R**. Boca Raton: Chapman & Hall/CRC, 2015. 810p.

BADDELEY, A.; TURNER, R. **Modelling Spatial Point Patterns in R**. In: Baddeley A., Gregori P., Mateu J., Stoica R., Stoyan D. (eds) Case Studies in Spatial Point Process Modeling. Lecture Notes in Statistics, vol 185. Springer, New York, 2006.

BADDELEY, A.; TURNER, R. spatstat: An R Package for Analyzing Spatial Point Patterns. **Journal of Statistical Software**, v.12, n.6, p.1-42, 2005.

BARTLETT, M.S. The spectral analysis of two dimensional point process. **Biometrika**, v.51, n.3/4, p.299-311, 1964.

CÂMARA, G.; CARVALHO, M.S. **Análise Espacial de Eventos**. 2005. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap2-eventos.pdf>. Acesso em: 06 maio 2015.

CÂMARA, G.; MONTEIRO, A.G.; CARVALHO, M.S. **Análise espacial e geoprocessamento**. In: Fuks, S.D.; M.S. Carvalho; G. Câmara; A.M.V. Monteiro (eds.), *Análise espacial de dados geográficos*. 2000. 209 p. Instituto Nacional de Pesquisas Espaciais: Divisão de Processamento de Imagens, São José dos Campos, SP. Disponível em: <<http://www.dpi.inpe.br/gilberto/livro/analise>>. Acesso em: 06 de maio 2014.

CRESSIE, N.A.C. **Statistics for spatial data**. New York: John Wiley & Sons, 1993. 900p.

CRESSIE, N.A.; HAWKINS, D. (1980). Robust estimation of the variogram. **Journal of the International Association for Mathematical Geology**, v.12, n2, 115-125, 1980.

DALEY, D.J.; VERE-JONES, D. **An Introduction to the Theory of Point Processes**. Second Edition, New York: Springer. 2002. 573p.

- DE BASTIANI, F.; CYSNEIROS, A.H.M.A.; URIBE-OPAZO, M.A.; GALEA, M. Influence diagnostics in elliptical spatial linear models, **TEST**, v.24, n.2, p.322-340, 2015.
- DIGGLE, P. **Spatial Statistics for Environmental Epidemiology**. Notas de Curso. 2000. 70p.
- DIGGLE, P. **Statistical Analysis of Spatial Point Patterns**. 2ed. New York: Oxford University Press Inc., 2003. 159p.
- DIGGLE, P.; RIBEIRO JR, P.J. **Model-Based Geostatistics**. Lancaster: Springer, 2007. 228p.
- GARCIA-PAPANI, F.; URIBE-OPAZO, M.A.; LEIVA, V.; AYKROYD, R.G. Birnbaum-saunders spatial modelling and diagnostics applied to agricultural engineering data, **Stochastic Environmental Research and Risk Assessment**, v.31, n.1, p.105-124, 2017.
- GBUR, E.E.; STROUP, W.W.; McCARCTER, K.S.; DURHAM, S.; YOUNG, L.J., CHRISTMAN, M.; WEST, M.; KRAMER, M. **Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences**. Madison: American Society of Agronomy, Soil Science of America, Crop Science Society of America. 2012. 299p.
- GOSHO, M.; HAMADA, C.; YOSHIMURA, I. Criteria for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data, **Communications in Statistics - Theory and methods**, v.40, n.21, p.3839-3856, 2011.
- GOTWAY, C.A.; STROUP, W.W. A generalized linear model approach to spatial data analysis and prediction, **Journal of Agricultural, Biological and Environmental Statistics**, v.2, n.2, p.157-178, 1997.
- GUEDES, L.P.C.; URIBE-OPAZO, M.A.; RIBEIRO JR, P.J. Influence of incorporating geometric anisotropy on the construction of thematic maps simulated data and chemical attributes of soil, **Chilean Journal of Agricultural Research**, v.73, n.4, p.414-423, 2013.
- GUAN, Y.T. **Nonparametric methods of assessing spatial isotropy**. 116p. Trabalho de conclusão de curso (Tese) - Philosophy, Texas A&M University - A&M, College Station, 2003.
- ILLIAN, J.; PENTTINEN, A.; STOYAN, H.; STOYAN, D. **Statistical Analysis and Modelling of Spatial Point Patterns**. England: John Wiley & Sons Ltda, 2008. 534p.
- ISAAKS, E.H.; SRIVASTAVA, R.M. **An Introduction to Applied Geostatistics**. New York: Oxford University Press. 1989. 561p.
- LI, Q.; RACINE, J.S. **Nonparametric Econometrics: Theory and Practice**. Princeton: University Press, 2007. 768p.
- LIANG, K-Y.; ZEGER, S.L. Longitudinal data analysis using generalized linear models, **Biometrika**, v.73, n.1, p.13-22, 1986.
- MARDIA, K.V.; MARSHALL, R.J. Maximum likelihood estimation of models for residual covariance in spatial regression, **Biometrika**, v.71, n.1, p.135-146, 1984.
- MATÉRN, B. **Lecture notes in statistics**. New York: Springer, 1986.

- MATHERON, G. Principles of geostatistics, **Economic Geology**, v.58, n.1, p.1246-1266, 1963.
- McCULLAGH, P. Quasi-likelihood functions, **Annals Statistic**, v.11, n.1, p.59-67, 1983.
- NAVA, D.T.; DE BASTIANI, F.; URIBE-OPAZO, M.A.; NICOLIS, O.; GALEA, M. Local influence for spatially correlated binomial data: An application to the *Spodoptera frugiperda* infestation in corn, **Journal of Agricultural, Biological and Environmental Statistics**, v.22, n.4, p.540-561, 2017.
- NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models, **Journal of the Royal Statistical Society**, v.135, n.3, p.370-384, 1983.
- OSHER, J. On estimators for the reduced second moment measure of point process. **Mathematische Operationsforschung und Statistik, series Statistics**, v.14, n.1, p.63-71, 1983.
- OSHER, J.; STOYAN, D. On the second-order and orientation analysis of planar stationary point process. **Biometrical Journal**, v.23, n.6, p.523-533, 1981.
- PAN, W. Akaike's information criterion in generalized estimating equations, **Biometrics**, v.57, n.1, p.120-125, 2001.
- PAULA, G.A. **Modelos de Regressão com apoio computacional**, IME-USP, São Paulo, 2013. 429p.
- PARDO, M.C.; ALONSO, R. Working correlation structure selection in GEE analysis. **Statistical Papers**, <https://doi.org/10.1007/s00362-017-0881-0>, 2017.
- POISSON, S.D. **Recherches sur la Probabilite des Jugements en Matiere Criminelle et en Matiere Civile, Precedees des Regles Generales du Calcul des Probabilites**, Paris: Bachelier, 1837. 442p.
- PRENTICE, R.L. Correlated binary regression with covariates specific to each binary observation, **Biometrics**, p.44, n.4, p.1033-1048, 1988.
- R CORE TEAM. **R: A language and environment for statistical computing**. R foundation for Statistical Computing, Vienna, Austria, 2017.
- RAJALA, T.A.; SÄRKKÄ, A.; REDENBACH, C.; SORMANI, M. Estimating geometric anisotropy in spatial point patterns. **Spatial Statistics**, v.15, n.1, p.100-114, 2016.
- RIPLEY, B.D. The second-order analysis of stationary point processes. **Journal of Applied Probability**, v.13, n.2, p.255-266, 1976.
- RIPLEY, B.D. Modelling spatial patterns (with discussion). **Journal Royal Statistics Society**, v.39, n.2, p.172-212, 1977.
- SOARES, A. **Geoestatística para as ciências da terra e do ambiente**. 2ed. Portugal: IST Press, 2006.

SRIVASTAVA, R.M.; PARKER, H.M. Robust measures of spatial continuity. In: **Geostatistics, on the 3rd. Geostatistical Congress**. Holland, Armstrong: [s.n.], 1989. p.295-308.

URIBE-OPAZO, M.; BORSSOI, J.; GALEA, M. Influence diagnostics in gaussian spatial linear models, **Journal of Applied Statistics**, v.39, n.3, p.615-630, 2012.

VEEN, A.; SCHOENBERG, F.P. Assessing Spatial Point Process Models Using Weighted K-functions: Analysis of California Earthquakes. In: Baddeley A., Gregori P., Mateu J., Stoica R., Stoyan D. (eds) **Case Studies in Spatial Point Process Modeling**. Lecture Notes in Statistics, vol 185. Springer, New York, 2006.

WEBSTER, R.; OLIVER, M. **Geostatistics for Environmental Scientists**. 2ed. Chichester: John Wiley & Sons, 2007. 315p.

WEDDERBURN, R.W.M. Quasi-likelihood functions, generalized linear models and the gauss-newton method, **Biometrika**, v.61, n.3, p.439-447, 1974.

ZEGER, S.; LIANG, K-Y. Longitudinal data analysis for discrete and continuous outcomes, **Biometrics**, v.42, n.1, p.121-130, 1986.

4 ARTIGOS

4.1 ARTIGO 1: Influência local para dados binomiais espacialmente correlacionados: Uma aplicação da infestação de *Spodoptera frugiperda* no milho

Resumo: Diagnósticos de influência são ferramentas importantes para o entendimento da influência dos dados e/ou das suposições do modelo nos resultados de uma análise estatística. Este artigo propõe influência local para a análise de dados binomiais espacialmente correlacionados. Considera-se um modelo espacial com distribuição marginal binomial e função de ligação logit. Equações de estimação generalizadas via escore de Fisher foram usadas para a estimação dos parâmetros. Apresentou-se uma aplicação da infestação espacial da *Spodoptera frugiperda*, onde foram utilizadas equações de estimação generalizadas para identificar possíveis observações influentes por análise de influência local. Comparou-se a predição espacial com e sem os pontos influentes. Os resultados mostram que a presença de observações influentes nos dados modificaram a inferência estatística, os valores preditos e os respectivos mapas. Um estudo de simulação, considerando diferentes cenários, foi realizado para representar a performance do método de diagnóstico de influência local.

Palavras-chave: Distribuição Binomial; escore de Fisher; família exponencial; outliers; quase-verossimilhança.

4.1.1 Introdução

O milho é um alimento nutritivo utilizado como fonte de alimento humano nas formas de óleo, farinha, amido, margarina, xarope de glicose e cereal, e como alimento animal para pássaros e suínos, dentre outras. Adicionalmente, ele pode ser usado na produção de dinamite, cosméticos, etanol, etc.

A lagarta do cartucho, cujo nome científico é *Spodoptera frugiperda*, é uma das pragas que mais ataca as plantações de milho, e contribui para uma redução de até 34% da produtividade dos grãos (Cruz e Turpin 1983). Ela é considerada a praga que mais causa danos às lavouras de milho, não apenas pelo dano que causa, mas também pela dificuldade para realizar seu controle. O inseto ataca o cartucho da planta, consome grande parte da área foliar antes que esta se abra e dependendo do estágio de crescimento da cultura, pode levar a planta à morte. A lagarta se alimenta em todas as fases de crescimento da planta, embora prefira as plantas mais jovens. Com o passar das fases da cultura, o controle da praga se torna mais complicado devido às dificuldades de acesso à lavoura.

Este artigo foi estruturado seguindo as normas da *Journal of Agricultural, Biological, and Environmental Statistics*, que encontram-se no Anexo 6.1. Está disponível online no link <https://link.springer.com/journal/13253>, com número DOI: 10.1007/s13253-017-0306-5.

A resistência aos inseticidas da lagarta do cartucho foi investigada por Yu (1991). Belay et al. (2012) verificaram a susceptibilidade da lagarta do cartucho em Porto Rico a diferentes inseticidas. Goergen et al. (2016) apresentam um relato de surtos de *Spodoptera frugiperda* chamando-a de uma nova praga invasora alienígena na África Ocidental e Central. Existem muitos estudos sobre esta lagarta, no entanto, são poucos os que consideram a probabilidade de uma planta ser infectada por este inseto considerando-se dados binomiais espacialmente correlacionados.

O uso de técnicas geoestatísticas quando os dados possuem distribuição espacial normal é amplamente conhecido. No entanto, quando têm-se dados binomiais espacialmente correlacionados, é importante considerar modelos que incorporem esta característica. Dados espaciais binomiais ocorrem em pesquisas ambientais e ecológicas, por exemplo quando os dados correspondem à presença ou ausência de certa espécie de planta, ou invasão de praga em alguma localização, ou quando os dados ocorrem em uma ou duas categorias, tais como dois tipos de solos (Haran 2011). A geoestatística na agricultura de precisão é importante para controlar a *Spodoptera frugiperda* na cultura do milho, pois fornece um sistema de informação de gerenciamento integrado e tecnologias baseadas no princípio de que a variabilidade espacial influencia o rendimento das culturas.

Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLGs), os quais permitem que a distribuição da variável resposta pertença à classe de distribuição da família exponencial, proporcionando assim diferentes possibilidades para a relação entre a média da variável resposta e o preditor linear. Uma generalização para os MLGs foi apresentada em 1974 por Wedderburn (1974), denominada de modelos de quase-verossimilhança. Esta metodologia especifica o relacionamento entre a média e a variância usando uma função de variância, em que a função da variância é uma função da média. Liang e Zeger (1986) consideram um método estreitamente relacionado à metodologia de quase-verossimilhança utilizando equações de estimação generalizadas (EEGs) para a análise de dados longitudinais. Posteriormente, este trabalho foi generalizado por Prentice (1988) e Prentice e Zhao (1991).

As EEGs assumem que o valor esperado da variável resposta é uma função linear das covariáveis e que a variância é uma função conhecida da média. A forma da estrutura de correlação é desconhecida e uma estrutura de correlação de trabalho (Liang e Zeger 1986) entre as unidades amostrais devem ser especificadas e geradas pela distribuição de probabilidades assumida. Albert e McShane (1995) sugeriram uma adaptação do método proposto por Prentice (1988) quando os dados são espacialmente correlacionados com distribuição binomial.

Na modelagem estatística, é importante investigar possíveis mudanças devido à influência dos dados e às suposições do modelo nos resultados (Cadigan e Farrel 2002). Cook (1986) apresentou um método que usa o afastamento da verossimilhança para medir influência local quando os modelos estatísticos estão sob pequenas perturbações. Cadigan e Farrel (2002) estudaram diagnósticos de influência local para modelos estatísticos no qual

a estimação envolve uma função objetivo, duas vezes diferenciável com respeito ao vetor de parâmetros e produz estimativas de parâmetros interiores únicos.

Na análise de dados espaciais, Militino et al. (2006) apresentaram um estudo acerca de outliers e análise de influência para dados multivariados. Uribe-Opazo et al. (2012) usaram técnicas de diagnóstico para avaliar a sensibilidade dos estimadores de máxima verossimilhança, funções de covariância e preditores lineares quando existem pequenas perturbações nos dados e/ou nas suposições do modelo espacial linear Gaussiano. De Bastiani et al. (2015) consideraram a família de distribuição elíptica para estimar a dependência espacial de dados georeferenciados e avaliaram a sensibilidade dos estimadores de máxima verossimilhança considerando-se técnicas de influência local. Garcia-Papani et al. (2017) apresentaram modelagem e diagnósticos para dados assimétricos espaciais através da distribuição espacial Birnbaum-Saunders com aplicações a dados de agricultura. No contexto de equações de estimação generalizadas, Jung (2008) definiu um afastamento para a quase-verossimilhança e sugeriu diagnósticos para estimar parâmetros de regressão baseados em influência local. Até o momento, há uma escassez de literatura que descreva a infestação espacial da *Spodoptera frugiperda* em lavouras de milho, e especialmente acerca de influência local para dados binomiais espacialmente correlacionados.

Neste artigo, considera-se um modelo com distribuição marginal binomial e função de ligação logit. Introduce-se uma pequena perturbação no modelo para observar a influência dos dados e/ou das suposições do modelo nos resultados da análise estatística. A Seção 4.1.2 apresenta modelos binomiais com correlação espacial. Na Seção 4.1.3, a influência local de observações em predição espacial é descrito. A seção 4.1.4 introduz o conjunto de dados reais e os resultados. Um estudo de simulação é mostrado na Seção 4.1.5. Algumas conclusões são descritas na Seção 4.1.6 e o Apêndice possui o conjunto de dados reais e detalhes do modelo.

4.1.2 Modelos binomiais com correlação espacial

Seja $\mathbf{Y} = (Y_1(s_1), \dots, Y_n(s_n))^T$ um vetor de dimensão $n \times 1$ de variáveis aleatórias binomiais associadas a n dados observados y_1, \dots, y_n , com m_i indivíduos no grupo i coletados nas localizações espaciais s_1, \dots, s_n , em que Y_i é o número de sucessos observados no grupo i e $\mathbf{X}(s) = \mathbf{X}$ é uma matriz $n \times p$ de variáveis exploratórias com j -ésima coluna $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$ para $j = 1, \dots, p$. Em MLGs, as observações não são necessariamente relacionadas ao parâmetro de localização, dado que o modelo é ajustado para alguma função deste parâmetro, $g(\cdot)$, chamado de função de ligação. O modelo espacial linear binomial correspondente tem

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\mu}, \\ \boldsymbol{\eta} &= g(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}, \end{aligned} \tag{4.1}$$

em que $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$, com $\mu_i = m_i \pi_i$ para $i = 1, \dots, n$, m_i é o número de plantas em cada parcela, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$ com π_i a probabilidade de que um sucesso é observado em qualquer indivíduo no grupo i , e $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$ é o preditor linear e $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é um vetor $p \times 1$ de parâmetros desconhecidos.

A função de ligação da Equação (4.1) é definida pela função de ligação logit, dada por

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip},$$

assim, $\pi_i = \mu_i / m_i = \exp(\eta_i) / (1 + \exp(\eta_i))$ e $1 - \pi_i = 1 / (1 + \exp(\eta_i))$.

A matriz de variância-covariância de \mathbf{Y} , denotada por \mathbf{V} , pode ser escrita como

$$\mathbf{V} = \left[v_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\alpha}) v_{\boldsymbol{\mu}}^{\frac{1}{2}} \right], \quad (4.2)$$

em que $v_{\boldsymbol{\mu}} = \text{diag}[m_i \pi_i (1 - \pi_i)]$, $\mathbf{R}(\boldsymbol{\alpha})$ é uma matriz $n \times n$ de correlação espacial simétrica positiva definida, e $\boldsymbol{\alpha}$ é um vetor de parâmetros caracterizando a estrutura de correlação. A matriz $\mathbf{R}(\boldsymbol{\alpha})$ é conhecida como matriz de correlação de trabalho. Em geoestatística, a matriz de correlação pode ser obtida a partir dos modelos de semivariograma, conforme os descritos em Cressie (1993). As Seções 4.1.2.1 e 4.1.2.2 explicam a metodologia utilizada para estimar os vetores de parâmetros desconhecidos $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$.

4.1.2.1 Básico em Geoestatística

Quando a variável resposta é medida em pontos no espaço e apresenta alguma característica de dependência espacial, a geoestatística deve ser utilizada. Uma das ferramentas utilizadas em geoestatística para estudar a estrutura de dependência espacial entre amostras vizinhas é o semivariograma. Este gráfico objetiva extrair a estrutura de dependência dos dados a partir da medida de correlação entre valores amostrados em dois pontos do espaço e o raio de dependência entre as amostras.

O semivariograma (Matheron 1963) denotado como $\gamma(h_{ij})$ pode ser estimado utilizando o estimador de Matheron, dado por

$$\hat{\gamma}(h_{ij}) = \frac{1}{2N(h_{ij})} \sum_{N(h_{ij})} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2,$$

em que $N(h_{ij})$ é o número de pares de valores medidos $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$, separados pela distância h_{ij} , onde $h_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ é a distância Euclideana entre as localizações \mathbf{s}_i e \mathbf{s}_j , e $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$ são os valores da i -ésima e j -ésima observações da variável regionalizada amostrada nas localizações \mathbf{s}_i e \mathbf{s}_j , $i, j = 1, \dots, n$.

No semivariograma, os parâmetros que definem a estrutura de dependência espacial são o efeito pepita, o alcance e o patamar. O efeito pepita, denotado como ϕ_1 , pode ser atribuído a erros de medição ou fontes de variação espacial a distâncias inferiores ao intervalo amostral ou ambos (Cressie 1993). O alcance, denotado como $a = c(\phi_3)$, representa a distância dentro da qual as amostras possuem dependência espacial. O alcance é uma

função do parâmetro ϕ_3 , e o patamar, obtido como $\phi_1 + \phi_2$, representa o limite do variograma quando as distâncias deste tendem ao infinito, denota-se o conjunto de parâmetros espaciais por $\alpha^\top = (\phi_1, \phi_2, \phi_3)^\top$, para maiores detalhes veja Uribe-Opazo et al. (2012).

Por exemplo, a função semivariância para a família Matérn (Matérn 1960) é dada por

$$\gamma(h_{ij}) = \begin{cases} 0, & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[\{1 - 2^{\kappa-1} \Gamma(\kappa)\}^{-1} \left(\frac{h_{ij}}{\phi_3}\right)^\kappa K_\kappa\left(\frac{h_{ij}}{\phi_3}\right) \right], & \text{if } h_{ij} > 0 \end{cases},$$

em que $\phi_j \geq 0$, $j = 1, 2, 3$ e o parâmetro de ordem $\kappa > 0$; $\Gamma(\kappa)$ é a função gama, e $K_\kappa(\cdot)$ é a função Bessel modificada de terceiro tipo de ordem κ . Para $\kappa = 0.5$, tem-se o modelo exponencial

$$\gamma(h_{ij}) = \begin{cases} 0 & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[1 - \exp\left(-\frac{h_{ij}}{\phi_3}\right) \right], & \text{if } 0 < h_{ij} \leq \phi_3 \\ \phi_1 + \phi_2, & \text{if } h_{ij} > \phi_3 \end{cases},$$

e quando $\kappa \rightarrow \infty$, corresponde à estrutura Gaussiana

$$\gamma(h_{ij}) = \begin{cases} 0, & \text{if } h_{ij} = 0 \\ \phi_1 + \phi_2 \left[1 - \exp\left[-\left(\frac{h_{ij}}{\phi_3}\right)^2\right] \right], & \text{if } 0 < h_{ij} \leq \phi_3 \\ \phi_1 + \phi_2, & \text{if } h_{ij} > \phi_3 \end{cases}.$$

Foram considerados diferentes valores para o parâmetro de ordem para a família Matérn, $\kappa = 0, 5; 1, 0; 1, 5; 2, 0; 2, 5$ e $\kappa \rightarrow \infty$.

As funções de covariância de $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$ para um processo estacionário dependem somente da distância espacial entre eles, isto é,

$$C(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = C(Y(\mathbf{s}_i), Y(\mathbf{s}_i + h_{ij})) = C(h_{ij}),$$

em que h_{ij} é a distância Euclideana entre \mathbf{s}_i e \mathbf{s}_j . Conseqüentemente, a função de correlação entre $Y(\mathbf{s}_i)$ e $Y(\mathbf{s}_j)$ é

$$r(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = \frac{C(h_{ij})}{C(0)} = r(h_{ij}),$$

em que $C(0)$ é a variância de Y . Além disso, para um processo estacionário isotrópico, tem-se $\gamma(h_{ij}) = C(0) - C(h_{ij})$; então,

$$r(h_{ij}) = 1 - \frac{\gamma(h_{ij})}{C(0)},$$

em que r_{ij} são os elementos da matriz de correlação de trabalho $\mathbf{R}(\alpha)$, $C(0) = \phi_1 + \phi_2$. Para maiores detalhes, veja Uribe-Opazo et al. (2012) e De Bastiani et al. (2015).

4.1.2.2 Equações de estimação de quase-verossimilhança

As equações de estimação generalizadas (EEGs) geralmente são utilizadas por estatísticos como uma ferramenta para estimar parâmetros de um modelo linear generalizado com uma possível correlação desconhecida entre os resultados. EEGs representam uma extensão do método de quase-verossimilhança. Veja mais detalhes em Zeger e Liang (1986)

e Wu (2010).

Zeger e Liang (1986) apresentaram uma classe de EEGs para estimar os parâmetros de regressão e conseqüentemente estimar a variância destes parâmetros. Prentice (1988) apresentou uma segunda classe de EEGs, na qual é possível estimar conjuntamente a média e a correlação dos parâmetros. Albert e McShane (1995) estudaram EEGs para estimar a média e estrutura de correlação e utilizaram modelos de semivariograma teóricos para parametrizar a estrutura de correlação, obtendo-se assim os valores iniciais para as EEGs. Gotway e Stroup (1997) propuseram o uso de EEGs para MLGs, cuja variável resposta não apresenta distribuição normal de probabilidades. Estes autores também propuseram um preditor espacial linear.

EEGs usando a matriz de correlação de trabalho foram originalmente propostas por Zeger et al. (1985) e posteriormente destacadas nos trabalhos de Zeger e Liang (1986) e Liang e Zeger (1986) sugerindo uma combinação da estrutura de MLGs com uma estrutura da matriz de correlação. Para tal, estima-se primeiramente um valor inicial para $\hat{\alpha}$ por aproximação geoestatística, conforme descrito na Seção 4.1.2.1. Então, considera-se a matriz $\mathbf{R}(\alpha)$ fixa para estimar o vetor β .

As equações de estimação resultantes são obtidas pela diferenciação da equação de quase-verossimilhança com respeito a β e são chamadas de equações de estimação generalizadas de primeira ordem (McCullagh 1983), dadas por

$$U_{\beta} = \mathbf{D}^{\top} \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \quad (4.3)$$

em que $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\beta$ e \mathbf{V} é definida na Equação (4.2). De acordo com McCullagh e Nelder (1989), a função de quase-verossimilhança para observações dependentes é dada por

$$Q(\boldsymbol{\mu}; \mathbf{Y}, \mathbf{t}(b)) = \sigma^{-2} \int_{\mathbf{Y}}^{\boldsymbol{\mu}} (\mathbf{Y} - \mathbf{t})^{\top} \{\mathbf{V}(\mathbf{t})\}^{-1} d\mathbf{t}(b),$$

em que $\mathbf{t}(b) = \mathbf{Y} + (\boldsymbol{\mu} - \mathbf{Y})b$, para $b \in [0, 1]$.

Liang e Zeger (1986) definiram uma estimativa de β como a solução do sistema de equações quase-escore diferencial conforme descrito na Equação (4.3). Isto constitui uma generalização do método de quase-verossimilhança de Wedderburn (1974) quando os dados são correlacionados.

Como as EEGs dependem de β bem como dos parâmetros de correlação α , e não possuem solução de forma fechada, um procedimento iterativo tal como o método do escore de Fisher é necessário para se obter a solução para o parâmetro β . Uma matriz de correlação de trabalho que é próxima à verdadeira estrutura de correlação implica maior eficiência para $\hat{\beta}$ (Albert e McShane 1995).

Este estimador é consistente e assintoticamente normal (Prentice 1988; Rao et al. 2008), e ele pode ser resolvido pelo seguinte processo iterativo:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \{\hat{\mathbf{D}}^{(t)\top} \hat{\mathbf{V}}^{-1} \hat{\mathbf{D}}^{(t)}\}^{-1} \hat{\mathbf{D}}^{(t)\top} \hat{\mathbf{V}}^{-1} \{\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(t)}\},$$

em que $\widehat{\mathbf{D}} = \partial \widehat{\boldsymbol{\mu}} / \partial \boldsymbol{\beta}$ e $\widehat{\mathbf{V}} = \left[v_{\widehat{\boldsymbol{\mu}}}^{1/2} \widehat{\mathbf{R}}(\widehat{\boldsymbol{\alpha}}) v_{\widehat{\boldsymbol{\mu}}}^{1/2} \right]$, $\widehat{\mathbf{R}}(\widehat{\boldsymbol{\alpha}})$ é a matriz $n \times n$ de correlação de trabalho calculada utilizando-se a metodologia apresentada na Seção 4.1.2.1. Para maiores detalhes, veja Prentice (1988) e McCullagh e Nelder (1989).

Um estimador robusto consistente da variância (Liang e Zeger 1986) é dado por

$$\widehat{\text{var}}_{rob}(\widehat{\boldsymbol{\beta}}) = \left(\widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}} \right)^{-1} \left(\widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{Z}} \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}} \right) \left(\widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}} \right)^{-1},$$

em que $\widehat{\mathbf{Z}}$ possui elementos $\widehat{\mathbf{Z}}(i, j) = (\mathbf{Y}(s_i) - \widehat{\boldsymbol{\mu}}(s_i))(\mathbf{Y}(s_j) - \widehat{\boldsymbol{\mu}}(s_j))$. A $\widehat{\text{var}}_{rob}(\widehat{\boldsymbol{\beta}})$ é consistente mesmo quando o modelo do semivariograma não é o melhor ajuste do conjunto de dados. O valor inicial para $\widehat{\boldsymbol{\beta}}$ pode ser encontrado quando se assume a independência espacial e se ajusta um MLG com função de ligação logit (Albert e McShane 1995).

4.1.2.3 Critério de seleção para a estrutura de correlação de trabalho

Existem diversos critérios apresentados na literatura que levam em consideração a medida da discrepância esperada do modelo ajustado com o verdadeiro modelo (Akaike 1974; Bozdogan 1987; Schwarz 1978) que nos permitem escolher o melhor modelo considerado.

Pan (2001) propôs ajustes ao método *AIC* para selecionar modelos em EEGs denominado de critério quase-verossimilhança sob o modelo independente, o qual denota-se por *QIC*. Este critério pode também ser utilizado para selecionar a melhor estrutura de correlação durante a análise de EEGs (Cui 2007).

A correção do *AIC* realizada por Pan (2001) consiste de um termo de penalização para as EEGs, a saber

$$QIC = -2Q(\widehat{\boldsymbol{\mu}}; \mathbf{I}) + 2tr(\widehat{\boldsymbol{\Omega}}_{\mathbf{I}} \widehat{\mathbf{V}}_r),$$

em que $Q(\cdot)$ é a função de quase-verossimilhança, \mathbf{I} é a matriz identidade, tr significa o traço de uma matriz, e $\widehat{\boldsymbol{\Omega}}_{\mathbf{I}}$, obtido como

$$\widehat{\boldsymbol{\Omega}}_{\mathbf{I}} = \widehat{\mathbf{D}}^\top \widehat{\mathbf{V}} \widehat{\mathbf{D}},$$

e

$$\widehat{\mathbf{V}}_r = \left(\widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}} \right)^{-1} \left\{ \widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\text{var}}(\mathbf{Y}) \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}}^\top \right\} \left(\widehat{\mathbf{D}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{D}} \right)^{-1},$$

em que $\widehat{\mathbf{D}} = \partial \widehat{\boldsymbol{\mu}} / \partial \boldsymbol{\beta}$, $\widehat{\text{var}}(\mathbf{Y}) = a(\phi)^{-1} v_{\widehat{\boldsymbol{\mu}}}$ e $\widehat{\mathbf{V}}$ é dado pela Equação (4.2).

Pan (2001) observou que o critério *QIC* pode ser utilizado para selecionar a melhor estrutura de correlação e o melhor modelo ajustado por análise de EEGs. Quando comparam-se duas ou mais estruturas de correlação para uma distribuição específica e função de ligação, a que apresenta o menor valor de *QIC* deve ser a preferível (Cui 2007). Para o caso particular, em que $tr(\widehat{\boldsymbol{\Omega}}_{\mathbf{I}} \widehat{\mathbf{V}}_r) \approx tr(\mathbf{I}) = p$ ocorre, o *QIC* se reduz ao *AIC*.

Um segundo critério de correlação, dado por

$$CIC(R) = tr(\widehat{\boldsymbol{\Omega}}_{\mathbf{I}} \widehat{\mathbf{V}}_r),$$

sugere selecionar o menor valor de *CIC(R)* (Hin e Wang 2009). E, outro critério é a estatística

χ^2 (Voinov et al. 2013), definida como

$$\chi^2 = \widehat{\mathbf{S}}^\top \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{S}},$$

em que $\widehat{\mathbf{S}} = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})$ e $\widehat{\mathbf{V}}$ possui a forma da Equação (4.2).

4.1.3 Influência local para observações em predição espacial

A análise de diagnóstico de influência mostra possíveis desvios das pressuposições estabelecidas sobre o modelo, os quais também permitem detectar a existência de observações influentes que interferem desproporcionalmente nas estimativas dos parâmetros do modelo. Uma observação que dista da nuvem de dados é denominada de outlier, enquanto uma observação que poderia causar mudanças substanciais no modelo predito, quando retirada, é chamada de observação influente.

Inicialmente introduzida por Cook (1986), a influência local tem por objetivo avaliar as mudanças nos resultados das análises quando ela incorpora pequenas perturbações no modelo. Estas pequenas perturbações são feitas por um vetor $\boldsymbol{\omega}$ de dimensão $n \times 1$, denominado de vetor de perturbações, o qual é utilizado para induzir pequenas modificações na matriz de covariância, nos parâmetros ou na variável resposta.

Seja $Q_\omega(\boldsymbol{\beta}; \mathbf{Y})$ a função de quase-verossimilhança perturbada. A influência da perturbação $\boldsymbol{\omega}$ no estimador de quase-verossimilhança pode ser avaliada pelo afastamento da quase-verossimilhança,

$$QD(\boldsymbol{\omega}) = 2\{Q(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) - Q(\widehat{\boldsymbol{\beta}}_\omega; \mathbf{Y})\},$$

em que $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_\omega$ são os estimadores EEGs de $\boldsymbol{\beta}$ sob os modelos não-perturbado e perturbado, respectivamente (Jung 2008).

Assume-se que $\boldsymbol{\omega} = \boldsymbol{\omega}_0 + d\mathbf{h}$, em que $\boldsymbol{\omega}_0$ é a perturbação nula, \mathbf{h} é um vetor de comprimento 1, e d é um escalar que determina a magnitude da perturbação. A curvatura normal da superfície $(\boldsymbol{\omega}^\top, QD(\boldsymbol{\omega}))^\top$ (Cook 1986), é dada por

$$C_h = 2|\mathbf{h}^\top \ddot{\mathbf{F}}\mathbf{h}|,$$

em que $\ddot{\mathbf{F}} = \boldsymbol{\Delta}^\top \{-\ddot{\mathbf{Q}}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}; \mathbf{Y})\}^{-1} \boldsymbol{\Delta}$ na qual $\boldsymbol{\Delta} = \partial^2 Q(\boldsymbol{\beta}; \boldsymbol{\omega}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^\top$ é uma matriz $p \times n$ avaliada em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, $\ddot{\mathbf{Q}}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) = \partial^2 Q(\boldsymbol{\beta}; \mathbf{Y}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top$ é uma matriz Hessiana $p \times p$ avaliada em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ com elementos dados por

$$\frac{\partial^2 Q(\boldsymbol{\beta}; \mathbf{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \mathbf{X}^\top \left\{ \frac{\partial(\dot{\mathbf{K}}\mathbf{V}^{-1})}{\partial \boldsymbol{\beta}_j} (\mathbf{Y} - \boldsymbol{\mu}) - \dot{\mathbf{K}}\mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}_j} \right\},$$

para $j = 1, \dots, p$. Note que $\partial \boldsymbol{\mu} / \partial \boldsymbol{\beta}_j = \dot{\mathbf{K}}\mathbf{x}_j$, com $\dot{\mathbf{K}} = \text{diag}[\partial \mu_i / \partial \eta_i]$ é uma matriz $n \times n$ em que $\partial \mu_i / \partial \eta_i = m_i \pi_i (1 - \pi_i)$, $i = 1, \dots, n$, e \mathbf{x}_j é a j -ésima coluna da matriz \mathbf{X} . Ainda,

$$\frac{\partial \dot{\mathbf{K}}}{\partial \boldsymbol{\beta}_j} = \text{diag} \left[m_i \pi_i \left(1 - \frac{\pi_i^2}{1 - \pi_i} \right) (1 - 2\pi_i) x_{ij} \right],$$

é uma matriz $n \times n$ e $\partial v_{\mu_i}^{1/2} / \partial \beta_j = 1/2 \left(\sqrt{v(\mu_i)} \right)^{-1} v'(\mu_i) \partial \mu_i / \partial \eta_i x_{ij}$.

Para o esquema de perturbação no preditor linear, $\eta_\omega = \eta + \omega$, tem-se

$$\frac{\partial^2 Q(\beta; \omega)}{\partial \beta \partial \omega_i} = \mathbf{X}^\top \left\{ \frac{\partial(\dot{\mathbf{K}}_\omega \mathbf{V}_\omega^{-1})}{\partial \omega_i} (\mathbf{Y} - \boldsymbol{\mu}_\omega) - \dot{\mathbf{K}}_\omega \mathbf{V}_\omega^{-1} \frac{\partial \boldsymbol{\mu}_\omega}{\partial \omega_i} \right\},$$

em que $\dot{\mathbf{K}}_\omega \mathbf{X} = \text{diag}[h'(\eta_{i\omega})] \mathbf{X}$, é uma matriz $n \times n$ com $\partial h'(\eta_{i\omega}) / \partial \omega_i = h''(\eta_{i\omega}) x_{ij}$, e uma matriz $n \times n$

$$\frac{\partial \dot{\mathbf{K}}_\omega}{\partial \omega_i} = \text{diag}[h''(\eta_{i\omega}) \mathbf{e}_i],$$

\mathbf{e}_i é a base canônica do \mathbb{R}^n , $h''(\eta_{i\omega}) = m_i \pi_{\omega_i} (1 - \pi_{\omega_i}^2 / (1 - \pi_{\omega_i})) (1 - 2\pi_{\omega_i})$ e $\partial v_{\mu_\omega}^{1/2} / \partial \omega_i = \text{diag}[a_i \mathbf{e}_i]$ é uma matriz $n \times n$ com $a_i = 1/2 \left(\sqrt{v(\mu_{i\omega})} \right)^{-1} v'(\mu_{i\omega}) h'(\eta_{i\omega})$, $i = 1, 2, \dots, n$. Ainda,

$$\frac{\partial \boldsymbol{\mu}_\omega}{\partial \omega_i} = h'(\eta_{i\omega}) \mathbf{e}_i.$$

Finalmente,

$$\boldsymbol{\Delta} = \frac{\partial^2 Q(\beta; \omega)}{\partial \beta \partial \omega^\top} = (\Delta_1, \dots, \Delta_i, \dots, \Delta_n),$$

é uma matriz $p \times n$, em que $\Delta_i = \partial^2 Q(\beta; \omega) / \partial \beta \partial \omega_i$, $i = 1, 2, \dots, n$.

Maiores detalhes acerca da metodologia desenvolvida são apresentados no Apêndice 4.1.7.

4.1.3.1 Predição

Seja $\{\mathbf{Y}(\mathbf{s}); \mathbf{s} \in \mathbb{D} \subset \mathbb{R}^2\}$ com conjunto domínio \mathbb{D} fixo e contínuo nas localizações $\mathbf{s}_1, \dots, \mathbf{s}_n$. Suponha que queiramos obter k variáveis aleatórias adicionais nas localizações $\mathbf{s}_{0,1}, \dots, \mathbf{s}_{0,k}$ não amostradas, nomeadamente, $Y(\mathbf{s}_{0,1}), \dots, Y(\mathbf{s}_{0,k})$. Seja o vetor \mathbf{Y}_0 de dimensão $k \times 1$ de variáveis aleatórias, cujos valores devem ser preditos, e

$$E(\mathbf{Y}) = \boldsymbol{\mu}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta}, \quad E(\mathbf{Y}_0) = \boldsymbol{\mu}(\mathbf{s}_0) = \mathbf{X}_0\boldsymbol{\beta}$$

em que $\boldsymbol{\mu}(\mathbf{s})$ é o vetor de médias $n \times 1$ associado às localizações $\mathbf{s}_1, \dots, \mathbf{s}_n$, e $\boldsymbol{\mu}(\mathbf{s}_0)$ é o vetor de médias $k \times 1$ associado às localizações de predições $\mathbf{s}_{0,1}, \dots, \mathbf{s}_{0,k}$ (Gotway e Stroup 1997). Além disso,

$$\text{var} \begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}_0 \end{pmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}} & \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}_0} \\ \boldsymbol{\Sigma}_{\mathbf{Y}_0\mathbf{Y}} & \boldsymbol{\Sigma}_{\mathbf{Y}_0\mathbf{Y}_0} \end{bmatrix},$$

em que $\boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}}$ e $\boldsymbol{\Sigma}_{\mathbf{Y}_0\mathbf{Y}_0}$ são matrizes positivas definidas $n \times n$ e $k \times k$, respectivamente, e $\boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}_0}$ e $\boldsymbol{\Sigma}_{\mathbf{Y}_0\mathbf{Y}}$ são matrizes positivas $k \times n$ e $n \times k$, respectivamente.

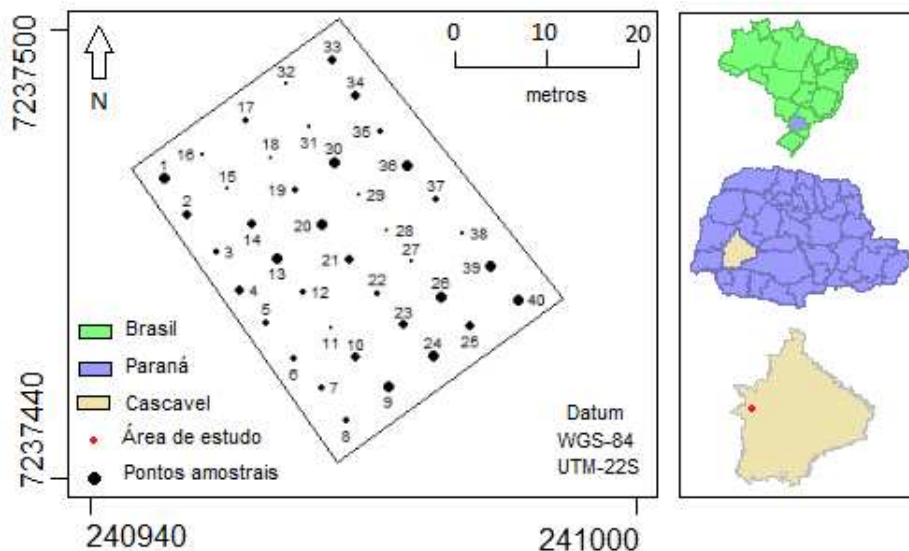
Gotway e Cressie (1993) usaram o melhor preditor linear heterogêneo como base para construir uma classe de preditores obtidos por qualquer estimador $\hat{\boldsymbol{\beta}}$, dado por

$$p(\mathbf{Y}; \mathbf{Y}_0) = \mathbf{X}_0 \hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\mathbf{Y}_0\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}}^{-1} (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}). \quad (4.4)$$

4.1.4 Aplicação ao conjunto de dados reais e resultados

O conjunto de dados é proveniente de um experimento em área comercial agrícola de 27ha, localizado no município de Cascavel, estado do Paraná no Brasil, cuja localização geográfica é aproximadamente de 24,95° latitude 53,57° longitude e altitude média de 650m. O solo é classificado como Latossolo Vermelho Distroférrico com textura argilosa (EMBRAPA 2006). O clima é classificado como temperado mesotérmico super-úmido, tipo CFA (Koeppen), e a temperatura média anual é 21°C. A Figura 2 representa a localização da área usada para o experimento.

Figura 2 Localização da área de estudo.



Para se obter cada coordenada geográfica, utilizou-se de um dispositivo receptor de sinal com sistema de posicionamento global, GPS GEOEXPLORER 3, com sistema de coordenadas Universal Transversa de Mercator (UTM). O experimento adotado é aproximadamente um retângulo de área 1000m², subdividido em 40 parcelas (5 × 8) com 234 pés de milho em cada parcela totalizando 9360 pés de milho contados, dado que a população de plantas conta com espaçamento de 0.45m entre linhas e aproximadamente 0.33m entre cada planta. Em cada parcela foi observado o número de plantas infectadas com *Spodoptera frugiperda* pelo número total de plantas por parcela. Para cada parcela, o ponto central foi georeferenciado. Os dados utilizados neste artigo estão dispostos na Tabela 12 do Apêndice 4.1.8. Realizou-se uma simplificação na magnitude dos valores de UTM originais com o objetivo de incorporar as coordenadas como possíveis covariáveis, entretanto, enfatiza-se que esta redução não altera as características originais dos dados.

O início do plantio ocorreu em 28 de Agosto e terminou em 31 de Agosto de 2015. O cultivar de milho observado no experimento foi o híbrido simples Pioneer 30F53YH, o qual é especialmente desenvolvido para as regiões Sul e Central do Brasil. Os dados amostrais foram coletados em 15 de Outubro de 2015. Para se obter a informação de quais e quantas plantas de milho foram infectadas pela lagarta do cartucho em cada parcela, escaneou-se toda a área

experimental.

Seja m o número de plantas em cada parcela e n o número de parcelas. Seja Y o vetor aleatório com distribuição binomial com índice m e parâmetro π , onde π é a razão entre o número de plantas infectadas por *Spodoptera frugiperda* pelo número total de plantas em cada parcela. Em nosso caso, tem-se $m = 234$, e um sucesso ocorre quando um pé de milho é detectado com a lagarta do cartucho dentro da parcela.

Considerou-se um modelo espacial com distribuição marginal binomial e função de ligação logit com vistas a analisar os dados dispostos na Tabela 12, c. Primeiramente, realiza-se um estudo descritivo da variável resposta Y , definida como o número de sucessos em cada uma das 40 parcelas. A Tabela 1 representa estes resultados.

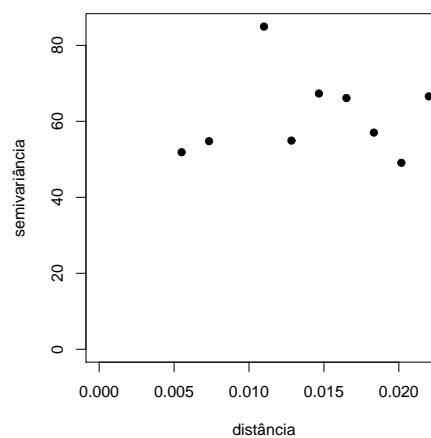
Tabela 1 Estatísticas descritivas para os dados de *Spodoptera frugiperda*.

Estatísticas	<i>Min</i>	<i>Max</i>	Q_1	Q_2	Q_3	\bar{y}	s^2	s	$CV(\%)$	A	K
Lagarta	15	46	25,75	33	38,25	32,42	60,25	7,76	23,94	-0,14	-0,87

Min : mínimo; *Max* : máximo; Q_1 : primeiro quartil; Q_2 : mediana; Q_3 : terceiro quartil; \bar{y} : média; s^2 : variância amostral; s : desvio padrão amostral; $CV(\%)$: coeficiente de variação; A : coeficiente de assimetria; K : coeficiente de curtose.

De acordo com o critério de plantio adotado pelo fazendeiro, em cada parcela existem $m_i = 234, i = 1, \dots, 40$ pés de milho. Nota-se que a parcela com maior quantidade de plantas atacadas pela *Spodoptera frugiperda* obteve 46 pés infectados. Não há parcela sem pelo menos uma planta infectada, dado que o mínimo foi de 15 infestações. O mínimo corresponde à parcela numerada por #38. Os dados aparentam simetria pela similaridade entre os valores de Q_2 e \bar{y} fato confirmado pelo coeficiente de assimetria. Pelo coeficiente de variação os dados podem ser considerados homogêneos, $CV(\%) = 23,94$.

Figura 3 Semivariograma experimental.



Como Y é um vetor de dados dependentes, a matriz de variância-covariância consiste da matriz de correlação para dados dependentes e a variância da distribuição binomial. A matriz $\hat{R}(\alpha)$ pode ser obtida a partir da escolha do melhor modelo geoestatístico entre os

modelos exponencial, Gaussiano e família Matérn com parâmetros de forma $\kappa = 1, 0; 1, 5; 2, 0$ e $2, 5$. A estrutura de dependência é obtida pelos dados do variograma representado pela Figura 3.

O ajuste do modelo geoestatístico foi realizado considerando-se os modelos teóricos e os resultados para os parâmetros α , que correspondem a cada modelo, estão dispostos na Tabela 2. O modelo Gaussiano apresenta o menor efeito pepita. O maior valor estimado para $\hat{\phi}_2$ corresponde ao modelo Gaussiano, e o parâmetro de dependência espacial, $\hat{\phi}_3$, estimado é cerca de 0,002.

Tabela 2 Estimativas para o vetor de parâmetros α considerando-se todos os dados.

Modelo geoestatístico	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Exponencial	20,4132	42,6212	0,0031
Matérn $k = 1$	23,7576	39,3556	0,0023
Matérn $k = 1, 5$	25,4911	37,0228	0,0018
Matérn $k = 2$	26,6551	35,5345	0,0014
Matérn $k = 2, 5$	28,4967	34,6964	0,0013
Gaussiano	8,1199	55,0519	0,0044

ϕ_1 : efeito pepita; ϕ_2 : patamar; ϕ_3 : parâmetro de dependência espacial.

Considerando-se $\hat{\alpha}$ e a estrutura de correlação obtida pelo modelo ajustado, conforme descrito na Tabela 2, estimou-se o vetor de parâmetros β por EEGs de acordo com o método de escore de Fisher. A partir dos dados da Figura 2, observou-se que a latitude e a longitude podem ser utilizadas como covariáveis no modelo (Gotway e Stroup 1997). Assim, depois de analisar os resíduos e os critérios dos modelos considerando-se as coordenadas como covariáveis, escolheu-se o modelo ajustado $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$, em que x_{i1} é a latitude, x_{i2} é a longitude e x_{i3} a interação entre latitude e longitude. A Tabela 3 representa os valores estimados bem como seus respectivos desvios padrões para cada modelo geoestatístico e para o modelo independente. Observa-se que os valores de β_i são similares para todos os modelos considerados. Notou-se também que os desvios padrões para o modelo independente são maiores do que para os demais modelos.

Tabela 3 Parâmetros estimados $\hat{\beta}$ e respectivos desvios padrões em parêntesis para o modelo considerando-se x_{i1} , x_{i2} and x_{i3} como covariáveis.

model	Exp	Mat1	Mat1,5	Mat2	Mat2,5	Gaus	Ind
$\hat{\beta}_0$	-17,9562	-17,9998	-17,9719	-17,9224	-17,9470	-17,9538	-17,9692
$sd(\hat{\beta}_0)$	(0,00037)	(0,00004)	(0,00011)	(0,00021)	(0,00016)	(0,00014)	(1,4657)
$\hat{\beta}_1$	276,1769	276,0800	276,0126	275,9794	275,9696	275,8586	275,8507
$sd(\hat{\beta}_1)$	(0,00116)	(0,00074)	(0,00046)	(0,00033)	(0,00032)	(0,00011)	(2,5224)
$\hat{\beta}_2$	33,5751	33,6850	33,6334	33,5294	33,5856	33,6186	33,6392
$sd(\hat{\beta}_2)$	(0,00091)	(0,00014)	(0,00026)	(0,00047)	(0,00036)	(0,00030)	(2,9516)
$\hat{\beta}_3$	-576,5038	-576,5407	-576,5769	-576,6012	-576,6012	-576,6512	-576,6531
$sd(\hat{\beta}_3)$	(0,00047)	(0,00034)	(0,00019)	(0,00012)	(0,00012)	(0,00003)	(1,2781)

Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1, 0$; Mat1,5: Matérn com $\kappa = 1, 5$; Mat2: Matérn com $\kappa = 2, 0$; Mat2,5: Matérn com $\kappa = 2, 5$; Gaus: modelo Gaussiano; e Ind: modelo independente,

Para escolher o melhor modelo ajustado à estrutura de dependência, considerou-se os critérios QIC , CIC e χ^2 conforme descritos na Seção 4.1.2.3. A Tabela 4 representa os resultados obtidos para estes critérios. O melhor modelo ajustado usando EEGs e a melhor estrutura de correlação é a que minimiza o critério. Portanto, escolhe-se a função de correlação Matérn com parâmetro de forma $\kappa = 2,0$ como a melhor estrutura para descrever a dependência espacial dos dados em estudo.

Tabela 4 Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} and x_{i3} como covariáveis.

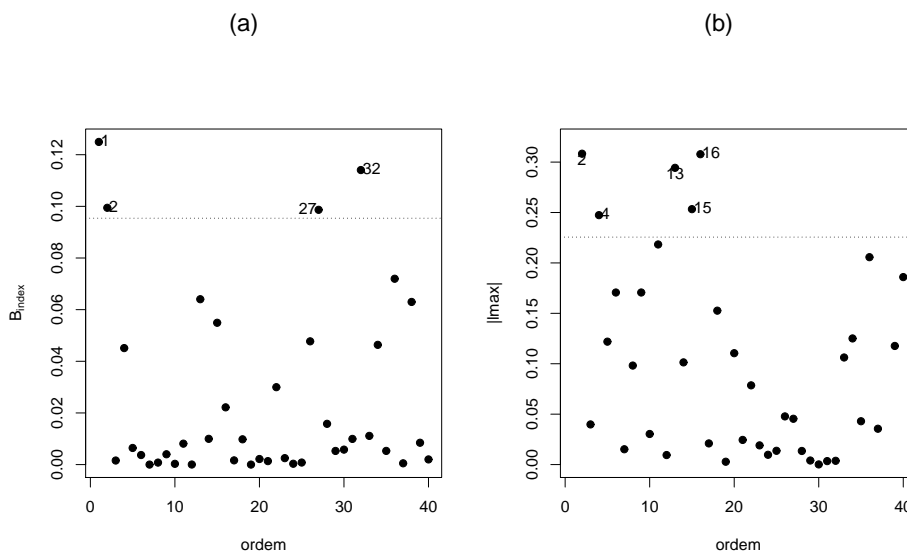
Modelo	Exp	Mat1	Mat1,5	Mat2	Mat2,5	Gaus	Ind
QIC	-94599,15	-94599,15	-94599,20	-94599,24	-94599,23	-94599,23	-94599,28
CIC	0,000011	0,000010	0,000004	0,000001	0,000002	0,000001	4,00
χ^2	82,43	85,28	83,31	79,91	82,27	87,25	76,50

Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1,0$; Mat1,5: Matérn com $\kappa = 1,5$; Mat2: Matérn com $\kappa = 2,0$; Mat2,5: Matérn com $\kappa = 2,5$; Gaus: modelo Gaussiano; e Ind: modelo independente,

O resultado para $\hat{\beta}$ implica que a probabilidade de ter-se uma planta infectada pela lagarta do cartucho é de cerca de 14%, um valor considerado elevado quando se leva em conta o grande dano que o inseto pode causar à cultura.

Com o interesse de avaliar a influência da função de quase-verossimilhança perturbada, realizou-se um estudo e influência local através da perturbação aditiva da variável resposta. A Figura 4 representa os gráficos index plot de B_i e $|l_{max}|$ para o modelo ajustado, respectivamente. Observa-se que os pontos #1, #2, #27 e #32 podem ser considerados influentes para B_i , e os pontos #2, #4, #13, #15 e #16 influentes para $|l_{max}|$.

Figura 4 Gráficos de influência B_i (a) e $|l_{max}|$ (b), do esquema de perturbação para o modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$ e estrutura de correlação Matérn com $\kappa = 2,0$.



Para verificar possíveis mudanças na inferência estatística, os potenciais pontos influentes #1, #2, #27 e #32 foram removidos do conjunto de dados e a análise realizada

novamente. Os resultados estão dispostos nas Tabelas 5-7. Em particular, a Tabela 5 representa os valores ajustados do vetor de parâmetros $\hat{\alpha}$ para cada modelo geoestatístico. Nota-se uma pequena redução nos valores ajustados quando comparados com os obtidos para o conjunto completo de dados, exceto para o parâmetro $\hat{\phi}_3$ que apresenta quase os mesmos valores para todos os modelos espaciais considerados.

Tabela 5 Estimativas para o vetor de parâmetros α sem as observações #1, #2, #27 e #32.

Modelo geoestatístico	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Exponencial	18,6951	38,5877	0,0031
Matérn $k = 1$	21,7084	35,4968	0,0022
Matérn $k = 1,5$	16,0054	40,5545	0,0015
Matérn $k = 2$	25,2940	31,8866	0,0016
Matérn $k = 2,5$	23,4850	33,5199	0,0014
Gaussiano	4,8873	52,6483	0,0044

ϕ_1 : efeito pepita; ϕ_2 : patamar; ϕ_3 : parâmetro de dependência espacial.

Utilizou-se o método de escore de Fisher para estimar $\hat{\beta}$ do conjunto de dados que considera a remoção de todos os pontos influentes, veja Tabela 6. Os valores do vetor de parâmetros $\hat{\beta}$ foram similares aos ajustados para todo o conjunto de dados completo.

Tabela 6 Parâmetros estimados $\hat{\beta}$ e, em parênteses, respectivos desvios padrões para o modelo considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem os pontos #1, #2, #27 e #32.

modelo	Exp	Mat1	Mat1,5	Mat2	Mat2,5	Gaus
$\hat{\beta}_0$	-19,1188	-19,1630	-19,0559	-19,1976	-19,1392	-19,0955
$sd(\hat{\beta}_0)$	(0,0036)	(0,0031)	(0,0039)	(0,0021)	(0,0026)	(0,0009)
$\hat{\beta}_1$	280,0635	280,0425	279,6962	280,0281	279,8991	279,8555
$sd(\hat{\beta}_1)$	(0,0330)	(0,0308)	(0,0267)	(0,0284)	(0,0271)	(0,0236)
$\hat{\beta}_2$	35,3292	35,4246	35,2609	35,4983	35,3984	35,3112
$sd(\hat{\beta}_2)$	(0,0036)	(0,0029)	(0,0050)	(0,0012)	(0,0023)	(0,0006)
$\hat{\beta}_3$	-574,5163	-574,5183	-574,6961	-574,5189	-574,5884	-574,6163
$sd(\hat{\beta}_3)$	(0,0159)	(0,0148)	(0,0130)	(0,0136)	(0,0130)	(0,0111)

Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1, 0$; Mat1,5: Matérn com $\kappa = 1, 5$; Mat2: Matérn com $\kappa = 2, 0$; Mat2,5: Matérn com $\kappa = 2, 5$ e Gaus: modelo Gaussiano,

A Tabela 7 dispõe dos critérios de seleção utilizados para escolher o melhor modelo. Observa-se que o melhor modelo é o Matérn com parâmetro de forma $\kappa = 1, 5$.

Tabela 7 Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem os pontos #1, #2, #27 e #32.

modelo	Exp	Mat1	Mat1,5	Mat2	Mat2,5	Gaus
QIC	-85100,2789	-85100,3446	-85101,7263	-85100,3888	-85100,9970	-85101,2552
CIC	0,0050	0,0048	0,0033	0,0046	0,0040	0,0036
χ^2	65,0518	66,9670	62,4963	69,6011	67,0859	69,8742

Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1, 0$; Mat1,5: Matérn com $\kappa = 1, 5$; Mat2: Matérn com $\kappa = 2, 0$; Mat2,5: Matérn com $\kappa = 2, 5$ e Gaus: modelo Gaussiano,

Uma segunda análise de influência local foi realizada consistiu na exclusão do possível ponto influente #1 do conjunto de dados com posteriores análises. A Tabela 8 representa as estimativas dos parâmetros de correlação geoestatística.

Tabela 8 Estimativas do vetor de parâmetros α sem o ponto #1.

Modelo geoestatístico	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Exponential	20,1791	42,5467	0,0032
Matérn $k = 1$	23,4119	39,3536	0,0024
Matérn $k = 1,5$	25,4167	37,1152	0,0019
Matérn $k = 2$	25,6429	35,4179	0,0013
Matérn $k = 2,5$	18,1837	43,3567	0,0012
Gaussian	5,1845	57,9202	0,0046

ϕ_1 : efeito pepita; ϕ_2 : patamar; ϕ_3 : parâmetro de dependência espacial.

A estimação de β foi realizada utilizando-se o método de escore de Fisher considerando-se a remoção do ponto influente #1 do conjunto de dados e as estimativas correspondentes estão dispostas na Tabela 9.

Tabela 9 Parâmetros estimados $\hat{\beta}$ e, em parênteses, respectivos desvios padrões para o modelo considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem o ponto #1.

modelo	Exp	Mat1	Mat1.5	Mat2	Mat2.5	Gaus
$\hat{\beta}_0$	-18.0206	-18.0724	-18.0779	-17.9750	-17.9775	-18.0672
$sd(\hat{\beta}_0)$	(0.0018)	(0.0021)	(0.0019)	(0.0003)	(0.0004)	(0.0016)
$\hat{\beta}_1$	278.0948	277.9880	277.8713	277.6188	277.6210	277.5057
$sd(\hat{\beta}_1)$	(0.0044)	(0.0041)	(0.0043)	(0.0060)	(0.0058)	(0.0048)
$\hat{\beta}_2$	33.3488	33.4776	33.5087	33.3300	33.3353	33.5478
$sd(\hat{\beta}_2)$	(0,0025)	(0,0032)	(0,0028)	(0,0007)	(0,0005)	(0,0022)
$\hat{\beta}_3$	-575,6127	-575,6528	-575,7055	-575,8399	-575,8384	-575,8755
$sd(\hat{\beta}_3)$	(0,0023)	(0,0022)	(0,0023)	(0,0028)	(0,0027)	(0,00240)

Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1, 0$; Mat1,5: Matérn com $\kappa = 1, 5$; Mat2: Matérn com $\kappa = 2, 0$; Mat2,5: Matérn com $\kappa = 2, 5$ e Gaus: modelo Gaussiano,

A Tabela 10 disponibiliza os valores para os critérios de escolha do melhor modelo ajustado. Observa-se que o modelo Matérn com parâmetro de forma $\kappa = 2, 0$ foi o melhor modelo espacial.

Tabela 10 Critérios para escolha do melhor modelo ajustado considerando-se x_{i1} , x_{i2} e x_{i3} como covariáveis sem o ponto #1.

modelo	Exp	Mat1	Mat1,5	Mat2	Mat2,5	Gaus
QIC	-92262,8124	-92262,8714	-92262,9461	-92263,0937	-92263,0892	-92263,0873
CIC	0,00042	0,00035	0,00029	0,00022	0,00022	0,00017
χ^2	79,8276	82,8729	83,3163	76,2425	76,8496	87,9648

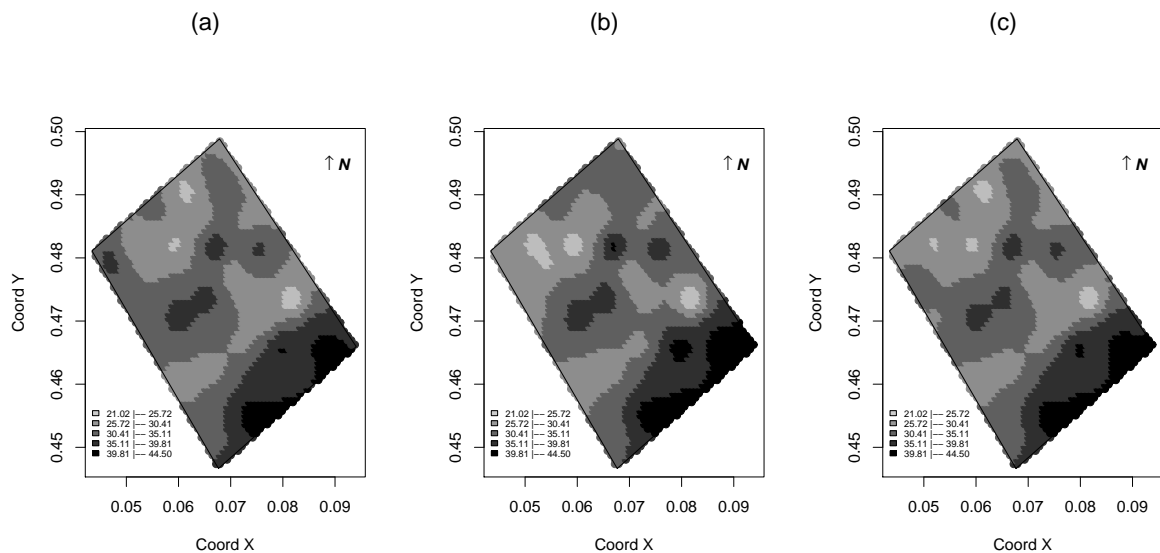
Exp: modelo exponencial; Mat1: Matérn com $\kappa = 1, 0$; Mat1,5: Matérn com $\kappa = 1, 5$; Mat2: Matérn com $\kappa = 2, 0$; Mat2,5: Matérn com $\kappa = 2, 5$ e Gaus: modelo Gaussiano,

A distribuição espacial da *Spodoptera frugiperda* em toda a região pode ser obtida por predição espacial linear para valores não amostrados. Utilizou-se o preditor definido pela

Equação (4.4) para prever estes pontos não amostrados. A área experimental foi dividida em 2700 pixels, com $0,7 \times 0,7m$ cada. A Figura 5 representa o número de insetos na área experimental sob os casos considerados.

A Figura 5(a) representa o mapa predito de infestação do número de *Spodoptera frugiperda* considerando-se todos os 40 valores amostrados. A Figura 5(b) representa o mapa obtido quando considera-se a remoção dos pontos influentes #1, #2, #27 e #32 e a Figura 5(c) o mapa predito quando da remoção do ponto influente #1. Para cada caso, utilizou-se a melhor função de correlação espacial selecionada. Os mapas preditos permitem observar a dispersão da lagarta do cartucho ao longo da área estudada. O número médio de lagartas preditas varia no intervalo de amplitude 21,02 a 44,50, cujos valores são próximos aos reais valores amostrados. Observa-se uma maior concentração da lagarta na região inferior e que a remoção dos pontos influentes altera os mapas.

Figura 5 Mapas preditos para todo o conjunto de dados (a), removendo-se os pontos #1, #2, #27 e #32 (b), e removendo-se o ponto #1 (c).



4.1.5 Simulações

Com vistas a analisar a performance da metodologia proposta para detectar pontos atípicos, conduziu-se um estudo de simulação considerando a grade de nossos dados e uma grade regular de distâncias 10 unidades entre os pontos horizontais e 100 unidades entre os pontos verticais. Ambos os casos estão baseados no conjunto de dados reais. Gerou-se um modelo binomial, $Y_i \sim Bin(234, \pi)$, $i = 1, \dots, 40$, com correlação espacial linear com $\mathbf{X} = \mathbf{1}_n$ e função de ligação logit para $\pi = 0,05; 0,14; \text{ e } 0,25$, considerando-se o modelo geoestatístico Matérn com parâmetro de forma $\kappa = 2$ e $\alpha = (26; 35; 0,01)^\top$. Após gerar o modelo binomial, perturbou-se uma observação para criar um outlier como $Y_1 \leftarrow Y_1 + \{[m\pi(1-\pi)]^{0,5}\} \delta$, em que $\delta = 0, 2, 4, 8$ e 16 . Por este esquema de perturbação, leva-se em conta a variabilidade dos

dados para o modelo sem covariáveis. Para cada cenário, foram geradas 1000 réplicas e um contador que adiciona 1 se a metodologia de influência local identifica a observação influente gerada, ou soma 0 em caso contrário. Esta metodologia nos permite verificar o comportamento do método quando a probabilidade de uma planta infectada pela lagarta do cartucho varia.

A Tabela 11 resume os percentuais de detecção do outlier utilizando-se a medida de influência de segunda ordem e nota-se a eficiência da metodologia. O número de identificações corretas do outlier cresce quando aumentado δ e/ou a probabilidade de uma planta estar infectada. Mesmo quando δ é pequeno, o método é capaz de identificar corretamente o ponto influente. Os resultados são muito similares para ambos os grids considerados.

Tabela 11 Percentual de vezes que a correta observação influente foi identificada sobre 1000 réplicas para o grid dos dados e para o grade regular.

	Grade dos dados			Grade regular		
	π			π		
δ	0,05	0,14	0,25	0,05	0,14	0,25
0	6,3	4,6	6,7	6,2	4,6	6,7
2	41,5	45,4	55,2	42,3	45,9	54,9
4	94,1	96,7	98,6	94,3	96,6	98,8
8	100	100	100	100	100	100
16	100	100	100	100	100	100

A rotina computacional foi escrita no software R versão 3.3.3 (R Core Team 2017).

4.1.6 Conclusão

Propôs-se um método para analisar um conjunto de dados reais de uma área agrícola utilizando-se a distribuição binomial. Discutiu-se metodologia geoestatística e EEGs para estimar os parâmetros, e, para a escolha da melhor estrutura de correlação que melhor descreve os dados, utilizaram-se os critérios QIC e CIC. O foco principal foi avaliar o efeito de possíveis outliers na variável resposta.

O estudo de influência local através de perturbação aditiva da variável resposta identificou potenciais pontos influentes, parcelas #1, #2, #27 e #32, as quais permitiram avaliar a estabilidade da função de quase-verossimilhança. Dois estudos que consistiram da remoção dos pontos influentes foram realizados. Primeiramente, removeram-se todos os pontos identificados como influentes, e posteriormente, removeu-se somente o ponto #1 e realizaram-se as análises novamente. Observa-se que a remoção de todos os pontos influentes do conjunto de dados alterou o mapa de predição mais do que a remoção do único ponto, #1. Portanto, a presença de pontos influentes afeta a probabilidade da dispersão da *Spodoptera frugiperda*. De acordo com a Figura 5, quando tem-se o mapa baseado na análise sem todas as observações influentes, nota-se que há mais pixels pertencentes às classes do mapa de menores valores preditos. A metodologia utilizada neste artigo abre a possibilidade para os fazendeiros visualizarem a dispersão da lagarta do cartucho ao longo da área de

estudo e proceder controle local da praga.

O estudo de simulação permitiu avaliar a eficiência da metodologia de influência local na detecção de observações que potencialmente podem representar influências nos resultados finais.

4.1.7 Apêndice A

A quase-verossimilhança para respostas dependentes é dada por

$$\frac{\partial Q(\beta; \mathbf{Y})}{\partial \beta} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right)^\top \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}),$$

em que $\partial \boldsymbol{\mu} / \partial \beta_j = \dot{\mathbf{K}} \mathbf{X}$, com $\dot{\mathbf{K}} = \text{diag} [\partial \mu_i / \partial \eta_i]$ uma matriz $n \times n$. Então

$$\frac{\partial Q(\beta; \mathbf{Y})}{\partial \beta} = \mathbf{X}^\top \dot{\mathbf{K}} \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}).$$

Estamos interessados em

$$\begin{aligned} \frac{\partial^2 Q(\beta; \mathbf{Y})}{\partial \beta \partial \beta_j} &= \frac{\partial (\mathbf{X}^\top \dot{\mathbf{K}} \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}))}{\partial \beta_j} \\ &= \mathbf{X}^\top \left\{ \frac{\partial (\dot{\mathbf{K}} \mathbf{V}^{-1})}{\partial \beta_j} (\mathbf{Y} - \boldsymbol{\mu}) - \dot{\mathbf{K}} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \beta_j} \right\}, \end{aligned}$$

para $j = 1, \dots, p$. Note que $\partial \boldsymbol{\mu} / \partial \beta_j = \begin{bmatrix} g'(\eta_1) \mathbf{x}_{1j} \\ g'(\eta_2) \mathbf{x}_{2j} \\ \vdots \\ g'(\eta_n) \mathbf{x}_{nj} \end{bmatrix}$ com $g'(\eta_i) = \partial \mu_i / \partial \eta_i$ e \mathbf{x}_j é a j -ésima

coluna da matriz \mathbf{X} . Além disso,

$$\frac{\partial (\dot{\mathbf{K}} \mathbf{V}^{-1})}{\partial \beta_j} = \frac{\partial \dot{\mathbf{K}}}{\partial \beta_j} \mathbf{V}^{-1} + \dot{\mathbf{K}} \frac{\partial \mathbf{V}^{-1}}{\partial \beta_j},$$

em que $\partial (\mathbf{V}^{-1}) / \partial \beta_j = -\mathbf{V}^{-1} \left(\partial v_{\boldsymbol{\mu}}^{1/2} / \partial \beta_j \mathbf{R} v_{\boldsymbol{\mu}}^{1/2} + \mathbf{R} v_{\boldsymbol{\mu}}^{1/2} \partial v_{\boldsymbol{\mu}}^{1/2} / \partial \beta_j \right) \mathbf{V}^{-1}$, com $\partial v_{\mu_i}^{1/2} / \partial \beta_j = 1/2 \left(\sqrt{v(\mu_i)} \right)^{-1} v'(\mu_i) \partial \mu_i / \partial \eta_i \mathbf{x}_{ij}$. $\partial \dot{\mathbf{K}} / \partial \beta_j = \text{diag} [m_i \pi_i (1 - \pi_i^2 / (1 - \pi_i)) (1 - 2\pi_i) \mathbf{x}_{ij}]$ é uma $n \times n$ matriz, e $\partial \mu_i / \partial \eta_i = m_i \exp(\eta_i) / (1 + \exp(\eta_i))^2 = m_i \pi_i (1 - \pi_i)$.

Por outro lado, para o modelo perturbado temos $\boldsymbol{\eta}_\omega = g(\boldsymbol{\pi}_\omega) = \boldsymbol{\eta} + \boldsymbol{\omega}$, e

$$\boldsymbol{\Delta} = \frac{\partial^2 Q(\beta; \boldsymbol{\omega})}{\partial \beta \partial \boldsymbol{\omega}^\top}.$$

Assim, a quase-verossimilhança perturbada é dada por

$$\frac{\partial Q(\beta; \boldsymbol{\omega})}{\partial \beta} = \left(\frac{\partial \boldsymbol{\mu}_\omega}{\partial \beta_j} \right)^\top \mathbf{V}_\omega^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\omega),$$

em que $\partial \mu_{i\omega} / \partial \beta_j = \partial h(\eta_i + \omega_i) / \partial \beta_j = h'(\eta_i + \omega_i) \mathbf{x}_{ij}$. Logo,

$$\frac{\partial^2 Q(\beta; \boldsymbol{\omega})}{\partial \beta \partial \omega_i} = \mathbf{X}^\top \left\{ \frac{\partial (\dot{\mathbf{K}}_\omega \mathbf{V}_\omega^{-1})}{\partial \omega_i} (\mathbf{Y} - \boldsymbol{\mu}_\omega) - \dot{\mathbf{K}}_\omega \mathbf{V}_\omega^{-1} \frac{\partial \boldsymbol{\mu}_\omega}{\partial \omega_i} \right\},$$

para $i = 1, \dots, n$, $\dot{\mathbf{K}}_{\omega} \mathbf{X} = \text{diag}[h'(\eta_{i\omega})] \mathbf{X}$ é uma matriz $n \times n$ com $\partial h'(\eta_{i\omega})/\partial \omega_i = h''(\eta_{i\omega}) x_{ij}$, e uma matriz $n \times n$

$$\frac{\partial \dot{\mathbf{K}}_{\omega}}{\partial \omega_i} = \text{diag}[h''(\eta_{i\omega}) \mathbf{e}_i],$$

em que \mathbf{e}_i é a base canônica do \mathbb{R}^n e $h''(\eta_{i\omega}) = m_i \pi_{\omega_i} (1 - \pi_{\omega_i}^2 / (1 - \pi_{\omega_i})) (1 - 2\pi_{\omega_i})$.

$$\frac{\partial \dot{\mathbf{K}}_{\omega} \mathbf{V}_{\omega}^{-1}}{\partial \omega_i} = \frac{\partial \dot{\mathbf{K}}_{\omega}}{\partial \omega_i} \mathbf{V}_{\omega}^{-1} + \dot{\mathbf{K}}_{\omega} \frac{\partial \mathbf{V}_{\omega}^{-1}}{\partial \omega_i},$$

com $\partial \mathbf{V}_{\omega}^{-1} / \partial \omega_i = -\mathbf{V}_{\omega}^{-1} \left(\partial v_{\mu\omega}^{1/2} / \partial \omega_i \mathbf{R} v_{\mu\omega}^{1/2} + v_{\mu\omega}^{1/2} \mathbf{R} \partial v_{\mu\omega}^{1/2} / \partial \omega_i \right) \mathbf{V}_{\omega}^{-1}$, e $\partial v_{\mu\omega}^{1/2} / \partial \omega_i = \text{diag}[a_i \mathbf{e}_i]$ é uma $n \times n$ matriz com $a_i = 1/2 \left(\sqrt{v(\mu_{i\omega})} \right)^{-1} v'(\mu_{i\omega}) h'(\eta_{i\omega})$, $i = 1, 2, \dots, n$. Ainda,

$$\frac{\partial \boldsymbol{\mu}_{\omega}}{\partial \omega_i} = h'(\eta_{i\omega}) \mathbf{e}_i.$$

Finalmente, $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_i, \dots, \Delta_n)$, uma matriz $p \times n$ em que,

$$\Delta_i = \frac{\partial^2 Q(\boldsymbol{\beta}; \boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \omega_i}, \quad i = 1, 2, \dots, n.$$

4.1.8 Apêndice B

A Tabela (12) dispõe dos dados amostrados.

Tabela 12 Dados da *Spodoptera frugiperda*.

Parcela	lat	long	Y	m	Parcela	lat	long	Y	m
1	47,275	480,046	40	234	21	69,078	470,491	35	234
2	50,036	475,829	38	234	22	72,245	466,457	26	234
3	53,363	471,472	30	234	23	75,356	462,791	37	234
4	56,189	466,829	36	234	24	78,830	459,169	40	234
5	59,234	463,007	29	234	25	83,207	462,717	36	234
6	62,442	458,875	28	234	26	79,799	466,186	44	234
7	65,720	455,373	33	234	27	76,385	470,245	21	234
8	68,738	451,575	33	234	28	73,438	473,885	25	234
9	73,568	455,561	45	234	29	70,058	478,101	25	234
10	69,780	458,992	34	234	30	67,277	482,007	45	234
11	66,955	462,530	25	234	31	64,372	486,131	25	234
12	63,524	466,623	31	234	32	61,577	491,115	20	234
13	60,441	470,723	41	234	33	67,038	493,902	35	234
14	57,620	474,690	35	234	34	69,830	489,696	37	234
15	54,642	478,779	24	234	35	72,722	485,605	28	234
16	51,885	482,781	22	234	36	75,731	481,607	41	234
17	56,874	486,774	32	234	37	79,194	477,568	30	234
18	59,711	482,462	21	234	38	82,247	473,602	15	234
19	62,632	478,668	30	234	39	85,591	469,695	39	234
20	65,808	474,604	40	234	40	88,916	465,784	46	234

Y : número de plantas infectadas em cada parcela, m : número de plantas em cada parcela, lat: latitude e long: longitude.

4.1.9 Referências

- Akaike, H. (1974), "A new look at statistical model identification," *IEEE Transactions on Automatic Control*, 19 (6), 716 - 723.
- Albert, P. S., McShane, L. (1995), "A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data," *Biometrics*, 51 (2), 627 - 638.
- Belay, D. K., Huckaba, R. M., Foster, J. E. (2012), "Susceptibility of the fall armyworm, *Spodoptera Frugiperda*, (Lepidoptera: Noctuidae), at Santa Isabel, Puerto Rico, to different insecticides," *The Florida Entomologist*, 95 (2), 476 - 478.
- Bozdogan, H., (1987), "Model selection and akaike's information criterion (aic): The general theory and its analytical extensions," *Psychometrica*, 52 (3), 345 - 370.
- Cadigan, N., Farrell, P. (2002), "Generalized local influence with applications to fish stock cohort analysis," *Applied Statistics*, 51 (4), 469 - 483.
- Cook, R. D. (1986), "Assessment of local influence," *Journal of the Royal Statistical Society, B* 48, 133 - 169.
- Cressie, N. (1993), *Statistics for Spatial Data* John Wiley, New York.
- Cruz, I., Turpin, F. T. (1983), "Yield impact of larval infestation of the fall armyworm *spodoptera irugiperda* (J. E. Smith) to mid-whorl growth stage of corn," *Journal of Economic Entomology*, 76 (5), 1052 - 1054.
- Cui, J. (2007), "QIC program and model selection in gee analyses," *The Stata Journal*, 7 (2), 209 - 220.
- De Bastiani, F., Cysneiros, A. H. M. A., Uribe-Opazo, M. A., and Galea, M. (2015), "Influence diagnostics in elliptical spatial linear models," *TEST*, 24 (2), 322 - 340.
- EMBRAPA. (2006), *Sistema brasileiro de classificação de solos* EMBRAPA.
- Garcia-Papani, F., Uribe-Opazo, M. A., Leiva, V., Aykroyd, R. G. (2017), "Birnbbaum-saunders spatial modelling and diagnostics applied to agricultural engineering data," *Stochastic Environmental Research and Risk Assessment*, 31 (1), 105 - 124.
- Goergen, G., Kumar, P. L., Sankung, S. B., Togola, A., Tamò, M. (2017), "First Report of Outbreaks of the Fall Armyworm *Spodoptera frugiperda* (J E Smith) (Lepidoptera, Noctuidae), a New Alien Invasive Pest in West and Central Africa," *PLoS One*, 11 (10), 1 - 9.
- Gosho, M., Hamada, C., Yoshimura, I. (2011), "Criteria for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data" *Communications in Statistics - Theory and methods*, 40 (21), 3839 - 3856.
- Gotway, C. A., Cressie, N. (1993), "Improved multivariate prediction under a general linear

model," *Journal of Multivariate Analysis*, 45 (1), 56 - 72.

Gotway, C. A., Stroup, W. W. (1997), "A generalized linear model approach to spatial data analysis and prediction," *Journal of Agricultural, Biological and Environmental Statistics*, 2 (2), 157 - 178.

Haran, M. (2011), Gaussian random fields models for spatial data. In: Brooks, S., Gelman, A., Jones, G. L., Meng, X.-L. (Eds.), *Handbook of Markov Chain Monte Carlo* Chapman and Hal/CRCI, New York, Ch. 18, pp. 449 - 473.

Hin, L. Y., Wang, Y. G. (2009), "Working-correlation-structure identification in generalized estimating equations," *Statistics in Medicine*, 28 (4), 642 - 658.

Jung, K-M. (2008), "Local influence in generalized estimating equations," *Scandinavian Journal of Statistics*, 35 (1), 286 - 294.

Krippendorff, K. (1980), *Content Analysis: an Introduction to its Methodology* Sage, Thousand Oaks.

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

Matérn, B. (1960), "Spatial variation," *Meddelanden fren Statens Skogséforskningsinstitut*, 49 (5).

Matheron, G. (1963), "Principles of geostatistics," *Economic Geology*, 58 (1), 1246 - 1266.

McCullagh, P. (1983), "Quasi-likelihood functions," *Annals Statistic*, 11 (1), 59 - 67.

McCullagh, P., Nelder, J. A. (1989), *Generalized Linear Models*, 2ed, Chapman e Hall, London.

Militino, A., Palacios, M., and Ugarte, M. (2006), "Outliers detection in multivariate spatial linear models," *Journal of Statistical Planning and Inference*, 136 (1), 125 - 146.

Nelder, J. A., Wedderburn, R. W. M. (1983), "Generalized linear models," *Journal of the Royal Statistical Society*, 135 (3), 370 - 384.

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

Prentice, R. L. (1988), "Correlated binary regression with covariates specific to each binary observation," *Biometrics*, 44 (4), 1033 - 1048.

Prentice, R. L., Zhao, L. P. (1991), "Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses," *Biometrics*, 47 (3), 825 - 839.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

- Rao, C. R., Toutenburg, H., Shalabh, Heumann, C. (2008), *Linear Models and generalizations: Least squares and alternatives*, 3ed, Springer, Berlin.
- Schabenberger, O., Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Chapman Hall, London.
- Schwarz, G. (1978), "Estimating the dimensional of a model," *Annals of Statistics*, 6 (2), 461 - 464.
- Uribe-Opazo, M., Borssoi, J., and Galea, M. (2012), "Influence diagnostics in Gaussian spatial linear models," *Journal of Applied Statistics*, 39 (3), 615 - 630.
- Voinov, V., Nikulin, M., Balakrishnan, N. (2013), *Chi-Squared goodness of fit tests with applications*, Elsevier, Waltham.
- Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models and the gauss-newton method," *Biometrika*, 61 (3), 439 - 447.
- Wu, L. (2010), *Mixed effects models for complex data*, Chapman Hall, Boca Raton.
- Yu, S.J (1991), "Insecticide resistance in the fall armyworm, *Spodoptera frugiperda* (J. E. Smith)," *Pesticide Biochemistry and Physiology*, 39 (1), 84 - 91.
- Zeger, S., Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, 42 (1), 121 - 130.
- Zeger, S., Liang, K.-Y., Self, S. G. (1985), "The analysis of binary longitudinal data with time-independent covariates," *Biometrika*, 72 (1), 31 - 38.

4.2 ARTIGO 2: Métodos estatísticos para identificar anisotropia na distribuição espacial da *Spodoptera frugiperda*

Resumo: O milho é um produto agrícola muito importante em que todas as partes da planta podem ser utilizadas. Por esta razão, é de grande interesse que a produtividade dos grãos seja a maior possível. Entretanto, algumas pragas podem causar grande dano à produtividade do milho, tal como a lagarta do cartucho (cujo nome científico é *Spodoptera frugiperda*) que impede a planta de crescer de maneira regular. Dado que o uso indiscriminado de pesticidas podem causar uma crescente resistência dos insetos além de danos ambientais, é importante estimar possíveis áreas e regiões dominantes em que o inseto pode se propagar. O principal objetivo deste artigo foi estudar a distribuição da lagarta do cartucho em uma área comercial agrícola na região Sul do Brasil. Para tal, considera-se um conjunto de dados da localização de cada planta de milho atacada pelo inseto. Em particular, assume-se que as localizações espaciais dadas pelas coordenadas geográficas constituem um padrão espacial pontual de um processo pontual de Poisson estacionário. Com vistas a detectar possíveis direções dominantes na distribuição da infestação da lagarta do cartucho, estudou-se as características de anisotropia dos dados a partir de técnicas de análise de padrões pontuais espaciais de segunda ordem, tal como os testes K direcional, baseados em ondaletas e contagem de quadrados. Todos os resultados mostram que a distribuição espacial da lagarta pode seguir um processo pontual de Poisson agrupado com evidente anisotropia, principalmente devido à forma da área experimental.

Palavras-chave: K direcional, padrão espacial pontual, processos cluster, teste de ondaletas.

4.2.1 Introdução

A lagarta do cartucho é uma das pragas que causam mais danos à cultura do milho, cujo prejuízo pode chegar a cerca de 34% da produção de grãos (Cruz et al., 1999). Os insetos se alimentam do miolo da planta desde sua emergência até a formação de espigas, e se não forem controlados podem causar a morte da planta. Mundialmente, o milho é uma das culturas de cereais mais valiosas, posto que é utilizado tanto para consumo humano quanto animal. O milho também é utilizado na produção de diversos produtos não alimentícios. É cultivado em grande parte do mundo, e o Brasil é um grande produtor mundial deste grão. Dada sua utilidade, é de extrema importância levar em consideração seu potencial produtivo, resistência às doenças, tipos de solo além das condições climáticas a fim de se otimizar o rendimento dos grãos.

Em geral, o controle da praga é realizado com a aplicação de pesticidas químicos, os quais além de poluírem o meio ambiente também induzem ao aumento da resistência do inseto. Por esta razão, técnicas de controle do inseto são bastante pesquisadas. Por exemplo,

Este artigo foi estruturado seguindo as normas da *Spanish Journal of Agricultural Research* e foi aceito para publicação com correções. As normas estão no Anexo 6.2.

Sena Jr et al. (2003) desenvolveram um algoritmo para identificar plantas de milho danificadas pela lagarta do cartucho usando imagens digitais coloridas. Yu et al. (2003) estudaram a capacidade de desintoxicação em corpos gordurosos larvais, dos intestinos médio e abdômen adulto das larvas para resistência e susceptibilidade a inseticidas da lagarta do cartucho. Os autores concluíram que a resistência ao inseticida observado ocorreu devido a múltiplos mecanismos de resistência. Niu et al. (2013) estudaram a resistência da lagarta do cartucho a certos tipos de proteína com vistas a propor um híbrido resistente à lagarta. Na Flórida e Lousiana, Yang et al. (2013) determinaram a susceptibilidade da *Spodoptera frugiperda* para um híbrido de milho contendo Agrisure[®]Viptera[™] 3111.

Métodos estatísticos baseados em processos pontuais espaciais produzem ferramentais úteis para pontos aleatoriamente distribuídos sobre o plano ou espaço em muitos campos do conhecimento, tais como ecologia, biologia, epidemiologia, sismologia, arqueologia, astronomia, e geografia. Entretanto, a maioria das análises são feitas assumindo-se isotropia do processo sem uma real confirmação (Guan et al., 2006). Este fato, geralmente, deve-se à falta de estudos que consideram a investigação da isotropia nestes tipos de dados. Na Ecologia, Sismologia e Biologia, Guan et al. (2006), Mateu e Nocolis (2015) e Rajala et al. (2016) apresentam alguns estudos para anisotropia. O primeiro propôs um teste formal para isotropia baseado na normalidade assintótica conjunta da função de intensidade de segunda ordem amostral. Mateu e Nocolis (2015) propuseram um teste baseado em ondaletas, e Rajala et al. (2016) desenvolveram um método não paramétrico em dois estágios para quantificar a anisotropia geométrica para padrões comprimidos ou esticados.

O uso de padrão espacial pontual na agricultura foi considerado no trabalho de Spósito et al. (2007), que estudaram a doença da mancha negra em plantação de laranjas no Brasil. Os autores avaliaram o padrão pontual das árvores infectadas de acordo com três diferentes métodos: o índice de dispersão binomial, a função K de Ripley e teste de Monte Carlo, e concluíram que a doença apresenta um comportamento agrupado na região sob estudo. Ribeiro Jr et al. (2009) utilizaram processos pontuais para detectar e modelar padrões específicos na ocorrência da doença tripes em plantação de cebolas. Os autores utilizaram o teste de Mantel para detectar o tipo de padrão pontual e posteriormente descrevê-lo como um processo espacial misto de Poisson com componentes geoestatísticos aleatórios. Eles também obtiveram mapas da predição dos níveis de susceptibilidade à infestação na área sob estudo.

Neste trabalho, estudou-se o espalhamento da lagarta do cartucho ao longo de uma área experimental agrícola no estado do Paraná, Brasil, comparando métodos de anisotropia por meio de padrões espaciais pontuais. O principal objetivo foi examinar a distribuição espacial dentro da área de estudo, investigando possíveis direções dominantes da infestação da lagarta do cartucho, e se o padrão apresenta características de agrupamento, inibição ou espalhamento aleatório. Este conhecimento possibilita ao agricultor um manejo de forma mais localizada da cultura quando se consideram as características da dispersão do inseto dentro da área em consórcio com técnicas de agricultura de precisão. A falta de literaturas que

incorporam processos pontuais espaciais e agricultura motivou este trabalho.

4.2.2 Materiais e métodos

O conjunto de dados

O experimento foi realizado em uma área agrícola comercial de 27ha, localizada na cidade de Cascavel, estado do Paraná, Brasil, cuja localização geográfica é de aproximadamente $24,95^\circ$ de latitude e $53,57^\circ$ de longitude, com altitude média de 650m. O solo é classificado como Latossolo Vermelho Distroférico com textura argilosa (EMBRAPA 2006). O clima é classificado como temperado mesotérmico super-úmido tipo CFA (Koeppen), e a temperatura média anual é 21°C . Uma representação do local de realização do experimento está disposta na Fig. 6.

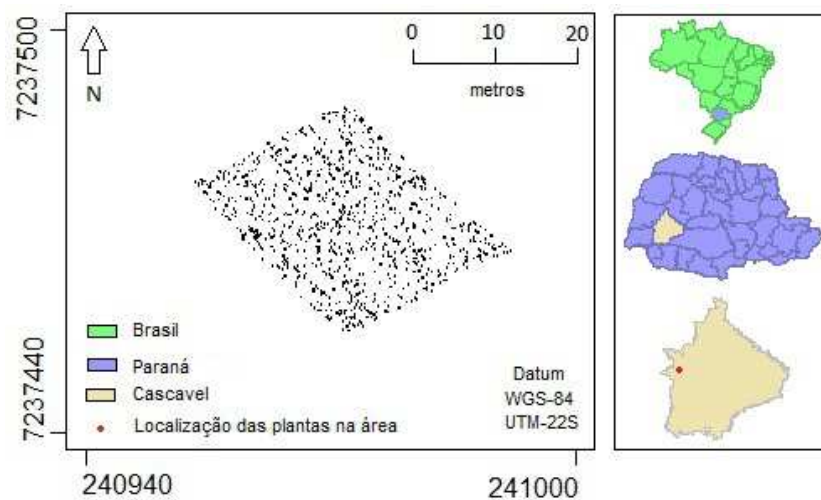


Figura 6 Área experimental em Cascavel, estado do Paraná (sul do Brasil).

Para a coleta dos dados, considerou-se como um possível evento cada planta de milho infectada pela lagarta do cartucho na área experimental durante o ano-safra 2015/2016. Para cada planta infectada observada, foram registradas as coordenadas geográficas como um evento de interesse, as quais foram obtidas utilizando-se o sistema de posicionamento global GEOEXPLORE 3 GPS com sistema de coordenadas UTM.

A área de estudo, representada pela Fig. 7, inclui 9360 plantas de milho, das quais 1303 estavam infectadas pela *Spodoptera frugiperda*. A área foi cultivada com a variedade híbrida Pioneer 30F53YH, desenvolvida pela companhia DuPont Pioneer, sob sistema de plantio direto com uma população de 82222 por hectare e fertilização 1095kg, que consiste de 8% de nitrogênio, 20% de fósforo e 18% potássio. O plantio teve início em 28 de agosto e término em 31 de agosto de 2015 e foi realizado com espaçamento entre linhas de 0,45m e entre plantas de aproximadamente 0,33m. A amostragem foi realizada no dia 15 de outubro de 2015, período em que a cultura estava no estágio vegetativo.

Análise estatística da anisotropia

Um padrão espacial pontual é um mecanismo estocástico que gera um conjunto

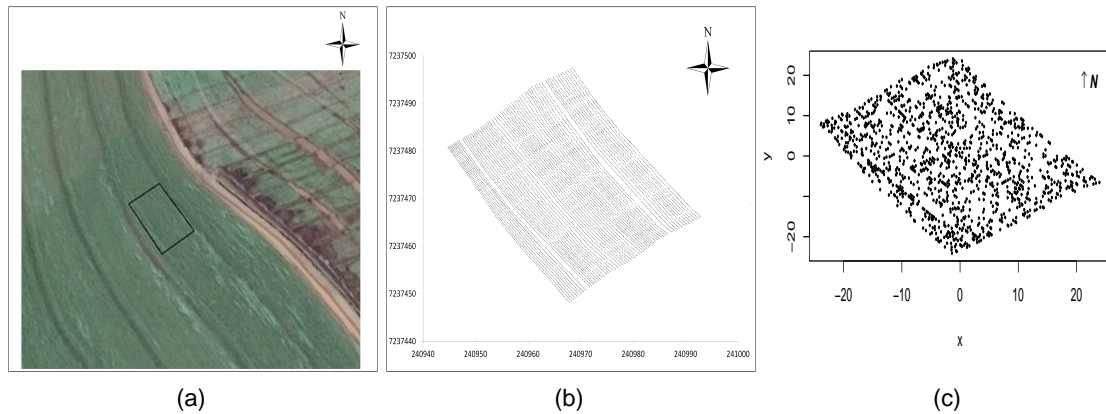


Figura 7 Imagem de satélite da área experimental com o polígono da área de estudo (a), localizações das plantas de milho (b), e localização das plantas infectas pela *Spodoptera frugiperda* (c).

contável de eventos x_i no plano (Diggle, 2003). Os eventos podem estar dispostos na região do plano de três formas, nomeadamente, completamente aleatório espacial (CAE), conforme representado pela Fig. 8a, regular como na Fig. 8b, e aglomerado conforme Fig. 8c. Em análise pontual espacial, o principal objetivo é identificar possíveis propriedades e formas de relacionamento do fenômeno sob estudo, levando-se em conta suas localizações espaciais, através das coordenadas geográficas.

Diz-se que um processo pontual espacial é estacionário sob translação se a distribuição não se altera quando a origem do conjunto índice é transladada (Illian et al., 2008; Diggle, 2003; Ripley, 1981), e é dito ser isotrópico se sua distribuição é invariante sob rotações em torno da origem. Em caso contrário, o processo é chamado anisotrópico. A suposição de estacionariedade simplifica drasticamente as estatísticas para padrões pontuais (Baddeley et al., 2006) porque a chance de se observar alguma configuração pontual em uma localização específica é independente da localização. Usualmente, a suposição de isotropia é feita por uma simples interpretação e facilidade das análises (Guan et al., 2006).

Em outras palavras, estacionariedade implica que é possível estimar algumas propriedades do processo a partir de uma única realização na região A , uma vez que estas propriedades são as mesmas em diferentes, mas geometricamente similares, subregiões de A . Isotropia significa que as propriedades são invariantes às diversas direções (Mateu e Nicolis, 2015). No que se segue, assume-se que o processo N é estacionário, mas não necessariamente isotrópico. O exemplo mais simples de processo pontual estacionário e isotrópico é o processo pontual de Poisson homogêneo.

Figs. 8a, 8b e 8c são padrões espaciais pontuais simulados CAE, regular e cluster, respectivamente. A diferença entre eles é a distância entre os pontos. Para um padrão regular a distância aparenta ser constante, enquanto para padrões cluster, existem grupos de pontos sobre a região. Em um padrão CAE, a distância entre os pontos é aleatória.

Considere N um processo pontual espacial bidimensional, em que cada evento é dado

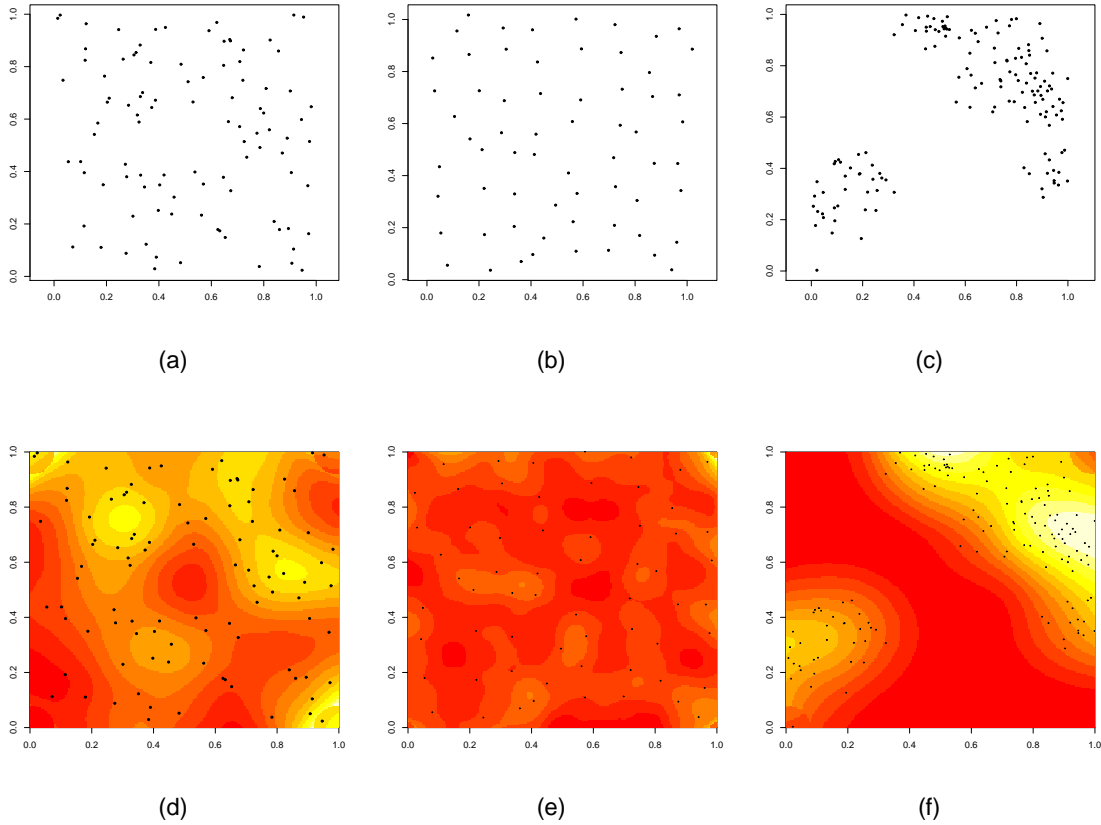


Figura 8 Simulações de processos pontuais espaciais de Poisson aleatório (a), regular (b) e cluster (c). Função de intensidade estimada para os padrões simulados aleatório (d), regular (e) e cluster (f).

por $\mathbf{x} \in \mathbb{R}^2$. Seja A a região onde os pontos são observados, denote por $|A|$ a área desta região, e por $N(A)$ o número de pontos aleatórios em A . Seja $d\mathbf{x}$ uma região infinitesimal que contém o ponto \mathbf{x} . A função de intensidade pode ser descrita pela propriedade de primeira ordem, dada por

$$\lambda(\mathbf{x}) \equiv \lim_{|d\mathbf{x}| \rightarrow 0} \left[\frac{E[N(d\mathbf{x})]}{|d\mathbf{x}|} \right], \quad (4.5)$$

em que $E[\cdot]$ é o valor esperado; $N(\cdot)$ é o número de eventos na região do plano; $d\mathbf{x}$ é uma pequena região que contém o evento \mathbf{x} ; $|d\mathbf{x}|$ é a área da região $d\mathbf{x}$. Para um processo estacionário, $\lambda(\mathbf{x}) \equiv \lambda$, o valor médio de eventos por unidade de área (Daley e Vere-Jones, 2002).

A estimação da propriedade de primeira ordem, ou intensidade do padrão pontual, é geralmente realizada por funções kernel (Baddeley et al., 2006; Baddeley et al., 2015),

$$f(\mathbf{x}) = \frac{1}{N(A)} \sum_{i=1}^{N(A)} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \quad (4.6)$$

em que $k(\cdot)$ é a função kernel com $\int k(u)du = 1$, e $h > 0$ o parâmetro de suavização. A estimativa da intensidade no ponto \mathbf{x} depende somente da relação espacial entre \mathbf{x} e os elementos da amostra \mathbf{x}_i ($i = 1, \dots, N(A)$), quantificados pela métrica incorporada na

função kernel. Para estimar $\lambda(\mathbf{x})$ utilizando-se a função kernel como descrito na Equação (4.6), utilizou-se a função `kernel2d` do pacote **splancs** (Rowlingson e Diggle, 2016) que está representada pelas Figuras 8d-8f. A estimativa kernel utilizando-se parâmetro de suavização quártica, $h = 0.2$, foi estimada para cada padrão simulado e está representada nas Figs. 8d, 8e e 8f o que corresponde às estimativas para a propriedade de primeira ordem para cada padrão pontual simulado. Vale a pena notar que a intensidade é quase constante para o padrão regular simulado, conforme Fig. 8e e, para o padrão cluster simulado, existem regiões com diferentes valores de intensidade, veja Fig. 8f.

As propriedades de segunda ordem de um processo pontual podem ser observadas utilizando-se da covariância entre eventos de duas áreas. A intensidade de segunda ordem possibilita uma forma de descrever o relacionamento entre pares de pontos, a qual é definida como

$$\lambda_2(\mathbf{x}, \mathbf{y}) \equiv \lim_{|d\mathbf{x}|, |d\mathbf{y}| \rightarrow 0} \left[\frac{E[N(d\mathbf{x})N(d\mathbf{y})]}{|d\mathbf{x}||d\mathbf{y}|} \right], \quad (4.7)$$

em que $d\mathbf{x}$ é uma pequena região contendo o evento \mathbf{x} ; $d\mathbf{y}$ é uma pequena região contendo o evento \mathbf{y} e $|d\mathbf{x}|$ e $|d\mathbf{y}|$ áreas das regiões $d\mathbf{x}$ e $d\mathbf{y}$, respectivamente. Para um processo estacionário, $\lambda_2(\mathbf{x}, \mathbf{y}) \equiv \lambda_2(\mathbf{x} - \mathbf{y})$, e para um processo estacionário e isotrópico, $\lambda_2(\mathbf{x} - \mathbf{y}) = \lambda_2(r)$ em que $r = \|\mathbf{x} - \mathbf{y}\|$ é a distância Euclideana (Daley e Vere-Jones, 2002).

Ripley (1976, 1977) apresenta uma caracterização alternativa da propriedade de segunda-ordem (Equação 4.7) para um processo estacionário e isotrópico, dado por

$$\lambda K(r) = E[N(b(o, r))].$$

O termo $\lambda K(r)$ é o número médio de pontos além do ponto típico em uma bola de raio r com centro no ponto típico. A função $K : [0, \infty) \rightarrow [0, \infty)$ é estimada como

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_{i=1}^{N(A)} \sum_{j \neq i} \frac{I(d_{ij} < r)}{N(A)},$$

em que d_{ij} é a distância entre o i -ésimo e o j -ésimo pontos, e $I(x)$ é a função indicadora de valor 1 se x é verdadeiro e 0 caso contrário, e $\hat{\lambda}$ é a estimativa para Equação (4.5).

Esta função tem mostrado sua importância como resumo de dependência espacial, dado que permite quantificar a dependência espacial entre diferentes regiões do processo, e é também conhecida como análise de segunda ordem reduzida. Ela tem sido amplamente utilizada em processos pontuais bidimensionais porque permite a detecção de padrões pontuais em diferentes escalas de distância simultaneamente, e o padrão pontual observado pode ser comparado a modelos de processos estocásticos conhecidos para diferentes configurações pontuais (Diggle, 2003).

A forma de $K(r)$ relativa à CAE fornece informações importantes acerca da distribuição do processo pontual (Illian et al., 2008), para um processo CAE, tem-se $K(r) = \pi r^2$. Para um processo cluster, os pontos possuem mais vizinhos do que esperado sob CAE, portanto, as estimativas para $K(r)$ serão maiores do que πr^2 . Por outro lado, para um padrão regular, as

estimativas de $K(r)$ serão menores que πr^2 (Diggle, 2003). Fig. 9 representa a função K e o intervalo de 95% de confiança para cada padrão pontual simulado nas Figs. 8a, 8b e 8c, respectivamente.

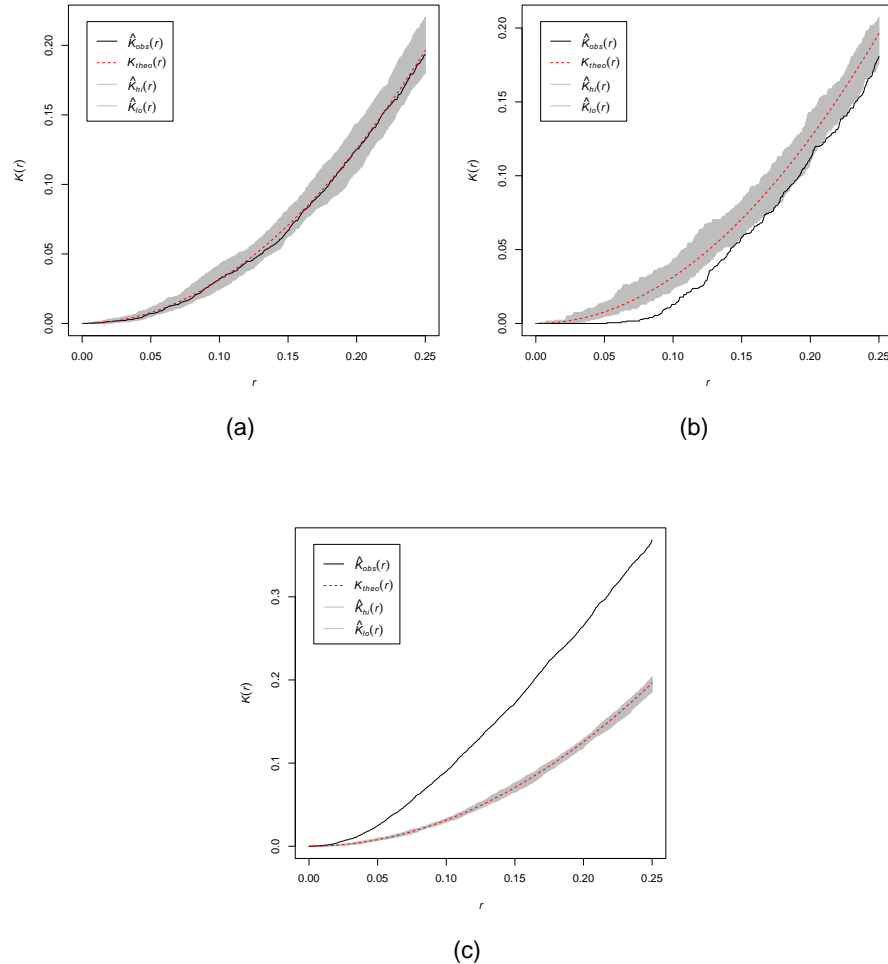


Figura 9 Função K de Ripley para padrão pontual aleatório (a), regular (b) e cluster (c). A linha pontilhada vermelha representa o caso teórico CAE, $K(r) = \pi r^2$ e a região cinza um envelope empírico de 95%. A linha contínua preta representa os dados observados.

Uma proposta direcional para a função K de Ripley foi desenvolvida por Rajala et al. (2016) como uma metodologia não paramétrica para estudar anisotropia em padrões espaciais pontuais. Ao invés de contar todos os eventos dentro de um círculo de raio r cujo centro é um ponto, pode-se utilizar outra forma geométrica (Baddeley et al., 2015). Para um vetor unitário \mathbf{u} , a função cônica K direcional é definida como

$$K_{cn}(r, \theta) = \frac{1}{\lambda^2 |A|} E \sum_{x_1, x_2 \in N}^{\neq} 1[x_1 \in A, x_2 - x_1 \in C_u(r, \theta)], \quad 0 < r, \quad 0 \leq \theta \leq \frac{\pi}{2} \quad (4.8)$$

em que λ é a intensidade e $C_u(r, \theta)$ denota um cone duplo na direção \mathbf{u} com uma inclinação de altura de comprimento r e um de ângulo o tamanho 2θ centrado em 0 (Redenbach et al., 2009).

A função é obtida pela contagem do número de pares de eventos do padrão que

possuem o ângulo do vetor diferença menor que $\epsilon = \beta - \alpha$ radianos a partir da direção 'u' e o comprimento do vetor diferença inferior ao alcance r . O procedimento consiste em comparar cada uma das direções, definidas pelo pesquisador, com a função sob CAE. Para tal, é necessário assumir que a anisotropia é resultado de achatamento ou alongamento e modelada por uma transformação linear composta de escala e rotação (Rajala et al., 2016).

Mateu e Nicolis (2015) desenvolveram um teste baseado em ondaletas utilizando-se o logaritmo empírico do escalograma direcional, assumindo que sob a hipótese nula de isotropia, um padrão pontual aleatório é esperado ter o mesmo valor do escalograma direcional para todas as possíveis direções. Seja $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^2$ uma função, a transformação contínua direcional das ondaletas para a escala a e orientação θ é dada por

$$W_f(a, \mathbf{b}, \theta) = \langle f, \Psi_{a, \mathbf{b}, \theta} \rangle = \int_{\mathbb{R}^2} f(\mathbf{x}) \overline{\Psi_{a, \mathbf{b}, \theta}(\mathbf{x})} d\mathbf{x},$$

em que a linha denota o conjugado complexo. Existem diversas ondaletas direcionais diferentes, $\Psi_{a, \mathbf{b}}(\mathbf{x}, \theta)$ propostas na literatura, veja por exemplo Neupauer e Powell (2005).

Kumar (1995) propôs a função

$$\eta(a, \theta) = \int |W_f(a, \mathbf{b}, \theta)|^2 d\mathbf{b} \quad (4.9)$$

com vistas a identificar o comportamento do processo em diferentes direções. $\eta(a, \theta)$ caracteriza a distribuição da energia em diferentes escalas e direções. O escalograma direcional é obtido a partir dos coeficientes das ondaletas para cada escala e cada direção, dados por $E[|W_{\Psi} f(\mathbf{b}, a, \theta)|^2]$.

Mateu e Nicolis (2015) aplicaram as transformações anisotrópicas das ondaletas de Morlet dadas em Neupauer e Powell (2005) para a função intensidade $\hat{\lambda}(\mathbf{x})$, estimada pelo método de contagem de quadrados, em que \mathbf{x} denota localizações. Sejam m possíveis direções $\theta_i \in (0, \pi]$ com $i = 1, \dots, m$; L possíveis escalas a_j com $j = 1, \dots, L$, e $N(A)$ pontos espaciais $\mathbf{b}_k, k = 1, \dots, N(A)$. As transformações das ondaletas direcionais $W_{\hat{\lambda}}(a_j, \mathbf{b}_k, \theta_i)$ são construídas para um alcance de escalas e orientações em todas as posições no domínio de $\hat{\lambda}(\mathbf{x})$. Denote por S_{θ_i, a_j} a variância do correspondente coeficiente das ondaletas para uma direção particular θ_i e escala a_j ,

$$\bar{S}_{\theta_i, a_j} = \frac{1}{N(A)} \sum |W_{\hat{\lambda}}(a_j, \mathbf{b}_k, \theta_i)|^2.$$

Simulações de Monte Carlo determinam a significância dos coeficientes das ondaletas direcionais. Assumindo que a hipótese nula é isotropia, um padrão pontual aleatório deve ser esperado ter o mesmo valor do escalograma direcional para todas as possíveis direções. Nicolis et al. (2010) consideraram o logaritmo da variância dos coeficientes das ondaletas sobre todas as escalas como teste estatístico,

$$T(\theta_i) = \frac{1}{L} \sum_{j=1}^L \log \bar{S}_{\theta_i, a_j}. \quad (4.10)$$

Um método alternativo para estudar o comportamento de isotropia do conjunto de dados é o histograma do ângulo do k -ésimo vizinho mais próximo. Para cada evento no padrão pontual, calcularam-se a distância do k -ésimo vizinho mais próximo e o ângulo azimute entre este pares de eventos, $[-\pi, \pi]$. O comportamento isotrópico dos dados é feito pela comparação com o valor teórico $1/2\pi$, em que picos acima deste valor teórico indica uma possível direção de ângulo dominante de anisotropia. Os histogramas podem ser utilizados para investigar em quais vizinhos mais próximos o processo alcança isotropia, isto é, quando não existem mais picos no histograma. Utilizou-se a função `knnangle` do pacote **Kdirectional** (Rajala, 2017) para a construção dos histogramas.

Para confirmar a forma do relacionamento do padrão pontual, realizou-se o teste de contagem de quadrados (Illian et al., 2008). Este é um teste para CAE que usa a estatística χ^2 baseada em contagem de quadrados em que a área de estudo A , é dividida em subregiões denominadas de quadrados (A_1, A_2, \dots, A_q) de igual área. O teste conta o número de pontos que estão em cada quadrado $n_j = N(A \cap A_j)$ para $j = 1, \dots, q$. Sob a hipótese nula de CAE, os n_j são variáveis aleatórias de Poisson independentes e identicamente distribuídas. O teste estatístico é dado por y

$$\chi^2 = (q - 1) \frac{s^2}{\bar{n}}, \quad (4.11)$$

em que s^2 é a variância amostral. Se a distribuição for Poisson, a variância é exatamente a média sob CAE, $s^2/\bar{n} = 1$. Além disso, se $s^2/\bar{n} < 1$ existe muito pouca variação entre os quadrados, os quais sugerem dispersão ao invés de aleatoriedade. Similarmente, se $s^2/\bar{n} > 1$ há muita variação entre os quadrados, sugerindo agrupamento.

Para a análise de dados utilizou-se o software R versão 3.3.3 (R Core Team, 2017) e os pacotes **spatstat** versão 1.47-0 (Baddeley e Turner, 2005) e **Kdirectional** versão 1.01 (Rajala, 2017).

4.2.3 Resultados

A Fig. 7c representa a localização das plantas de milho infestadas pela *Spodoptera frugiperda* na área de estudo. O espalhamento dos pontos ao longo da região não aparenta nenhuma direção predominante para o ataque da lagarta do cartucho.

Quando se aplica o método de kernel dado pela Equação 4.6 para estimar a função intensidade, observam-se alguns clusters correspondentes aos ângulos de 135 e 45 graus, aproximadamente (Fig. 10). O método permite ver como a concentração muda ao longo da área de estudo, entretanto, não é muito útil para detectar padrões anisotrópicos dado que o método está baseado em funções kernel isotrópicas.

Um primeiro estudo de anisotropia é realizado pelo histograma de ângulo vizinho mais próximo, o qual representa um comportamento descritivo do padrão pontual. Fig. 11 representa os histogramas para o primeiro e terceiro ângulo vizinho mais próximo, respectivamente. O histograma para o primeiro ângulo vizinho mais próximo (Fig. 11a) mostra duas direções principais dadas pelos picos mais altos. A primeira direção corresponde

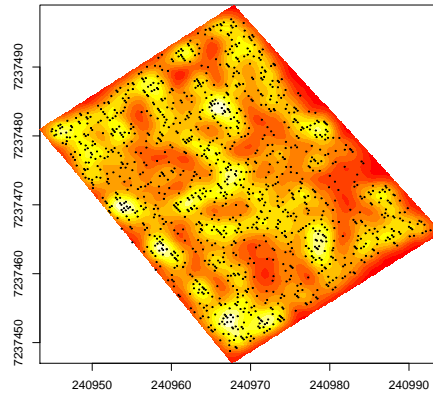


Figura 10 Função de intensidade estimada usando kernel quártico, parâmetro de suavização $h = 2$ e uma grade regular de 400×400 pixels.

aproximadamente aos ângulos de 135 e 315 graus, e o segundo pico mais alto corresponde aos ângulos 45 e 217 graus. Vale a pena notar que estas direções correspondem às direções utilizadas pela semeadora. Quando considera-se o histograma dos terceiros ângulos vizinhos mais próximos (Fig. 11b), existem somente duas direções dominantes principais, 135 e 315 graus, entretanto há uma redução do valor da densidade. Nota-se ainda que o processo tende a permanecer abaixo da linha teórica para todos os outros graus.

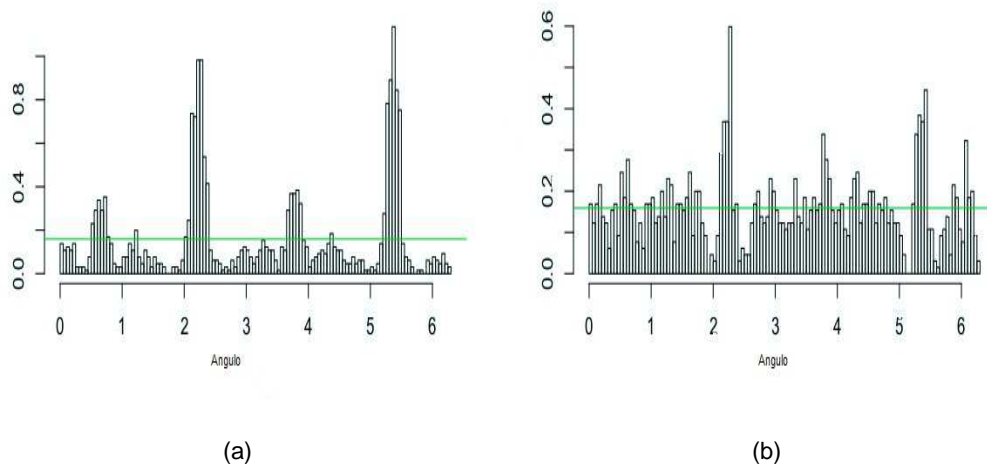


Figura 11 Histograma para os ângulos mais próximos para o primeiro vizinho (a) e terceiro vizinho (b). Neste gráfico, 0 corresponde a 0^0 ; 0,79 a 45^0 ; 1,57 a 90^0 ; 2,36 a 135^0 ; 5,50 a 315^0 e 6,28 a 360^0 .

Rotacionou-se o conjunto de dados em 45 graus no sentido horário, com o propósito de corrigir a anisotropia observada na análise dos histogramas. O padrão pontual rotacionado está representado pela Fig. 12b. E, a fim de eliminar qualquer possível efeito de borda apresentado pela forma da área experimental, selecionou-se um número de pontos no centro

da área, o qual é representado pela Fig. 12c.

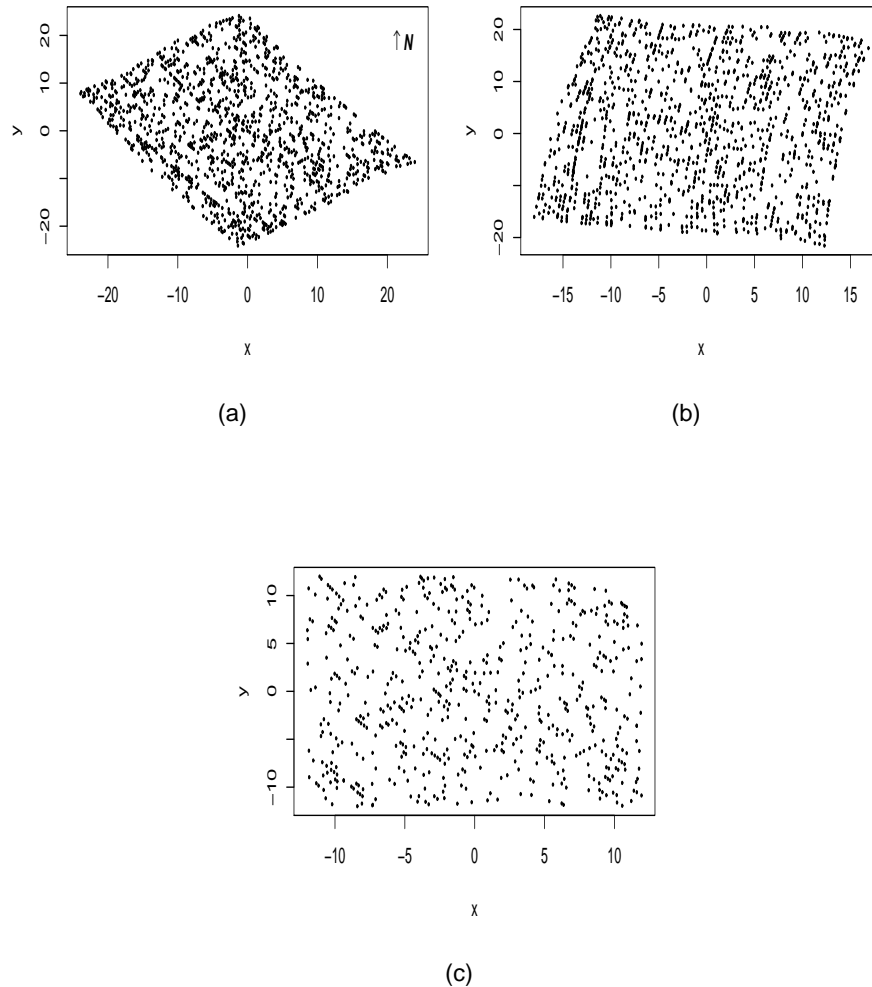


Figura 12 Padrão pontual original (a), padrão pontual rotacionado 45° no sentido horário (b) e recorte da região central do padrão rotacionado (c).

No que segue, calculou-se a função K direcional dada pela Equação 4.8 e o teste ondaletas com Equação 4.10 para os três conjuntos de dados comparando as características de isotropia/anisotropia. A propriedade de segunda ordem, conforme definida na Equação 4.7, é estimada pela função K direcional. Usou-se a forma de cone com direções $2\theta = 0, 45, 90$ e 135 graus com respeito ao eixo x , dado que estas direções geralmente são utilizadas na literatura para estudar anisotropia. A Fig. 13 representa a função K direcional para cada padrão e cada uma das direções consideradas com um envelope de 95% simulado sob a hipótese nula de CAE. É evidente para o conjunto de dados originais uma forte clusterização em cada direção, conforme visualizado pela função K direcional representada pela Fig. 13a, observa-se que a direção mais dominante de anisotropia é 135° , seguida por 45° . As duas outras direções (0° e 90°) apresentam quase o mesmo comportamento. Considerando o padrão rotacionado mostrado pela Fig. 13b, observa-se uma redução da anisotropia e uma fraca clusterização. A direção mais dominante de anisotropia é 90 graus, seguida pela

direção de 0° . E, para o padrão pontual recortado, representado pela Fig. 13c, vê-se leve agrupamento e anisotropia muito fraca. Entretanto, todas as direções consideradas das funções K direcionais do padrão original estão fora do envelope, então, pode-se assumir que o processo não é completamente aleatório.

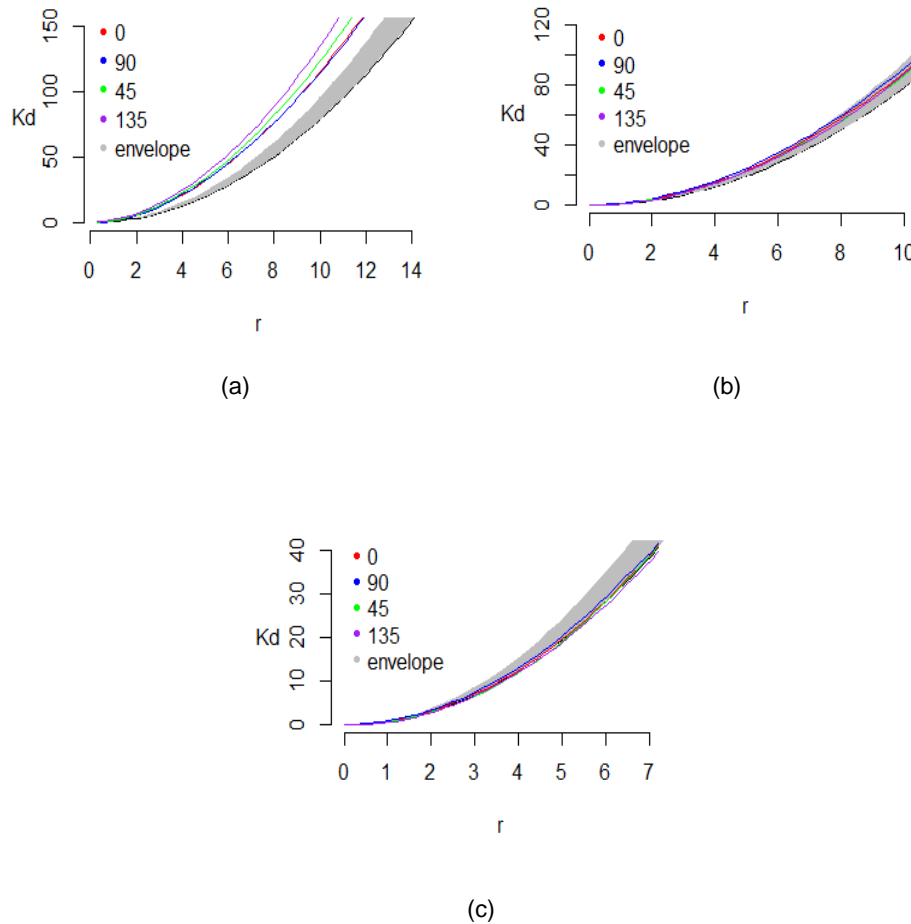


Figura 13 Função K direcional para o padrão original (a), rotacionado (b) e recortado (c) com envelopes empíricos simulados de 95% sob a hipótese nula de CAE para cada um. A distância r no eixo x é dada em metros.

A Fig. 14 representa os escalogramas direcionais para os conjuntos de dados quando aplica-se a transformação das ondaletas direcional Morlet para cada grau (de 1 a 180) e escala (de 1 a 30). Para o padrão pontual original (Fig. 14a), vê-se uma anisotropia evidente na direção de aproximadamente 135 graus no nível mais alto de resolução e espalhamento em um ângulo maior (entre 80 e 140 graus) nos níveis menores (próximo de $a = 30$). Para o padrão rotacionado, Fig. 14b, o escalograma direcional apresenta fraca anisotropia com uma direção dominante em aproximadamente 80 graus. Finalmente, o escalograma avaliado para o padrão pontual recortado, representado pela Fig. 14c, não mostra qualquer direção dominante.

Em seguida, realizou-se o teste utilizando-se o logaritmo empírico do escalograma

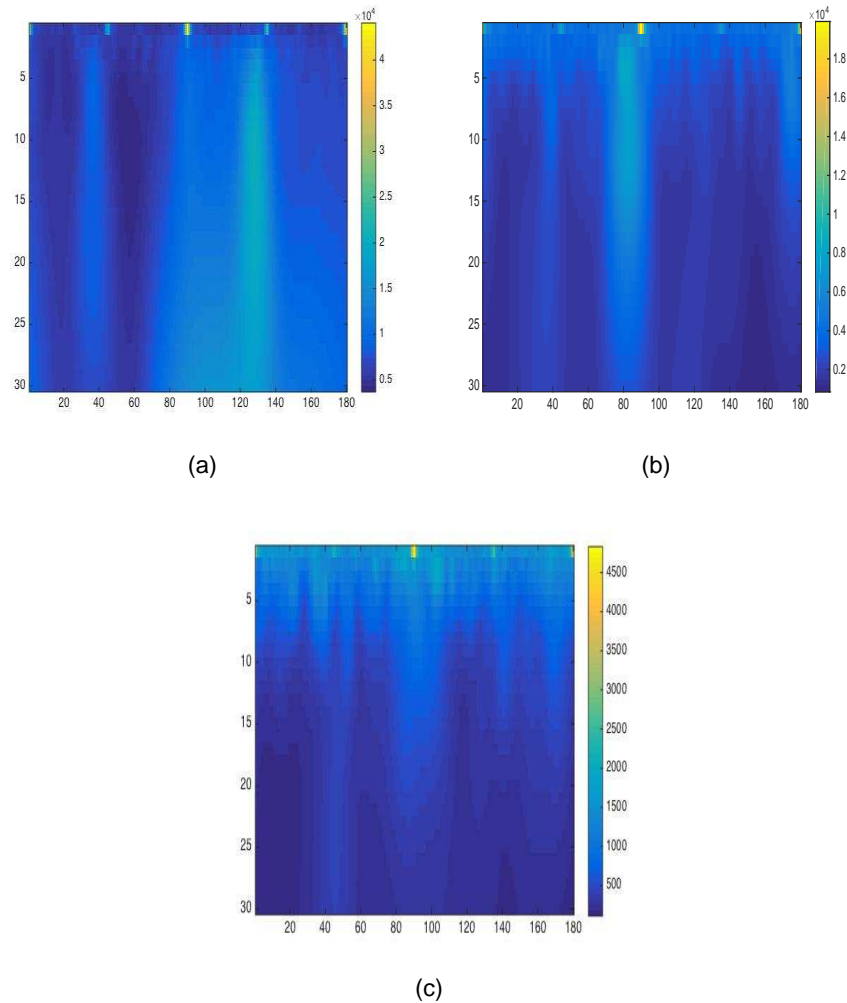


Figura 14 Escalograma direcional para o padrão original (a), padrão rotacionado (b) e padrão recortado (c) usando a ondaletas de Morlet anisotrópica . As transformações das ondaletas foram avaliadas em uma grade de dimensão 64×64 pixels para cada direção $\theta \in [0, 180]$ (eixo x), e $a = 1 \dots 30$ escalas (eixo y).

direcional conforme descrito na Equação 4.9. A linha azul da Fig. 15 representa os valores de $T(\theta_i), i = 1, \dots, 180$ para o conjunto de dados de *Spodoptera frugiperda*. Com o objetivo de testar se o processo é CAE, foram simulados 1000 padrões pontuais isotrópicos de Poisson e avaliaram-se os 2.5% e 97.5% do total da distribuição empírica (linhas vermelhas da Fig. 15). Para os dados originais (Fig. 15a), a estatística T está fora do intervalo empírico de 95% de confiança para todos os ângulos, e mostra um pico entorno do ângulo 135 graus. Neste caso, o teste claramente rejeita a hipótese de que o processo é isotrópico. Quando se realiza a rotação do conjunto de dados por 45 graus, o resultado da estatística T (representado na Fig. 15b) não apresenta direções dominantes significativas, exceto para os graus de alcance 80 – 85, que estão ligeiramente fora do intervalo de confiança empírico. Ao se observar a estatística T avaliada no padrão recortado (Fig. 15c), não há direção dominante e o padrão pontual considerado aparenta ser completamente aleatório, confirmando os resultados obtidos pela análise K –direcional. Talvez a anisotropia do conjunto original de dados deva ter sido devido à forma da área experimental, pois parece que os insetos tendem a se propagar nas plantas

mais próximas. Com a correção da anisotropia, a análise das ondaletas mostra uma direção dominante perto de 80 graus que não é evidente no teste K -direcional. Dado que o conjunto de dados recortados aparenta CAE. Assim, é possível que a anisotropia da Fig. 15b esteja localizada próxima à borda da área experimental. A principal vantagem do teste baseado em ondaletas com respeito aos demais métodos estatísticos é que este seja capaz de detectar possíveis anomalias dos dados em diferentes níveis de resolução.

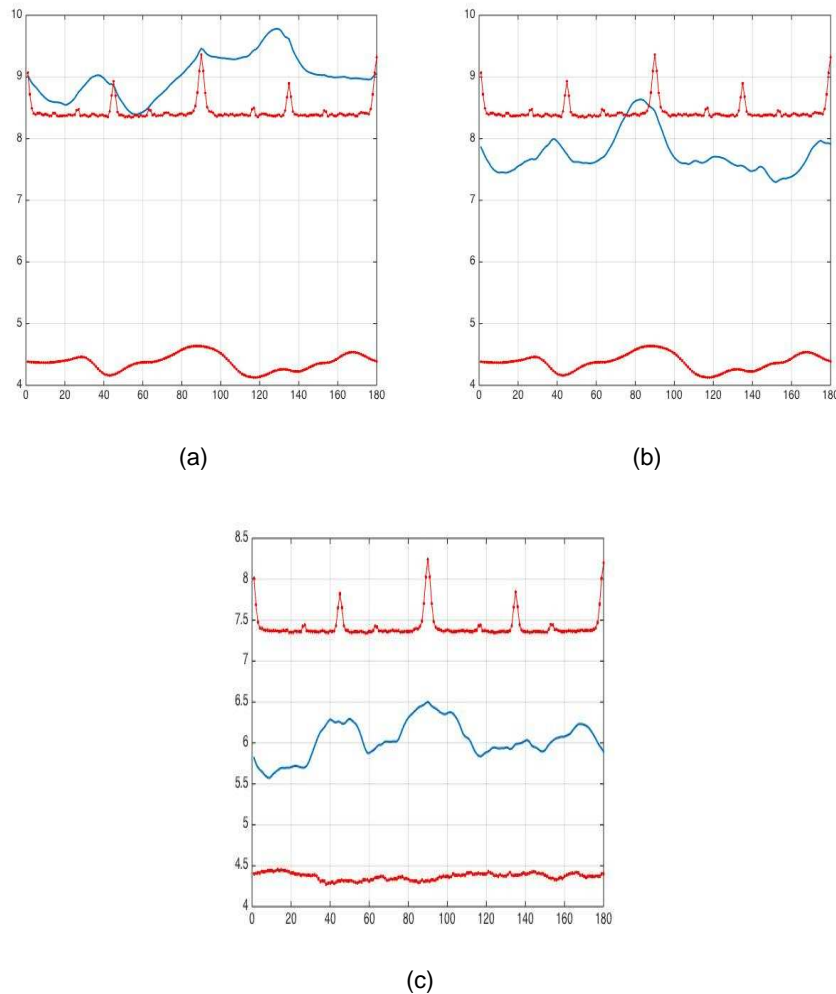


Figura 15 Teste baseado em ondaletas direcional. Valores da estatística $T(\theta_i)$, $i = 1, \dots, 180$ para os dados de *Spodoptera frugiperda* (linha contínua) e intervalo empírico de 95% de confiança (linhas vermelhas) obtido pela simulação de 1000 padrões pontuais sob a hipótese de CAE, com $\lambda = 1303$ para o conjunto de dados originais (a) e rotacionados (b), e $\lambda = 557$ para o conjunto de dados recortado (c).

A fim de analisar as características de clusterização dos dados, aplicou-se o teste de contagem de quadrados, conforme descrito pela Equação 4.11. A Tabela 13 possui os resultados obtidos, e de acordo com eles, deve-se rejeitar a hipótese de CAE e concluir que o padrão pontual é agrupado. Os resultados mostram que, para a fase de crescimento vegetativa da cultura do milho, na qual os dados foram coletados, a lagarta do cartucho apresentou comportamento agrupado.

Tabela 13 Resultados do teste de contagem de quadrados, considerando-se diferentes quantidades de quadrados

quadrats	χ_{orig}^2	d.f.	$\chi_{1-\alpha/2}^2$	χ_{rot}^2	d.f.	$\chi_{1-\alpha/2}^2$	χ_{cent}^2	d.f.	$\chi_{1-\alpha/2}^2$
8	135,36	45	65,41	82,38	63	86,83	75,79	63	86,83
12	210,15	93	121,57	198,57	142	176,88	151,25	143	178,00
16	335,78	158	194,70	335,15	252	297,86	267,98	255	301,12

d.f.: graus de liberdade; χ_{orig}^2 , χ_{rot}^2 e χ_{cent}^2 são as estatísticas de teste para o padrão original, rotacionado e recortado, respectivamente; $\alpha = 0.05$.

4.2.4 Discussões

O conhecimento do espalhamento da população de *Spodoptera frugiperda*, o inseto que mais danos causa à cultura do milho, é extremamente importante para o desenvolvimento de técnicas eficientes tanto para o monitoramento quanto para o manejo da cultura. Nesta pesquisa, foram aplicadas técnicas de processos pontuais espaciais para estudar o espalhamento da lagarta do cartucho em uma área comercial agrícola.

A partir da estimação da função de intensidade dos dados da lagarta do cartucho (Fig. 10), observa-se uma maior concentração deste inseto na região próxima à borda esquerda da área experimental. Entretanto, outras pequenas regiões parecem ser caracterizadas por clusters. O agrupamento do padrão pontual da lagarta do cartucho foi também confirmado pelo teste de contagem de quadrados (Tabela 13).

Nossos resultados também mostraram anisotropia quando considerado o conjunto de dados originais. Ambos os métodos, o K -direcional e o baseado em ondaletas, detectaram anisotropia nas direções de 45 e 135 graus, provavelmente devido ao formato da área experimental e da semeadora. Com vistas a corrigir esta anisotropia, rotacionou-se o padrão pontual em 45 graus no sentido horário. O padrão pontual rotacionado resultante não apresentou direção dominante quando utilizado o método K -direcional. Em contra partida o método baseado em ondaletas detectou uma leve anisotropia em torno do ângulo 80 graus. O resultado diferente obtido pelo método de ondaletas com respeito ao K -direcional é provavelmente devido à habilidade da transformação das ondaletas em analisar o padrão pontual em cada ângulo e escala de resolução.

Com o objetivo de evitar qualquer efeito de borda, selecionou-se um subconjunto central dos dados rotacionados, e estimou-se a função K -direcional e o escalograma direcional para este novo padrão pontual. Neste caso, ambos os métodos não detectaram qualquer direção dominante e mostraram completa aleatoriedade do processo. A diferença entre os estudos de anisotropia apresentados aqui consiste no número de escalas de resolução, dado que o teste de ondaletas é capaz de identificar direções dominantes em escalas mais finas do que a função K -direcional. Entretanto, quando compararam-se ambos os resultados, temos a mesma conclusão. Em nosso conhecimento, não existem resultados na literatura acerca do estudo de anisotropia para padrões espaciais pontuais em agricultura, por esta razão não é possível realizar comparações com outros resultados. Entretanto, existem

alguns trabalhos que analisam a propriedade de clusterização da *Spodoptera frugiperda* (veja, Farias et al., 2008 e Rios et al., 2014). Pensamos que nossos resultados de agrupamento são em geral coerentes com os obtidos por Farias et al. (2008) e Rios et al. (2014), embora em nosso estudo não se tenha levado em consideração o tamanho da lagarta do cartucho. Em particular, Farias et al. (2008) encontraram que o espalhamento da *Spodoptera frugiperda* para pequenas larvas apresentam comportamento agregado, tornando-se cada vez mais aleatório com o crescimento da larva. Hernández-Mendoza et al. (2008) também concluíram que a distribuição espacial da lagarta do cartucho esteve fortemente associada ao estágio fenológico da cultura do milho. Dado que a área experimental considerada neste trabalho consiste em uma pequena parte de toda a área cultivada com a cultura do milho, acredita-se que resultados mais consistentes em anisotropia podem ser obtidos se estudada uma área maior. Também considera-se que melhores resultados sobre a distribuição espacial da *Spodoptera frugiperda* poderiam ser obtidos se considerado o tamanho do inseto. Dado que este banco de dados não possui esta informação, considerar-se-ão estas questões em trabalhos futuros, juntamente com imagens de satélite de alta resolução ou imagens de drones com a finalidade de estudar áreas de maior dimensão. Finalmente, planeja-se melhorar a estimação da função intensidade utilizando-se técnicas baseadas em kernel direcional e funções baseadas em ondaletas direcionais.

4.2.5 Agradecimentos

O primeiro autor agradece pela bolsa recebida da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior para realizar doutorado sanduíche na Universidad de Valparaíso-Chile. Ainda, agradece pelo suporte financeiro da Fundação Araucária do estado do Paraná, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil.

4.2.6 Referências

- Baddeley A, Turner R, 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. **Journal of Statistical Software** 12(6):1-42
- Baddeley A, Rubak E, Turner R, 2015. **Spatial Point Patterns - Methodology and Applications with R**. CRC Press, Florida. 810p.
- Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D, 2006. **Case studies in spatial point process modelling**. Springer, New York. 310p.
- Cruz I, Figueiredo MLC, Oliveria AC, Vasconcelos CA, 1999. Damage of *Spodoptera frugiperda* (Smith) in different maize genotypes cultivated in soil under three levels of aluminium saturation. **International Journal of Pest Management** 45(4):293-296
- Daley DJ, Vere-Jones D, 2002. **An introduction to the theory of point processes**. 2ed. New York: Springer. 536p.

- Diggle P, 2003. **Statistical Analysis of Spatial Point Patterns**. 2ed. New York: Oxford University Press Inc. 155p.
- EMBRAPA, 2006. **Sistema brasileiro de classifica o de solos** EMBRAPA.
- Farias PRS, Barbosa JC, Busoli AC, Overal WL, Miranda VS, Ribeiro SM, 2008. Spatial analysis of the distribution of *Spodoptera frugiperda* (J.E. Smith)(Lepidoptera: Noctuidae) and losses in maize crop productivity using geostatistics. **Neotropical Entomology** 37(3):321-327
- Guan Y, Sherman M, Calvin JA, 2006. Assessing isotropy for spatial point processes. **Biometrics** 62:119-125
- Hernandez-Mendoza JL, Lopez-Barbosa EC, Garza-Gonzalez E, Mayek-Perez N, 2008. Spatial distribution of *Spodoptera frugiperda*(Lepidoptera: Noctuidae) in maize landraces grown in Colima, Mexico. **International Journal of Tropical Insect Science** 28(3):126-129
- Illian J, Penttinen A, Stoyan H, Stoyan D, 2008. **Statistical Analysis and Modelling of Spatial Point Patterns**. England:John Wiley & Sons Ltda. 560p.
- Kumar P, 1995. A wavelet based methodology for scale-space anisotropic analysis. **Geophysical Research Letters** 22(20):2777-2780
- Mateu J, Nicolis O, 2015. Multiresolution analysis of lineary oriented spatial point patterns. **Journal of Statistical Computation and Simulation** 85:621-637
- Neupauer RM, Powell KL, 2005. A fully-anisotropic Morlet wavelet to identify dominant orientations in a porous medium. **Computer and Geosciences** 31:465-471
- Nicolis O, Mateu J, D'Ercole R, 2010. **Testing for anisotropy in spatial point processes**. Proceedings of the Fifth International Workshop on Spatio-Temporal Modelling (METMA5). Gonzalez-Manteiga et al. (Eds.), Publisher Unidixital. ISBN:978-840934272-8-3. Deposito Legal: C 1990-2010.
- Niu Y, Meagher Jr RL, Yang F, Huang F, 2013. Susceptibility of field populations of the fall armyworm (Lepidoptera: Noctuidae) from Florida and Puerto Rico to purified Cry1f protein and corn leaf tissue containing single and pyramided Bt genes. **Florida Entomologist** 96(3):701-713
- R Core Team, 2017. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna.
- Rajala TA, 2017. **Analysis of anisotropy in point patterns using second order statistics**. <https://github.com/antiphon/Kdirectional/blob/master>
- Rajala TA, Sarkka A, Redenbach C, Sormani M, 2016. Estimating geometric anisotropy in spatial point patterns. **Spatial Statistics** 15(1):100-114
- Redenbach C, Sarkka A, Freitag J, Schladitz K, 2009. Anisotropy analysis of pressed point processes. **AStA Advances in Statistical Analysis** 93(3):237-261.

Ribeiro Jr PJ, Viola DN, Demétrio CGB, Manly BF, Fernandes OA, 2009. Spatial Pattern detection modelling of Thrips (*Thrips tabaci*) on onion fields. **Scientia Agricola** 66(1):90-99

Rios ES, Martins ICF, Noronha MP, Silva JA, Silva Filho JG, Badji CA, 2014. Spatial distribution of Spodoptera frugiperda in the wasteland of southern Pernambuco state, Brazil. **Amazonian Journal of Agricultural and Environmental Sciences** 57:297-304.

Ripley BD (1981) **Spatial Statistics**. New Jersey: John Wiley & Sons. 272p.

Ripley BD, 1976. The second-order analysis of stationary point processes. **Journal of Applied Probability** 13(2):255-266

Ripley BD, 1977. Modelling Spatial Patterns. **Journal of the Royal Statistical Society** 39(2):172-212.

Rowlingson B, Diggle P, 2016. **splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-39**. <https://CRAN.R-project.org/package=splancs>

Sena Jr DG, Pinto FAC, Queiroz DM, Viana PA, 2003. Fall armyworm damaged maize plant identification using digital images. **Biosystems Engineering** 85(4):449-454

Spósito MB, Amorin L, Ribeiro Jr PJ, Bassanezi RB, Krainski ET, 2007. Spatial Pattern of trees affected by Black spot in Citrus Groves in Brazil. **Plant Disease** 91(1):36-40

Yang F, Huang F, Qureshi JA, Leonard BR, Niu Y, Zhang L, Wangila DS, 2013. Susceptibility of Louisiana populations of Spodoptera frugiperda (Lepidoptera: Noctuidae) to transgenic Agrisure[®] Viptera[™] 3111 corn. **Crop Protection** 50(1):37-39

Yu SJ, Nguyen SN, Albo-Elghar GE, 2003. Biochemical characteristics of insecticide resistance in the fall armyworm, Spodoptera frugiperda (J.E. Smith). **Pesticide Biochemistry and Physiology** 77(1):1-11

5 CONSIDERAÇÕES FINAIS

Os resultados desta tese foram apresentados em forma de artigos. Os dados utilizados nos artigos são provenientes de um experimento implantado em uma área comercial agrícola na região Sul do Brasil, cultivada com a cultura do milho. Apresentam-se aqui as principais contribuições verificadas durante o desenvolvimento do trabalho:

- O artigo 1 definiu 40 parcelas identificadas por coordenadas geográficas, em área cultivada com a cultura do milho. O conjunto de dados é resultante da contagem do número de plantas infectadas pela *Spodoptera frugiperda* em cada parcela, e o interesse foi a distribuição do número de plantas infectadas em cada parcela. Neste artigo, desenvolveu-se uma nova metodologia para análise de influência local de primeira ordem para dados com distribuição marginal binomial e função de ligação logit. Mapas foram obtidos para representar a predição espacial com e sem os pontos influentes. Observou-se que a presença de pontos influentes nos conjunto de dados alteram a inferência estatística, os valores preditos e os respectivos mapas. Também realizou-se um estudo de simulação a fim de se avaliar a performance da metodologia de diagnóstico de influência local proposta.
- No artigo 2 estudou a distribuição espacial da lagarta do cartucho ao longo da área experimental. Foram observados tanto o número de plantas infectas pelo inseto em toda a área como suas coordenadas geográficas foram estudadas. Com o objetivo de detectar possível direção dominante de infestação, estudou-se as características de anisotropia a partir de técnicas de análise de padrões pontuais espaciais de segunda ordem, tal como a função K direcional, teste baseado em ondaletas e contagem de quadrados. A conclusão que se observa é que a distribuição espacial da lagarta segue um processo pontual de Poisson agrupado com evidente anisotropia principalmente devido à forma da área experimental.

Além das contribuições apresentadas acima, surgiram ao longo do desenvolvimento da tese alguns possíveis problemas a serem abordados em estudos futuros, como realizar o estudo da tendência da distribuição espacial da lagarta do cartucho em área de maior dimensão. Ainda, futuramente, pretende-se estudar a influência local para outras distribuições de probabilidades com diversas funções de ligação aplicadas em outros experimentos.

6 ANEXOS

6.1 Normas do Journal of Agricultural, Biological and Environmental Statistics

Information for authors

General

The Journal of Agricultural, Biological, and Environmental Statistics (JABES) publishes papers of immediate and practical value to applied researchers and statistical consultants in the agricultural sciences, the biological sciences (including biotechnology), and the environmental sciences (including those dealing with natural resources). Only papers that address applied statistical problems will be considered. Interdisciplinary papers as well as papers that illustrate the application of new and important statistical methods using real data are strongly encouraged. For regular papers, a motivating example should be presented early in the paper. The statistical development should then be presented, and the results applied to the example. Expository, review, and survey articles addressing broad-based statistical issues are encouraged. Presentation should be accurate, clear, and comprehensible to readers with a background in statistical applications. When necessary, detailed proofs, computer code, and other lengthy technical portions of a manuscript should be placed in an appendix so that they will not interfere with the primary focus on the paper, which is to be a discussion of the statistical methods and issues being addressed. Real data should almost always be used to illustrate the statistical applications being discussed. Articles that have been previously published in a refereed journal or articles that are under review by another journal may not be submitted to JABES.

Manuscript Submission

Manuscripts submitted in LATEX should generally use the “article” style and should not use any special macros. Submission of a manuscript implies: that the work described has not been published before; that it is not under consideration for publication anywhere else; that its publication has been approved by all co-authors, if any, as well as by the responsible authorities - tacitly or explicitly - at the institute where the work has been carried out. The publisher will not be held legally responsible should there be any claims for compensation. When submitting a manuscript, please DO NOT include any of your personal information (name, affiliation, etc.) or acknowledgments anywhere in the manuscript or on the cover page. This is to ensure a double-blinded review process. Thank you.

Online Submission

Authors should submit their manuscripts online. Electronic submission substantially reduces the editorial processing and reviewing times and shortens overall publication times. Please follow the hyperlink “Submit online” on the right and upload all of your manuscript files following the instructions given on the screen. For queries regarding submission, e-mail

jabes.editor@gmail.com. Each manuscript should be completely double-spaced, with one-inch margins and 12- point size letters.

Abstract

Each manuscript should contain an abstract of about 200 words. Authors must pay great attention to the content and language of the abstract. The first part of the abstract should succinctly describe the manuscript's motivation and statistical contribution. For the benefit of JABES's broad readership, the remainder of the abstract should amplify and illustrate, preferably using concrete examples and interesting special cases.

Keywords

Please provide three to six key words or phrases (that do not appear in the title) in alphabetical order. For published manuscripts, these key words will be used, in addition to the title, for indexing in BIOSIS, Current Index to Statistics, Current Contents/Agriculture, Biology and Environmental Sciences (CC/AB&ES), The Science Citation Index, Science Citation Index Expanded, Agricultural On Line Access (AGRICOLA), and Compumath Citation Index.

Length of Manuscripts

Normally, papers exceeding 25 pages in the style described above will be returned to the authors by the editor without review. Please note that the page limits include the appendixes, but not tables and figures. Papers appearing in the journal should rarely have more than six (6) tables or figures combined. Authors are strongly encouraged to be judicious in the use of tabular and graphical displays and should not combine what ought to be several tables or figures into very large single ones. Papers with an extreme number of tables and/or figures may be returned by the editor without review. It is recognized, however, that graphical depictions or images are essential for conveying the message in some substantive areas. In such circumstances, more figures than in the typical submission may be appropriate, and authors should note this explicitly in a covering letter. References to Web sites where additional figures, tables, or mathematical material (such as proofs) may be found are permitted.

General Comments on Style

Do not use footnotes, and avoid abbreviations. References within the text should give the authors' names, followed by the year of publication in parentheses. Mathematical material in the manuscript should be set using standard mathematical typography. This is normally done using LATEX (and its variants) or mathematical typesetting options such as the Equation Editor or MathType in Word. Please ensure that matrices and vectors are always set in bold. Complicated exponentials should be represented by the symbol $\exp()$. Fractions in the body of the text (and when possible in displayed expressions) and fractions occurring in the numerators or denominators of fractions should be written with the solidus-for example, $(a + b)/(c + d)$.

Figures and Tables

Figures and tables should be numbered consecutively at the top with Arabic numerals and clearly titled and labeled. Figures must be visually clear and capable of withstanding reduction. Once a manuscript is accepted for publication, authors should submit electronic files of all figures. The following formats are accepted: .PS, .EPS, .TIF, .BMP, .JPG, or

.PCX. Other formats are also accepted; please contact the ASA production office with any questions. Normally figures will be printed in gray scale or black and white. Color figures are extremely expensive to produce and JABES encourages authors to avoid color unless absolutely necessary. Extra page charges may be levied for color reproduction. Detailed guidelines for preparation of figures and tables are provided in the ASA Style Guide.

References

References should be listed in alphabetical order at the end of the manuscript after any appendixes, but before any tables and figures. Each reference should contain, in the following order, the last name and initials of the author followed by those of coauthors, year of publication, title of reference, source, volume number, and page. Do not abbreviate journal titles. Reference to books must include the publisher's name and location. Only list references actually cited in the text. The use of inessential, unpublished or obscure references is discouraged. Personal communications are listed in the text only. Other details for listing and citing references are given in the ASA Style Guide.

More details

<http://www.springer.com/statistics/journal/13253>

6.2 Normas do Spanish Journal of Agricultural Research

Author guidelines

Aims and scope

The Spanish Journal of Agricultural Research (SJAR) is a quarterly international journal that accepts research articles, reviews and short communications of content related to agriculture. Research articles and Short communications must report original work not previously published in any language and not under consideration for publication elsewhere.

SJAR publishes papers that report research findings on the following topics: agricultural economics; agricultural engineering; agricultural environment and ecology; animal breeding, genetics and reproduction; animal health and welfare; animal production; plant breeding, genetics and genetic resources; plant physiology; plant production (field and horticultural crops); plant protection; soil science; and water management. SJAR is not publishing articles whose topic is “food science and technology”. Original research on genetic association studies must report variance estimates explained by those genetic variants/alleles in study. Articles on local research will only be publishable if they show methodological innovation or results that can be extrapolated to other areas.

Manuscript preparation

Manuscripts should be written in Times New Roman 12-point font, with 1.5 line spacing. The four margins will be 2.5 cm. Section headings should be written 14-point font in bold print. All pages should be numbered consecutively, and line numbers should be printed on each page (starting with 1 on each page) to facilitate ease of reference for the reviewers. Each paragraph should begin with an indentation of 1-cm. Tables, figures and annexes must be included on separate sheets (but in the same Word document), one per page, following the References section. Use Word compatible software. Separate figure files will be required later if the manuscript is accepted.

Language: Manuscripts should be written in concise, legible English, which must be carefully reviewed by the authors for correctness of language and content. English spelling can be British or American, but it must be consistent throughout. Authors whose first language is not English are strongly advised to have their manuscripts checked by a proficient third party prior to submission.

Papers reporting field research. Field research should indicate replications in sites and years in connexion to the addressed aims of the experiment(s). In many cases, several years (a minimum of three in general) may be required to account for all variations in factors that affect plant growth and development, in particular for those researches dealing with production and yield. Since the lack of this requirement can be cause of manuscript rejection during the preliminary inspection, if this recommendation is not fulfilled, arguments supporting the validity of the results must be stated and included in the letter addressed to the Editorial Office.

Research articles should not be longer than 28 pages (or 8,500 words), including front page, references, tables and figures, and excluding the supplementary material. Research

articles should contain a front page, an abstract, up to seven keywords and the abbreviations used. The main text should have the following sections (see suggested layout below): introduction; material and methods; results; discussion (avoid combining the results and discussion sections); acknowledgments; references; tables and figures; and supplementary material, if any.

Layout

The following layout is strongly recommended:

Front page. The first page must include the following:

Title of the work. The title must be clear, short and concise. Avoid terms such as “Study of...”, “Observations...”, or “Contribution to...”. The title should preferably not exceed 20 words.

Authors’ names: We recommend hyphenating in the case of more than one surname (e.g., Luisa M. Torres-Cerezo). When authors are associated with different institutions, each author should be marked with a superscript number indicating the corresponding author’s mailing address.

Affiliations. Name and full postal address of the institution(s).

Corresponding author’s name and e-mail address.

Author contributions. Brief but clear account of the contributions of the different authors is mandatory. Contributions can be: conception or design; acquisition, analysis, or interpretation of data; drafting of the manuscript; critical revision of the manuscript for important intellectual content; statistical analysis; obtaining funding; administrative, technical, or material support; supervising the work; coordinating the research project, etc.

Number of tables and figures. If any, number of supplementary tables and/or figures, with the following sentence: “Supplementary material accompanies the paper on SJAR’s website”. In the text, they will be cited as: “Table S1/Fig. S1 [supplementary]”.

Running title of the work, used in the heading of the pages of the printed article, should not exceed 90 characters (including spaces).

Topic, selected from the following twelve: agricultural economics; agricultural engineering; agricultural environment and ecology; animal breeding, genetics and reproduction; animal health and welfare; animal production; plant breeding, genetics and genetic resources; plant physiology; plant production (field and horticultural crops); plant protection; soil science; water management.

Funding. Indicate the sources of financing the study, or write “The author(s) received no specific funding for this work”

Competing interests, if any, or write “The authors have declared that no competing interests exist”

Abstract, keywords, and abbreviations.

Special attention should be paid to the title and abstract, as these will influence readers’ decisions to proceed with the text.

Abstract. The abstract length is 250 words maximum. The style must be concise and must not contain references. A typical abstract structure can be as follows: i) describe the relevance of

the study and establish the goal or the specific objectives; ii) a brief description of the materials and methods; crops or organisms involved must be identified as well as soil type, chemicals, and other details that may be important to interpret the results; iii) list and discuss relevant results (including numeric values of experimental results); and iv) one or two closing sentences addressing the most relevant findings and implications.

Additional key words. A maximum of seven key words should be included. These should not repeat words that appear in the title.

Abbreviations used. Include a list of all non-standard abbreviations used in the paper and their meaning.

Text of the article

The text of the article should contain the following sections:

Introduction. The introduction should contain sufficient background information about the work to allow it to be placed in the context of other research and to allow the reader to understand the relevance, proposed objectives and evaluation of the results. The introduction should conclude with one or two sentences that define the objectives and the essence of the article. Authors presenting articles to the “agricultural economics” section are invited to present introductions shorter than usual in journals of the Social Science area, not exceeding two pages and avoiding the use of subchapters.

Material and methods. Sufficient information should be provided to enable experiments to be repeated. For routine methods, a brief description and literature reference will be enough. New methods must be described in detail and, in the case of rarely used chemical products or equipment, the manufacturer’s name and address should be given.

Results. In general, this section should not include literature references; it should only describe the results of the experiments. Interpretations of the experimental data should be reserved for the Discussion section. The explanations provided in the figure and table captions should not be repeated in the text.

Discussion. The discussion should not be limited to describe experimental results and drawing conclusions; it should also be analytical and interpretative and should establish an association between the results obtained and other published works. The discussion may describe conflicting opinions and the results of other authors and indicate the value of these results for future works. This section should conclude with a few sentences that summarise the most relevant conclusions and implications. Conclusions do not contain references or enumerated/bulleted paragraphs, but provide a brief and precise summary of the most important findings of the work, their limitations, importance and future research needs.

Avoid combining the Results and Discussion sections into a single section (except in the case of Short communications). SJAR’s policy is to keep manuscripts merging the sections Results and Discussion only in those cases when this practice is strictly necessary, or adds some value to the work. In these cases, we require a formal statement by the authors explaining their reasons to do it so.

Acknowledgements. When it is considered necessary, acknowledgements should be made

to the people, centres or bodies that have collaborated or supported the carrying out of the work. Authors are responsible for obtaining the necessary permission of the people or bodies mentioned, given that the readers might infer that they endorse the data and conclusions of the article. Contributors who meet fewer than all four of the above mentioned criteria for authorship should not be listed as authors, but they should be acknowledged.

References. When references are cited in the text, the author's surname should be provided in parentheses, followed by a comma and the year of publication. If there are two authors, the surnames should be followed by "&". If there are three or more authors, include the surname of the first author followed by "et al." and a comma. When several references are cited, they should be ordered from oldest to most recent; if they are from the same year, sort them alphabetically. If there are two authors with the same surname and year of publication include the initial.

Tables and figures should be cited consecutively in the text, numbered independently with Arabic numerals and self-explanatory. Figures and tables must be very high quality and must be received in a suitable form and condition to be reproduced.

Tables should be headed by a number and title. Explanatory notes that facilitate the interpretation of the tables should be included at the bottom of the tables. Tables should have defined cells and must not be created using the space bar and/or tab keys.

Figures may correspond to diagrams or photographs. The figure number and legend should be presented at the bottom of the figure. After the acceptance of the paper, photographs should be sent separately as image files (jpg, tiff or similar) with a finished size of at least 300 dpi. Only under well-justified circumstances will colour photographs be admitted. Figures prepared with Excel or a similar program should be included in the text as MS Office Objects or sent separately in the format of their source program.

More details

<http://revistas.inia.es/index.php/sjar/about/submissions#authorGuidelines>