Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelos lineares generalizados espaciais mistos aplicados em estudos de insetos

Marcello Neiva de Mello

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

Piracicaba
2020

Marcello Neiva de Mello
Bacharel em Estatística

Modelos lineares generalizados espaciais mistos aplicados em estudos de insetos

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

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

Piracicaba
2020

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

Mello, Marcello Neiva de

Modelos lineares generalizados espaciais mistos aplicados em estudos de insetos / Marcello Neiva de Mello. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2020 .

42 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Produção 2. Unidade experimental 3. Parcela 4. Imputação múltipla 5. Dependência Espacial 6. Análise espaço-temporal . I. Título.

à Marcella e Isadora

AGRADECIMENTOS

Agradeço à Deus por todas as coisas boas e ruins que aconteceram nesse período, porque estas serviram de aprendizado e amadurecimento.

À minha esposa Lorena e minhas filhas Marcella e Isadora. Tudo é por vocês.

Um agradecimento para toda minha família, em especial à Ana (Mãe), José (Pai), Michelle (irmã) e Marcus (irmão), Lucita, Maria e Stelia (avós) por todo amor, cuidado e respeito.

À família Moreira pelo acolhimento, carinho e toda ajuda nessa caminhada.

Gostaria também de agradecer ao meu orientador, Carlos Tadeu, por sua amizade, apoio e conselhos, desde 2012.

Agradeço ao professor Guilherme Ludwig, Departamento de Estatística da Universidade Estadual de Campinas, pelas contribuições essenciais no Capítulo 2 e por me dar instruções em toda a tese.

Ao professor Elias Silva de Medeiros, Faculdade de Ciências Exatas e Tecnologia da Universidade Federal de Grande Dourados, pelas contribuições no Capítulo 3.

Não esquecendo de agradecer aos meus amigos Pedro Amoedo, Ana Cristina Garcêz, Djair Durand, Leomir, Hiron e Paula, Ricardo e Maria Cristina, Siglea, Arnaldo e Carol, Edilan e Ezequiel por todos os bons momentos em Piracicaba.

A todos os professores do Departamento de Ciências Exatas da ESALQ / USP, em especial a Sônia Maria de Stefano Piedade, César Gonçalves de Lima e Idemauro Antônio Rodrigues de Lara.

Aos funcionários do Departamento de Ciências Exatas da ESALQ / USP, Luciane Brajão, Solange Sabadin e Eduardo Bonilha.

À Universidade Federal Rural da Amazônia (UFRA), campus Capanema-PA, em especial aos professores e amigos Ebson Pereira, Geraldo Melo, Ivan Martins e ao seu projeto que forneceu o conjunto de dados.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), pelo financiamento em parte deste estudo - Código de Financiamento 001.

EPÍGRAFE

"A vida é muito curta para que se perca tempo numa existência medíocre."
Leandro Karnal

SUMÁRIO

RESUMO

Modelos lineares generalizados espaciais mistos aplicados em estudos de insetos

Experimentos são ferramentas comumente utilizadas para comparar ou melhorar alguma variedade de determinada cultura de produção. Milho, cana-de-açucar, café e laranja são alguns exemplos de culturas que muito são estudadas. Em particular, o milho é subsídio para alimentação humana e animal em forma de ração. E como toda produção em larga escala, está sujeita à presença de doenças e pragas. Portanto o monitoramento desses cultivos são ações importantes para prevenção e controle. Assim, coletas de uma mesma amostra, em determinados períodos acabam se tornado rotina e gera o que chamamos de série temporal e uma análise adequada deve ser feita, o que muitas vezes pesquisadores de áreas aplicadas não fazem. Portanto, uma análise espaço-temporal será abordada com a utilização de uma distribuição de Poisson inflacionadas de zeros (ZIP) para modelagem do componente de tendência, além da utilização de um variograma espaço temporal para captar o efeito não explicado por esse modelo. Um capítulo inteiro servirá de base à esses profissionais. É comum também em experimentos de campo existirem perdas de unidades observacionais, parcelas ou até mesmo de linhas inteiras por vários fatores que não são ou não podem ser controlados. Neste sentido, métodos de imputação são alternativas valiosas para preencher o valor de uma ou mais amostras que foram perdidas. Utilizando métodos de imputação múltipla já conhecidos na literatura como o MICE e o Hotdeck múltiplo, além do modelo de interpolação espacial SGLMM, a ideia foi fornecer cenários de perdas coerentes e comparar os diferentes métodos de imputação. O diferencial, em relação ao que já existe na literatura, é que agora levamos em conta a dependência espacial entre as amostras. Foi verificado no primeiro artigo que dentro de amostra, o MICE apresentou desempenho próximo ao SGLMM, todos os métodos apresentaram dificuldades de previsão, SGLMM apresenta melhor desempenho de extrapolação e Hotdeck múltiplo apresentou melhor desempenho em perdas limítrofes. No segundo artigo, verificou-se que habitats não cultivados podem favorecer a abundância de pragas, conforme colônias de pragas começaram a aparecer e a aumentar, houve o aumento de inimigos naturais que saíram de áreas adjacentes de floresta e que as moscas podem servir como indicador biológico no controle em culturas para amenizar o uso de agroquímicos nas lavouras de produção.

Palavras-chave: Produção, Unidade experimental, Parcela, Imputação múltipla, Dependência espacial, Análise espaço-temporal

ABSTRACT

Spatial generalized linear mixed models applied in insects studies

Experiments are tools commonly used to compare or improve some variety of a particular production culture. Corn, sugar cane, coffee and orange are some examples of cultures that are studied. In particular, maize is a subsidy for human and animal food in the form of feed. And like all large-scale production, it is liable to the presence of diseases and pests. Therefore, the monitoring of these crops are important actions for prevention and control. Thus, collections from the same sample, in certain periods, end up becoming routine and generate what we call a time series and an adequate analysis must be made, which many times researchers in applied areas do not do. Therefore, a space-time analysis will be approached with the use of an inflated Poisson distribution of zeros (ZIP) for modeling the trend component, in addition to the use of a space-time variogram to capture the effect not explained by this model. An entire chapter will serve as a basis to these professionals. It is also common in field experiments to have losses of observational units, plots or even entire lines due to various factors that are not or cannot be controlled. In this sense, imputation methods are valuable alternatives to fill the value of one or more samples that have been lost. Using multiple imputation methods already known in the literature, such as MICE and multiple Hotdeck, in addition to the spatial interpolation SGLMM model, the idea was to provide coherent loss scenarios and compare the different imputation methods. The difference, in relation to what already exists in the literature, is that we now take into account the spatial dependence between the samples. It was verified in the first article that within the sample, MICE performed close to SGLMM, all methods showed difficulties in forecasting, SGLMM presented better extrapolation performance and multiple Hotdeck showed better performance in borderline losses. In the second article, it was found that uncultivated habitats can favor the abundance of pests, as pest colonies began to appear and increase, there was an increase in natural enemies that left adjacent forest areas and that flies can serve as a biological indicator in crop control to mitigate the use of agrochemicals in production crops.

Keywords: Production, Experimental unit, Plot, Multiple imputation, Spatial dependence, Spatio-temporal analysis

1 INTRODUÇÃO

A produção agrícola é de suma importância, tanto de forma econômica como social. Os ganhos de produtividade, pós Segunda Guerra Mundial, geraram uma abundância de alimentos e fibras a um custo relativamente baixo para os consumidores domésticos e volume crescente de exportações para paises produtores (Capalbo e Antle, 2015). Ao longo dos anos foram desenvolvidos métodos melhores para preparar o solo e proteger plantas contra insetos, além da utilização de máquinas para maximizar a colheita e reduzir o trabalho humano. A implantação do GPS permitiu o melhor monitoramento da produção e de pragas, transformando as informações em mapas coloridos dando ao agricultor uma visão sobre várias técnicas de produção, condições climáticas e tipos de solos (Lowenberg-DeBoer, 2015).

A coleta e estudos dessas variáveis foram iniciadas pelo sul africano Daniel Krige (Krige, 1951), em jazidas minerais e posteriormente, o engenheiro francês George Matheron utilizou o termo Geoestatística, levando em consideração a localização geográfica e a dependência espacial dos dados, a partir da teoria das variáveis regionalizadas (Matheron, 1963), que tem por definição, ser uma função numérica, com uma continuidade espacial aparente e cuja variação de um local para outro deve ser incluída em tal modelo. Hoje variáveis dependentes espacialmente são amplamente utilizadas em áreas como agricultura, botânica, zoologia, hidrologia, transporte e até mesmo em estudos que utilizam o corpo humano como referência.

Dependendo da natureza da variável, isto é, se este é discreto ou contínuo, deve-se utilizar um modelo específico. Modelos de distribuição de espécies tornaram-se uma ferramenta fundamental em ecologia e tem amplas aplicações na avaliação das relações entre as espécies, o meio ambiente e o impacto das mudanças ecológicas (Franklin, 2013). A maioria dos estudos anteriores utilizavam dados de origem contínua, e quando discretos, utilizava-se transformações para gerar uma distribuição Gaussiana para a análise dos dados como pode ser visto em Cressie (1993). A partir de Diggle et al. (1998) extensões foram propostas o que ficou conhecido como model-based geostatistics. A proposta utiliza abordagem espacial para analisar dados de contagem e proporção, conhecidos como modelos lineares generalizados.

O desenvolvimento da teoria introduzida por Nelder e Wedderburn (1972) chamada de modelos lineares generalizados (MLG) unificou técnicas e metodologias que, até então, eram estudadas separadamente e lentamente se tornaram conhecidos e amplamente utilizados (Lindsey, 2000). Essa estrutura permite especifcar, por um lado, o primeiro e segundo momentos apenas, e por outro lado todas as pressuposições da distribuição. O MLG pressupõe que, seja Y uma variável aleatória que segue distribuição pertencente à família exponencial,

$$f(y;\theta,\phi) = e^{\phi^{-1}[y\theta - b(\theta)] + c(y,\phi)},$$

em que $b(.)$ e $c(.)$ são conhecidos, θ é o parâmetro canônico e ϕ é o parâmetro de dispersão, ambos são desconhecidos. É possível demonstrar que o valor esperado e a variância de Y com distribuição na família exponencial são $E(Y) = \mu = b^{'}(\theta)$ e $Var(Y) = \phi b^{''}(\theta)$. São distribuições pertencentes à família exponencial a distribuição Normal, Poisson, Binomial, Binomial Negativa, Gama e Normal Inversa (Molenberghs et al., 2007).

Uma extensão desses modelos são os modelos lineares generalizados espaciais mistos (SGLMM). O primeiro elemento dessa classe de modelos é o processo gaussiano estacionário $S(x)$. O campo aleatório gaussiano estacionário, $S(x)$, é um modelo Gaussiano se a distribuição conjunta de $S(x_1), \ldots, S(x_n)$ for uma Normal multivariada para qualquer valor de k inteiro no conjunto de localizações x_i. O processo é dito estacionário se a esperança de $S(x)$ for a mesma em toda área e a correlação entre dois pontos depender somente da distância entre eles, ou seja $||x_i - x_j||$, com $i \neq j$.

Considere n locais distintos $\{x_1, \ldots, x_n\} \in I$ e suponha que observemos a realização $y = (y_1, \ldots, y_n)^T$ de $Y = (Y_1, \ldots, Y_n)^T$. Supondo que $S = (S_1, \ldots, S_n)^T$ segue uma distribuição Normal com

vetor de médias $D\beta$ e matriz de covariâncias $\sigma^2\rho(u;\phi)+\varepsilon$. O SGLMM assume que $Y_1,\ldots,Y_n$ são condicionalmente independentes, dado o processo gaussiano $S(x)$, e distribuídos com densidade $f(.;\mu(s_i),\phi)$, sendo que $\mu(s_i)$ está relacionado ao preditor linear por meio da função de ligação $g(.)$ e ϕ é o parâmetro de dispersão, o qual aparecerá como parâmetro extra na função de verossimilhança. Logo o modelo é dado por

$$\begin{aligned} Y(x)|S(x) &\sim Poisson(\mu_i) \\ g(\boldsymbol{\mu}) = log(\mu_i) &= D\boldsymbol{\beta} + S(\mathbf{x}) = D\boldsymbol{\beta} + \sigma R(\mathbf{x};\phi) + \varepsilon. \end{aligned}$$

em que o processo $S(\mathbf{x}) \sim N(\mathbf{0},\Sigma)$ e pode ser decomposto em uma parte espacialmente dependente representado pelo termo $\sigma R(\mathbf{x};\phi)$ e outra espacialmente independente representado por $\boldsymbol{\varepsilon} \sim N(0,\tau^2 I)$.

O preditor linear consiste na soma de três elementos: os efeitos fixos $D\beta$, no efeito aleatório espacial correlacionado $R(\mathbf{x},\phi)$ e em uma parte independente ε. Impomos que, $R(\mathbf{x},\phi)$ é uma variância unitária com função de correlação $\rho(u,\phi)$, onde $u = ||x_i - x_j||$ é a distância Euclidiana entre dois pontos no espaço. Comumente, as funções de correlação presumem estacionariedade e são determinadas pelo variograma.

Sendo T como o domínio temporal, podemos expadir o processo para $S(x,t)$, um processo Gaussiano estacionário que modela desvios espacial e temporalmente (Gräler et al., 2015). Seja uma amostra $s = S(x_1,t_1),\ldots,S(x_n,t_n)$ observada em um conjunto de localizações espaço-temporais distintas $(x_1,t_1),\ldots,(x_n,t_n) \in X \times T \subseteq \Re^2 \times \Re$ que admite medições simultâneas em múltiplos locais no espaço. A função geral de covariância espaço temporal é dada por $C_{st} = Cov(S(x_i,t_i),S(x_j,t_j))$ para uma distância separável u e a distância temporal h e qualquer par de pontos $(x_i,t_i),(x_j,t_j) \in S \times T$ com $u = ||x_i - x_j||$ e $h = |t_i - t_j|$.

Modelos de covariâncias espaço temporal foram propostos inicialmente por Rouhani e Hall (1989), chamado modelo linear e é composta pela soma das covariâncias marginais. Mais tarde, De Cesare et al. (1997) propôs um modelo produto que utiliza o produto das covariâncias marginais. A desvantagem desses modelos é que eles não incorporam a interação e são ditos separáveis (covariâncias marginais não correlacionadas). Porém, modelos não separáveis levam em consideração essa interação espaço-tempo. Temos como exemplo o modelo produto soma e sua generalização (produto soma generalizado) ambos propostos por De Cesare et al. (2001). Uma descrição mais detalhada dos modelos está no anexo B.

Tanto no aspecto espacial quanto no espaço-temporal, perdas de observações podem ocorrer por diversos fatores. Os mais comuns são erros de digitação, destruição da parcela, plantas danifcadas, etc, gerando assim um conjunto de dados incompletos ou missing data. Alguns softwares são sensíveis a este tipo de situação eliminando um ou mais indivíduos da análise, caso alguma característica atribuída ao mesmo esteja *not available* ou NA.

Técnicas que possibilitem tratar tais tipos de dados foram propostas, dentre elas a imputação. Trata-se do preenchimento dos dados ausentes com valores plausíveis para uma posterior análise dos dados completos, sendo simples quando somente um valor é colocado para cada dado ausente, ou múltipla quando há mais de um valor em cada dado ausente (Van Buuren, 2018).

Neste contexto, no capítulo 2 será apresentado resultados de um estudo de simulação de diferentes métodos de imputação para diferentes cenários com perdas amostrais. No capítulo 3 será realizada uma abordagem espaço-temporal, o que não é comum nas áreas mais aplicadas. No capítulo 4 será apresentada uma conclusão geral.

Referências

Capalbo, S. M. e Antle, J. M. (2015). *Agricultural productivity: measurement and explanation*. Routledge.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

De Cesare, L., Myers, D., e Posa, D. (1997). Spatial-temporal modeling of so2 in milan district. In *5th International Geostatistical Congress*, volume 2, pages 1031–1042. Kluwer Academic Press.

De Cesare, L., Myers, D., e Posa, D. (2001). Estimating and modeling space–time correlation structures. *Statistics & Probability Letters*, 51(1):9–14.

Diggle, P. J., Tawn, J. A., e Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Franklin, J. (2013). Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions*, 19(10):1217–1223.

Gräler, B., Pebesma, E., e Heuvelink, G. (2015). Spatio-temporal geostatistics using gstat.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Lindsey, J. K. (2000). *Applying generalized linear models*. Springer Science & Business Media.

Lowenberg-DeBoer, J. (2015). The procision agriculture revolution. *Foreign Aff.*, 94:105.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.

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

Rouhani, S. e Hall, T. J. (1989). Space-time kriging of groundwater data. In *Geostatistics*, pages 639–650. Springer.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

2 COMPARAÇÃO DE IMPUTAÇÕES MÚLTIPLAS EM MODELOS LINEARES GENERALIZADOS ESPACIAIS MISTOS

Resumo

Estudos de campo estão expostos à possibilidade de perdas de unidades experimentais ou parcelas. Metodologias para imputação de dados perdidos são amplamente estudados pela literatura especializada, pois se não analisadas de forma coerente, podem trazer problemas na inferência do modelo final. O objetivo deste estudo qualitativo é fornecer uma visão de possíveis cenários de perdas de observações em experimentos de campo, utilizando imputações com a Média, MICE e Hotdeck múltiplo, junto ao modelo linear generalizado espacial misto (SGLMM) para melhorar estimativas no contexto espacial, para modelos de contagem. Para tal foram feitas simulações com diferentes sementes gerando cenários de experimentos cuja resposta é Poisson espacialmente estruturadas. Os cenários foram reproduzidos e imputados. Finalmente uma aplicação foi realizada em um experimento de produção de milho, realizado no município de igarapé-açu, Pará e a validação cruzada utilizada para investigação do comportamento em dados reais.

Palavras-chave: Imputação múltipla; MCMC; Campo aleatório Gaussiano; Modelo espacial.

2.1 Introdução

A modelagem espacial inicialmente foi introduzida por Krige (1951), para estudos de concentração de ouro na África. A modelagem que leva em consideração dados georreferenciados provenientes de contagem ou proporções é sempre mais complicada de ser analisada, pois requer a resolução de uma integral de alta dimensão para os efeitos aleatórios espacialmente correlacionados e necessitam de métodos numéricos para estimação dos parâmetros, (Bonat e Ribeiro Jr, 2016).

Como em todo estudo de campo, perdas amostrais são muito comuns em diversas áreas de pesquisa e estão envidenciadas em Liu et al. (2019), Zhang et al. (2019), Dai e Müller (2019), Naegele et al. (2017), entre outros. Essas interferências na área experimental, diferente de estudos em laboratório com ambientes controlados, nem sempre podem ser previstas e evitadas. Existem vários fatores abióticos como excesso de chuvas, ventos e enxurradas. Também bióticos como, por exemplo, tratores ou animais que destroém uma ou mais armadilhas, pessoas diferentes fazendo amostragem, doenças ou pragas que matam a plantação, deriva de produtos químicos de áreas adjacentes, interações ecológicas, dentre outras que podem gerar um *missing data*.

Porém, a falta de dados é um problema comum e recorrente e o manuseio indevido de dados perdidos pode comprometer seriamente a validade das inferências de um estudo (Lang e Little, 2018). Na intenção de contornar essa situação, alguns *softwares* eliminam totalmente um ou mais indivíduos da análise, caso falte alguma informação. Técnicas que possibilitem tratar tais tipos de dados foram propostas, dentre elas a imputação. Trata-se do preenchimento dos dados ausentes com valores plausíveis para uma posterior análise dos dados completos, sendo dita simples quando somente um valor é colocado para cada dado ausente ou múltipla quando há mais de um valor em cada dado ausente (Van Buuren, 2018).

Neste sentido, o presente trabalho tem como objetivo comparar resultados de um estudo de simulação de diferentes métodos de imputação para diferentes cenários com perdas amostrais, fornecendo ao pesquisador uma ferramenta de decisão quanto à viabilidade de utilização ou não das técnicas. Diferentes cenários serão utilizados para simular situações que comumente podem ocorrer em áreas experimentais.

2.2 Material e Métodos

2.2.1 Modelos espaciais

Modelos lineares generalizados (GLM) foram introduzidos por Nelder e Wedderburn (1972) unificando técnicas e metodologias que, até então, eram estudadas separadamente e lentamente se tornaram conhecidas e usadas (Lindsey, 2000). No GLM, uma variável resposta $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ é assumida, sendo mutuamente independente com valor esperado relacionado à um preditor linear $E[\boldsymbol{Y}] = g^{-1}(\boldsymbol{d}^{'}\boldsymbol{\beta})$, em que $\boldsymbol{\beta}$ é um vetor de parâmetros de regressão desconhecido, $\boldsymbol{d}$ são as variáveis explanatórias conhecidas e $g(.)$ é uma função conhecida chamada de função de ligação.

Uma extensão dos GLM é o modelo linear generalizado espacial misto (SGLMM), em que as variáveis resposta são consideradas independentes umas das outras condicionalmente aos valores de um conjunto de variáveis latentes, que nesse caso, derivam de um processo espacial (Diggle et al., 1998).

Considere n locais $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, em que $\boldsymbol{x}_i \in R^2$ e assuma que a realização $\boldsymbol{y} = (y_1, \ldots, y_n)$ de variáveis aleatórias $\boldsymbol{Y}$ é observada. Seja $S(\boldsymbol{x})$ um processo Gaussiano com esperança $E[S(\boldsymbol{x}_i)] = 0$ e covariância $Cov[S(\boldsymbol{x}_i), S(\boldsymbol{x}_j)] = \sigma^2\rho(u, \phi) + \tau^2\mathbf{1}$, em que $\rho(u, \phi)$ é a função de correlação e $u = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$ é a distância Euclidiana entre um par de pontos no espaço, ϕ é o parâmetro de escala que controla a velocidade na qual as correlações espaciais se aproximam de 0 à medida que a distância entre os locais aumenta. O parâmetro τ^2 é o efeito pepita. Supõe-se que os componentes em $\boldsymbol{Y}(\boldsymbol{x})$ sejam condicionalmente independentes, dado um processo espacial Gaussiano $S(\boldsymbol{x})$ e $E[Y(x_i)|S(x_i)] = g^{-1}(S(x_i))$.

Para estimar os parâmetros de um SGLMM $\boldsymbol{\theta} = (\beta, \sigma^2, \tau^2, \phi)$, é necessário maximizar a função de verossimilhança marginal, obtida pela integração dos efeitos aleatórios $S(\boldsymbol{x})$ da distribuição conjunta dos termos aleatórios no modelo, $L(\boldsymbol{\theta}) = \int f(\boldsymbol{y}(\boldsymbol{x})|S(\boldsymbol{x}))f(S(\boldsymbol{x}))dS(\boldsymbol{x})$, em que a verossimilhança é o produto de duas distribuições e não é possível obter uma forma fechada. Observe que a função condicional é uma $Poisson(\mu_i)$ com função de ligação $g(.)$, portanto o SGLMM assume o formato

$$\begin{aligned} y(\boldsymbol{x})|S(\boldsymbol{x}) &\sim Poisson(\mu) \\ g(\mu) &= log(\mu) \\ Cov[S(\boldsymbol{x}_i), S(\boldsymbol{x}_j)] &= exp(-u/\phi) \end{aligned}$$

Um caso especial é quando a função condicional é gaussiana e a probabilidade assume uma forma fechada. Cadeia de Markov e Monte Carlo (MCMC) é um método alternativo para ajustar modelos espaciais a dados não gaussianos (Brown et al., 2015). Diggle et al. (1998) introduziu o MCMC para cálculo dos parâmetros espaciais num contexto Bayesiano, aparecendo em mais estudos como Christensen e Waagepetersen (2002), Diggle et al. (2003), Christensen (2004). A descrição à seguir será baseada em Diggle e Ribeiro (2007).

O método MCMC consiste em gerar amostras das distribuições à posteriori associadas a modelos de verossimilhança com informação à priori sobre os parâmetros. Seja $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ vetores já definidos anteriormente, $\boldsymbol{S}$ o vetor de valores observados de $S(\boldsymbol{x}_i)$, $\boldsymbol{Y}$ o vetor correspondente da variável resposta $\boldsymbol{Y}$ e $\boldsymbol{S}^*$ o vetor de valores preditos do campo aleatório Gaussiano $S(\boldsymbol{x})$. O MCMC gera amostras à posteriori da distribuição de $[\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{Y}]$ para estimação dos parâmetros e $[\boldsymbol{S}^*|\boldsymbol{Y}]$ para previsão. O algoritmo consiste em:

- Passo 1: selecionar valores para $\boldsymbol{\theta}, \boldsymbol{\beta}$ e $\boldsymbol{S}$;
- Passo 2: atualizar todos os componentes de $\boldsymbol{\theta}$:
  - (i) escolha um novo valor para $\boldsymbol{\theta}^{'}$ amostrando de uma distribuição de candidatos definida no espaço paramétrico;

(ii) aceite $\boldsymbol{\theta}^{'}$ com probabilidade $\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^{'}) = min\left\{\frac{p(\boldsymbol{S}|\boldsymbol{\theta}^{'})}{p(\boldsymbol{S}|\boldsymbol{\theta})}, 1\right\}$, caso contrário não altere $\boldsymbol{\theta}$.

- Passo 3: Atualize $\boldsymbol{S}$,

  (i) escolha um novo valor proposto, $\boldsymbol{S}_i^{'}$, para o i-ésimo componente de $\boldsymbol{S}$ da densidade de probabilidade condicional Gaussiana univariada $p(\boldsymbol{S}_i^{'}|\boldsymbol{S}_{-i}, \theta)$, onde $\boldsymbol{S}_{-i}$ denota $\boldsymbol{S}$ com seu i-ésimo elemento removido;

  (ii) aceite $\boldsymbol{S}_i^{'}$ com probabilidade $\Delta(\boldsymbol{S}_i, \boldsymbol{S}_i^{'}) = min\left\{\frac{p(y_i|s_i^{'}, \boldsymbol{\beta})}{p(y_i|s_i, \boldsymbol{\beta})}, 1\right\}$, caso contrário não altere $\boldsymbol{S}_i$;

  (iii) repita (i) e (ii) para todo $i = 1, \ldots, n$.

- Passo 4: Atualize todos os elementos de $\boldsymbol{\beta}$:

  (i) escolha um novo valor para $\boldsymbol{\beta}^{'}$ da densidade condicional $p(\boldsymbol{\beta}^{'}, \boldsymbol{\beta})$;

  (ii) aceite $\boldsymbol{\beta}^{'}$ com probabilidade $\Delta(\boldsymbol{\beta}, \boldsymbol{\beta}^{'}) = min\left\{\frac{\prod_{j=1}^{n} p(y_j|s_j, \boldsymbol{\beta}^{'}) p(\boldsymbol{\beta}, \boldsymbol{\beta}^{'})}{\prod_{j=1}^{n} p(y_j|s_j, \boldsymbol{\beta}) p(\boldsymbol{\beta}^{'}, \boldsymbol{\beta})}, 1\right\}$, caso contrário não altere $\boldsymbol{\beta}$.

O conhecido *burn-in* é a repetição dos passos 2-4. Todas essas iterações iniciais são descartadas, até que se julgue que a cadeia alcançou sua distribuição de equilíbrio. Para fins de previsão, agora deve-se gerar amostras da distribuição condicional $[(\boldsymbol{S}, \boldsymbol{S}^*)|\boldsymbol{Y}] = [\boldsymbol{S}|\boldsymbol{Y}][\boldsymbol{S}^*|\boldsymbol{S}, \boldsymbol{Y}]$, utilizando o mesmo algoritmo anteriormente apresentado acrescentando apenas mais uma etapa, passo 5, que consiste em desenhar uma amostra aleatória da distribuição gaussiana multivariada $[\boldsymbol{S}^*|\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{S}]$, em que $(\boldsymbol{\theta}, \boldsymbol{S}, \boldsymbol{\beta})$ são os valores gerados nas etapas 2 a 4.

2.2.2 Métodos de imputação

Hotdeck múltiplo, introduzido por Cranmer e Gill (2013) e MICE (Van Buuren e Oudshoorn, 2000) são exemplos dos diversos métodos de imputação existentes na literatura. Estes podem ser classificados como métodos de imputação simples, pois o valor ausente é preenchido apenas uma vez, ou imputação múltipla livre de distribuição (Bergamo et al., 2008), porém menos explorado.

A imputação múltipla (Rubin, 1978) se baseia no preenchimento dos valores ausentes m vezes e representa a incerteza sobre o valor a ser alocado. Tal procedimento é composto por três etapas. A primeira é a imputação, em que os valores ausentes são estimados m ($m > 1$) vezes, gerando m conjuntos de dados completos. No segundo passo, os m conjuntos de dados completos são analisados e parâmetros estimados. Na última fase, os m resultados obtidos serão agrupados (Figura A.1, apêndice I).

O algoritmo MICE (Multivariate Imputation by Chained Equation) se baseia em MCMC, que amostra as distribuições condicionais para obter amostras da distribuição conjunta. O método proposto por Van Buuren e Groothuis-Oudshoorn (2011) utiliza o seguinte algoritmo: Seja $(Y, \boldsymbol{d})$ a variável completa a ser estudada e tendo como Y o vetor da variável resposta, e $\boldsymbol{d}$ o vetor de covariáveis (coordenadas geográficas), em que na parte obervada temos $(Y_i^O, \boldsymbol{d}_i^O)$, com $i = 1, \ldots, n$, e não observada representada por $(Y_i^M, \boldsymbol{d}_i^M)$, com $i = n+1, \ldots, n+m$. Logo, temos $(Y, \boldsymbol{d})$ como um vetor multivariado o qual será utilizado para a imputação pelo método. O j-ésimo conjunto de dados à ser imputado é Y^j com $j = 1 \ldots, m$. Assumindo que Y seja completamente especificado por $\boldsymbol{\theta}$, um vetor de parâmetros desconhecidos, temos que $P(Y|\boldsymbol{\theta})$ seja a distribuição. O algoritmo MICE obtém a distribuição à posteriori de $\boldsymbol{\theta}$ por amostragem iterativa a partir das distribuições condicionais da forma

$$\begin{aligned} &P(Y|d_1, d_2, \theta) \\ &P(d_1|Y, d_2, \theta) \\ &P(d_2|Y, d_1, \theta) \end{aligned} \tag{2.1}$$

O vetor de parâmetros θ não é necessariamente o produto de uma fatoração da distribuição conjunta $P(Y|\theta)$, mas é específico de cada distribuição em 2.1. A t-ésima iteração em cada cadeia é um amostrador de Gibbs que gera

$$\begin{aligned} \theta^{*(t)} &\sim P(\theta|Y^O, d_1^{(t-1)}, d_2^{(t-1)}) \\ Y^{*(t)} &\sim P(Y|Y^O, d_1^{(t-1)}, d_2^{(t-1)}) \end{aligned}$$

em que $Y_j^{(t)} = Y_j^{*(t)}$ é o j-ésima variável imputada na iteração t. O método utilizado foi o ppm (predictive mean matching), o qual forma um pequeno conjunto de candidatos a doadores de todos os casos completos que possuem valores preditos mais próximos do valor predito para a observação ausente. Um doador é sorteado aleatoriamente dos candidatos, e o valor observado do doador é retirado para substituir o valor ausente.

No método Hotdeck Múltiplo (Cranmer e Gill, 2013), valores observados em todo conjunto de dados são usados para preencher os valores ausentes, apresentando a vantagem de manter as propriedades de uma variável discreta, por exemplo, caso esta seja imputada. A imputação consiste em pesquisar coluna(s) com dados ausentes e, caso um valor ausente seja encontrado, calcular um vetor de pontuações de afinidade α para esse valor ausente. Os melhores doadores serão selecionados aleatoriamente para produzir um vetor de imputações e esse valor será atribuído a cada um dos m conjunto de dados duplicados. A pontuação de afinidade mede a proximidade dos dados ausentes com os dados observados. Essa pontuação varia de 0 a 1 e mede o grau de similaridade que cada receptor i tem para cada potencial doador j e é dado por

$$\alpha_{ij} = \frac{k - q_i - z_{ij}}{k - q_i},$$

em que k é o número de variáveis, q_i é o número de valores ausentes na i-ésima posição, z_{ij} é o número de variável(is) para a(s) qual(is) o potencial doador j e o destinatário i possui(em) valor(es) diferente(s). É importante salientar que z_{ij} os autores apresentam uma distância multivariada e no nosso caso essa distância é no espaço, a distância Euclidiana $||x_i - x_j||$.

Devido a dificuldade de atender suposições de distribuições e pelos métodos de imputação múltiplas serem mais robustos, o presente estudo selecionou o método da Média, MICE e Hotdeck Múltiplo, sendo que estes são métodos flexíveis quanto à suposição de distribuição. O primeiro é considerado uma forma rápida e simples, visando substituir os valores faltantes pela média geral da variável. Após imputado, a análise é realizada e os parâmetros são estimados. É importante destacar que o método da média é considerado ruim quando utilizado para imputações (Van Buuren e Groothuis-Oudshoorn, 2011).

As imputações múltiplas, realizadas no MICE e Hotdeck múltiplo foram realizadas com $m = 5$, ou seja, 5 bancos foram imputados de forma independente, realizadas análises e posteriormente combinadas. O software `R` (R Core Team, 2019) possui um pacote chamado `mice` (Van Buuren e Groothuis-Oudshoorn, 2011) e este pacote possui ainda outros métodos de imputação univariada, além do pacote `hot.deck` (Cranmer et al., 2016) o qual foi utilizado para realização do método hotdeck múltiplo.

2.2.3 Validação do modelo

Duas abordagens foram aplicadas para avaliar a precisão das previsões feitas pelas imputações e pelo modelo SGLMM. São elas o sample mean square error (SMSE) e o prediction mean square error (PMSE). No SMSE queremos calcular o erro, levando em consideração somente os valores observados na amostra, para ver como se comporta a estimação do modelo SGLMM com perdas. No PMSE, calculamos apenas o erro associado às observações perdidas e imputadas (Y_i^{imp}) (somente no estudo de simulação). Seja n o número de observações na amostra e n_{mis} o número de observações perdidas, as fórmulas para cálculo das medidas são dadas respectivamente por

$$SMSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

$$PMSE = \frac{1}{n_{mis}}\sum_{i=1}^{n_{mis}}(Y_i^{imp} - \hat{Y}_i^{imp})^2.$$

2.2.4 Estudo de simulação e cenários

O estudo de simulação tem como principal objetivo verificar, por meio do SMSE e PMSE, o comportamento dos diferentes cenários com perdas, imputá-los e compará-los com cenários sem perdas. Pela pesquisa que fizemos, estudos qualitativos de cenários são escassos na literatura de análise de dados georreferenciados, portanto a ideia é testar o quão robusto é o método e como funciona em tipo de perdas diferentes.

Um grid regular com 100 parcelas foi adotado como área de estudo e será detalhado na próxima sessão. Utilizou-se as 100 primeiras sementes do `R`, para gerar campos aleatórios Gaussianos com parâmetros fixos de covariância em $\phi = 100$ e $\sigma^2 = 5$, além do efeito pepita $\tau^2 = 0.5$, utilizando um modelo de covariância exponencial e não levando em consideração possíveis tendências. Esses campos foram condicionados para gerar um vetor resposta $y \sim Poisson(\mu)$. Os cenários foram replicados e os pontos removidos. Cada método de imputação foi utilizado para completar os conjuntos de dados, além da estimação por SGLMM (sem imputação) que possibilitou fazer previsões aos pontos retirados e servirá como referência. O processo esta representado na Figura 2.1. As estimativas dos parâmetros ($\hat{\boldsymbol{\theta}}$) foram calculadas com o algoritmo MCMC existente no pacote `geoRglm` (Christensen e Ribeiro Jr, 2002) com *burn-in* de 100.000 e *thining* de 20. Devido a instabilidade numérica de todos os métodos utilizados utilizou-se um corte de 20% para exclusão das piores estivativas que ocorreram por conta de convergências ruins das cadeias, gerando assim uma visualização mais agradável.

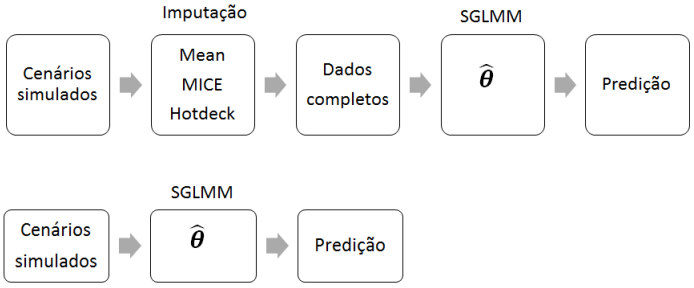

Figura 2.1. Diagrama representativo do processo de simulação e estimação para métodos de imputação e SGLMM

Assim, foram propostos nove cenários diferentes com tipos e percentuais diferentes de perdas apresentados na Figura 2.2. Os pontos em destaque preto simbolizam as amostras perdidas. O cenário 1 explora a situação de perda de uma linha inteira na borda do grid. No 2 e no 3 cenário temos perdas de blocos na borda e blocos na parte central, respectivamente, simulando pragas que agem de forma agregada. Já nos cenários 4, 5, 6 e 7 apresentam perdas aleatórias de 5, 10, 12 e 15%, respectivamente. Os dois últimos apresentam perdas mais expressivas de 30%, e tendem a ser mais extremas e servirão para ver como se comportarão os métodos em casos atípicos.

A Figura 2.3 (a) apresenta boxplots para a raiz quadrada dos erros quadráticos médios amostrais das simulações e os diferentes tipos de imputação para nove cenários. Nela verifica-se que no cenário 1,

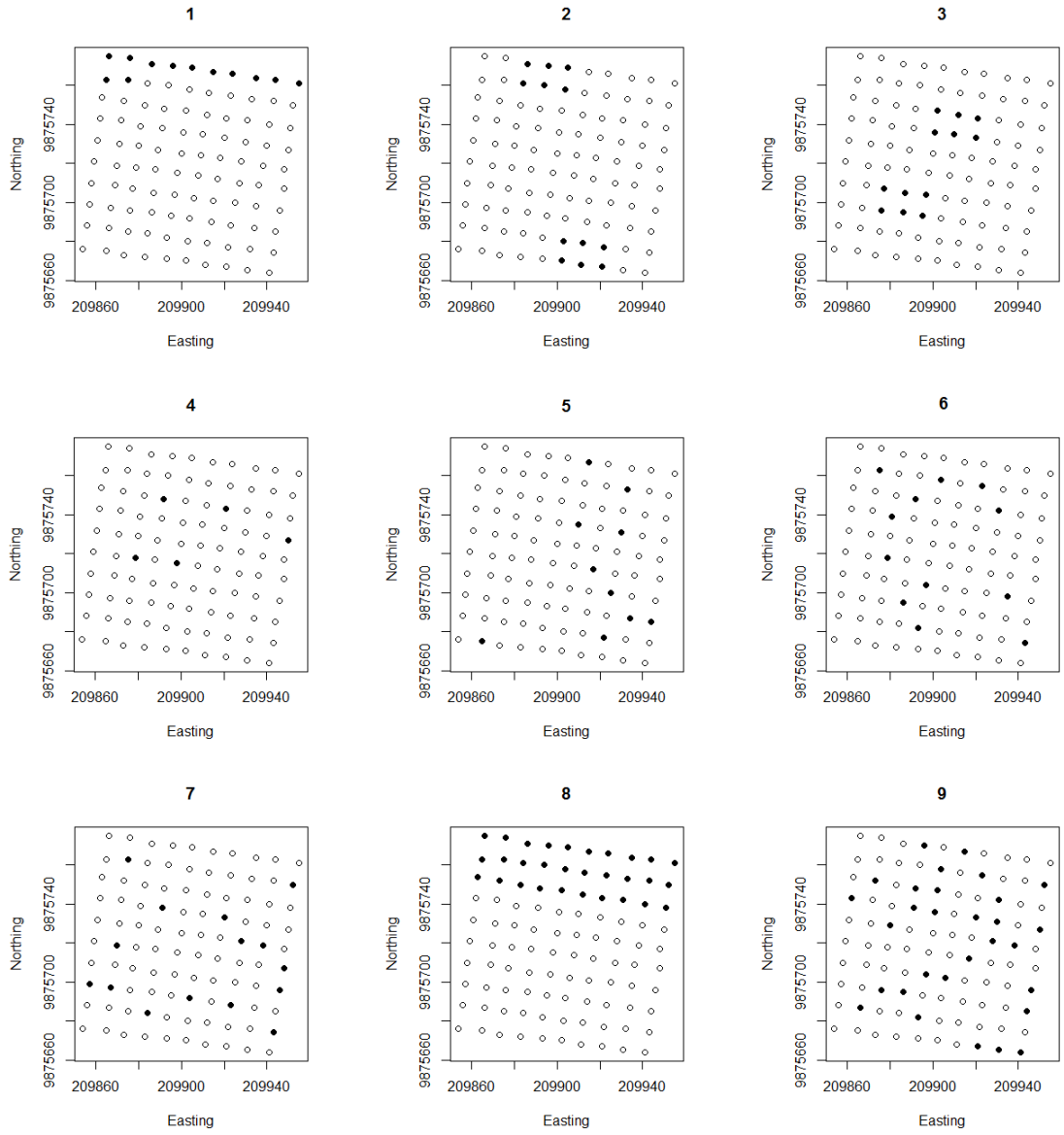

Figura 2.2. Cenários gerados com perdas de pontos amostrais em experimento de campo, sendo os pontos em destaque preto as amostras perdidas. (1) Perda de pontos na borda; (2) perda de blocos na borda; (3) perda de blocos no centro; (4) perda de 5% de pontos aleatórios; (5) perda de 10% de pontos aleatórios; (6) perda de 12% de pontos aleatórios; (7) perda de 15% de pontos aleatórios; (8) perda de 30% na borda; (9) perda de 30% de pontos aleatórios

o MICE e Hotdeck múltiplo apresentaram comportamento semelhante ao SGLMM, sendo o SGLMM ligeiramente melhor. Já nos cenários 2, 3 e 6 o método Hotdeck múltiplo apresenta média menor e o MICE se iguala ao SGLMM. Nos cenários 4 e 5 todos parecem ter comportamentos semelhantes com algumas alterações de variabilidade. Em 7 temos o MICE com comportamento próximo ao SGLMM e o Hotdeck múltiplo apresenta média mais baixo do que o método da média. Porém, com o aumento das perdas em 8 e 9, a média se mostra ineficiente, Hotdeck múltiplo e MICE apresentam variabilidade maior do que SGLMM.

A Figura 2.3 (b) apresenta boxplots para a raiz quadrada dos PMSE das simulações e os tipos de imputação para nove cenários. Nela verifica-se que o método Hotdeck múltiplo apresenta melhor desempenho na predição de parcelas limitrofes (cenários 1, 2 e 8), porém no cenário 3, 4, 5, 6 e 7 o SGLMM apresenta melhor comportamento e, por fim, os cenários 8 e 9, apesar da instabilidade dos métodos, todos estão semelhantes. Vale destacar que, em praticamente em todos os cenários, o MICE apresentou dificuldades de predição.

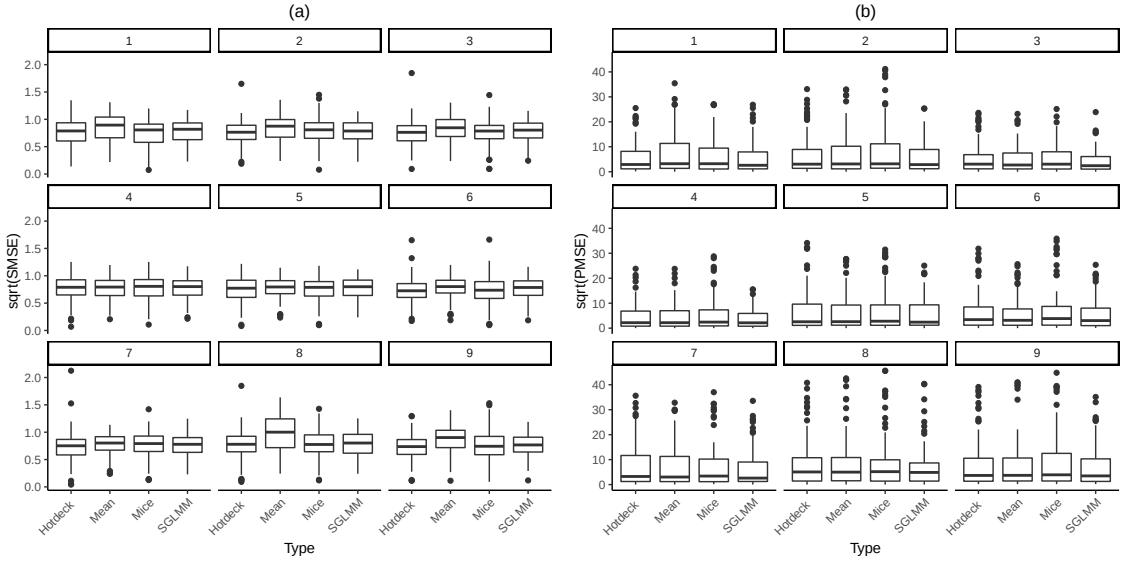

Figura 2.3. Boxplot: (a) SMSE square root and (b) PMSE square root para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

Examinaremos agora a performance em relação à estimativa dos parâmetros simulados. A Figura 2.4 apresenta diagramas de dispersão para (a) raiz quadrada do SMSE *versus* viés do parâmetro $\hat{\phi} - \phi$, (b) raiz quadrada do SMSE *versus* log diferença de $\hat{\tau}^2 - \tau^2$ e (c) raiz quadrada do SMSE *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$ para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação. Em 2.4 (a) verifica-se uma discreta relação positiva entre o viés do patamar e o erro amostral. Em 2.4 (b) uma discreta relação negativa entre o viés do estimador $\hat{\tau}^2 - \tau^2$ e o erro amostral. Por fim, em 2.4 (c) o erro amostral e o viés da variabilidade espacial aparentam não ter relação.

A Figura 2.5 apresenta diagramas de dispersão para (a) raiz quadrada do PMSE *versus* log diferença de $\hat{\tau}^2 - \tau^2$, (b) raiz quadrada do PMSE *versus* viés do parâmetro $\hat{\phi} - \phi$ e (c) raiz quadrada do PMSE *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$ para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação. Na Figura 2.5 (a) podemos verificar que não existe relação aparente entre a raiz quadrada do PMSE e o viés da variância aleatória ($\hat{\tau}^2 - \tau^2$), ou seja o erro de previsão não possui relação aparente com a combinação de erros de amostragem e variações que ocorrem em escalas menores que a distância entre os pontos amostrados (Curran (1988), Cressie (1993)). Pela Figura 2.5 (b) é possível verificar que, de forma geral, à medida que o viés da amplitude ($\hat{\phi} - \phi$) aumenta, o erro de predição

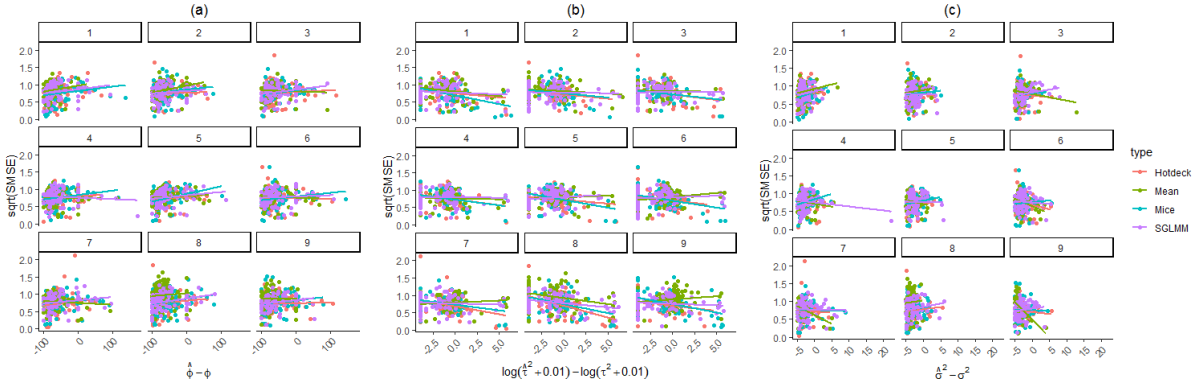

Figura 2.4. Diagramas de dispersão: (a) raiz quadrada do SMSE *versus* viés do parâmetro $\hat{\phi} - \phi$; (b) raiz quadrada do SMSE *versus* log diferença de $\hat{\tau}^2 - \tau^2$; (c) raiz quadrada do SMSE *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$, para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

também aumenta, ou seja, quanto maior a distância em que os dados deixam de serem correlacionados (Silveira et al., 2017), maior o erro de previsão. Também na Figura 2.5 (c) pode-se verificar uma aparente correlação positiva, ou seja, à medida que o viés da variação explicada pela estrutura espacial ($\hat{\sigma}^2 - \sigma^2$) aumenta, o erro de previsão tende a aumentar.

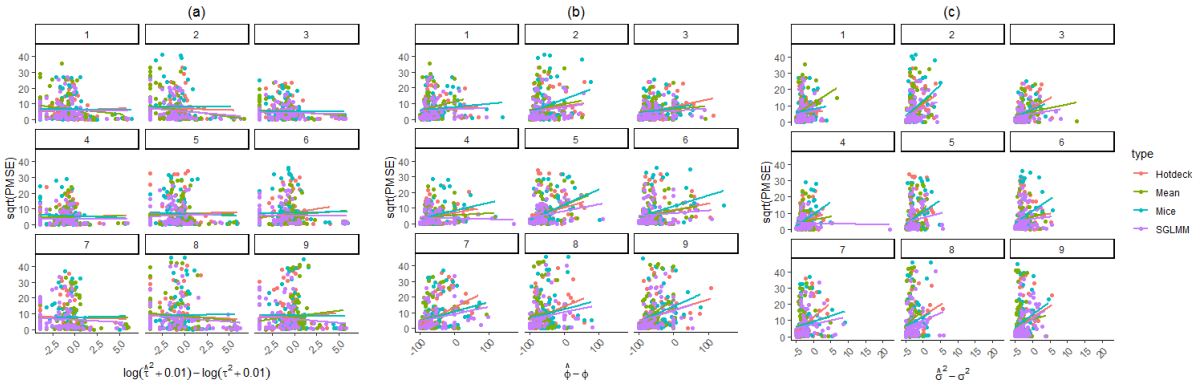

Figura 2.5. Diagramas de dispersão: (a) raiz quadrada do PMSE *versus* log diferença de $\hat{\tau}^2 - \tau$; (b) raiz quadrada do PMSE *versus* viés do parâmetro $\hat{\phi} - \phi$; (c) raiz quadrada do PMSE *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$, para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

A Figura 2.6 apresenta diagramas de dispersão para (a) viés do parâmetro $\hat{\phi} - \phi$ *versus* log diferença de $\hat{\tau}^2 - \tau^2$, (b) viés do parâmetro $\hat{\sigma}^2 - \sigma^2$ *versus* log diferença de $\hat{\tau}^2 - \tau^2$ e (c) viés do parâmetro $\hat{\phi} - \phi$ *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$ para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação. É possível verificar em 2.5 (a) uma discreta correlação negativa entre $\hat{\phi} - \phi$ e $\hat{\tau}^2 - \tau^2$, ou seja, à medida que o patamar aumenta o erro combinado tende a diminuir. Em 2.6 (b), à medida que o erro combinado aumenta, a variabilidade espacial tende à diminuir. Por fim, a correlação entre $\hat{\phi} - \phi$ e $\hat{\sigma}^2 - \sigma^2$ é positiva, ou seja, à medida que o patamar aumenta, a variabilidade espacial também aumenta.

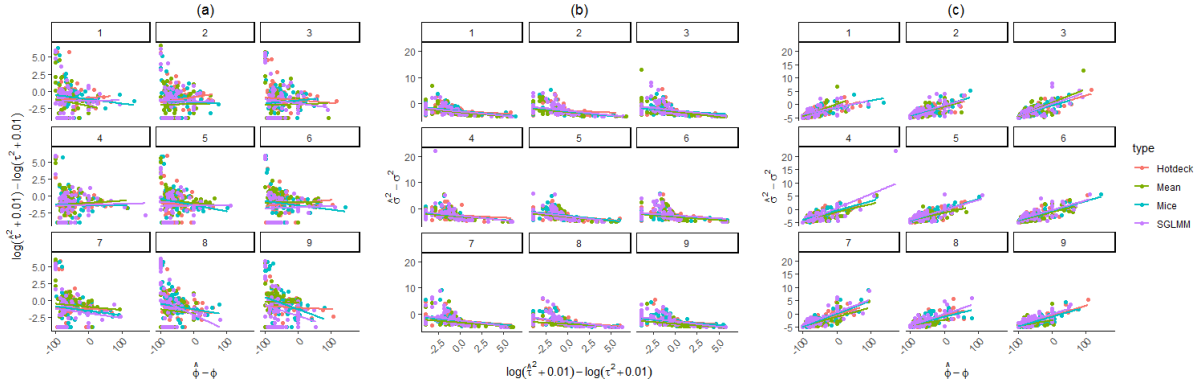

Figura 2.6. Diagramas de dispersão: (a) viés do parâmetro $\hat{\phi} - \phi$ *versus* log diferença de $\hat{\tau}^2 - \tau^2$; (b) viés do parâmetro $\hat{\sigma}^2 - \sigma^2$ *versus* log diferença de $\hat{\tau}^2 - \tau^2$; (c) viés do parâmetro $\hat{\phi} - \phi$ *versus* viés do parâmetro $\hat{\sigma}^2 - \sigma^2$, para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

2.3 Aplicação em contagem de moscas sirfídeos

2.3.1 Área de estudo e amostragem

Esse estudo foi realizado em 2016 na Fazenda Experimental de Igarapé-açu (FEIGA) no município de Igarapé-açu, estado do Pará, Brasil (01º07'33"S; 47º37'27''W; altitude 39 m). A área possui 10.000 m^2 e foi dividida em 100 parcelas de $10 \times 10m^2$ cada. A cultivar de milho utilizada foi híbrida. A semeadura foi realizada em 10/04/16 com espaço entre linhas de 0,90 m e 0,15 m por planta e sem intervenção química. A amostragem foi realizada em 22/05/16. Em cada parcela, 10 plantas foram selecioandas aleatoriamente, totalizando 1.000 plantas. A ocorrência de inimigos naturais foi avaliada visualmente e o número de indivíduos por planta foi determinado.

2.3.2 Resultados

A partir da Figura 2.7, é possível verificar uma análise exploratória dos dados de moscas sirfideos, as quais são inimigos naturais de pragas que ocorrem na produção de milho, os pulgões-do-milho. Essas moscas são possíveis indicadoras biológicas da ocorrência dessas pragas. Na Figura 2.7 (a), a sequência ($\circ$, $\triangle$, $+$, $\times$) segue ordem crescente, portanto é possível verificar uma aparente dependência espacial entre as ocorrências de moscas sirfídeos. Nas Figuras 2.7 (b) e (c) é possível visualizar uma leve tendência das observações e, por fim, em 2.7 (d) é possível identificar um comportamento semelhante ao de uma distribuição Poisson.

Assim, o conjunto de dados foi submetido ao processo apresentado no estudo de simulação. Algumas observações foram retiradas, os cenários foram gerados e os métodos de imputação foram aplicados, além do método de interpolação com o modelo SGLMM, sem imputação.

A Figura 2.8 apresenta os valores observados *versus* valores preditos de cada cenário, para cada método de imputação e método de interpolação com SGLMM. É possível verificar no cenário 1, que todos os método apresentaram problemas para estimação, sendo a média a que apresentou menor contribuição. No cenário 2, a média continua tendo uma péssima contribuição, já o Hotdeck múltiplo e o MICE parecem prever melhor os valores retirados. No cenário 3 e 6, o método Hotdeck múltiplo apresentou melhor desempenho aparente, estando mais próximo à reta. No cenário 4, todos apresentaram valores acima da reta, indicando uma possível superestimação para perdas aleatórias menores. No cenário 5, o método MICE e o SGLMM apresentaram maior proximidade à reta. Já com o cenário 7, o qual apresenta perda maior (15%), os métodos apresentaram dificuldades mais expressivas, quando comparados com perdas

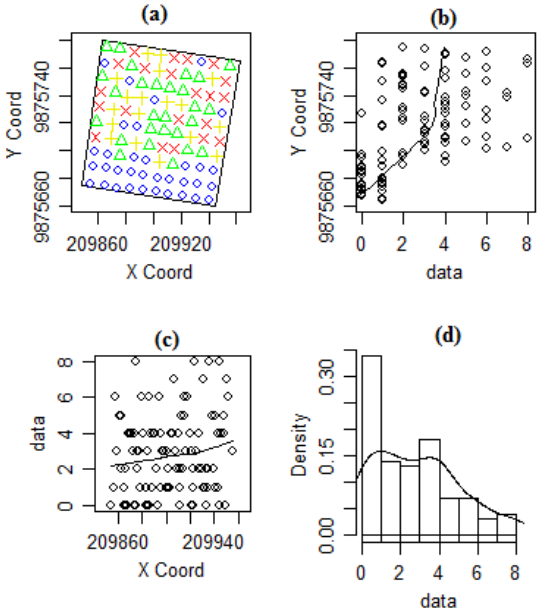

Figura 2.7. Análise exploratória: (a) localização dos dados; (b) dados *versus* coordenadas Y; (c) dados *versus* coordenadas X e (d) histograma

semelhantes ao cenário 5 (10%) e 6 (12%), porém o SGLMM aparenta um desempenho melhor. Perdas sistemáticas como a do cenário 8 revela dificuldades de todos os métodos, tanto de imputação, como de interpolação, porém em perdas aleatórias relativamente grandes, representado no cenário 9, houve um interessante desempenho do MICE em relação aos demais.

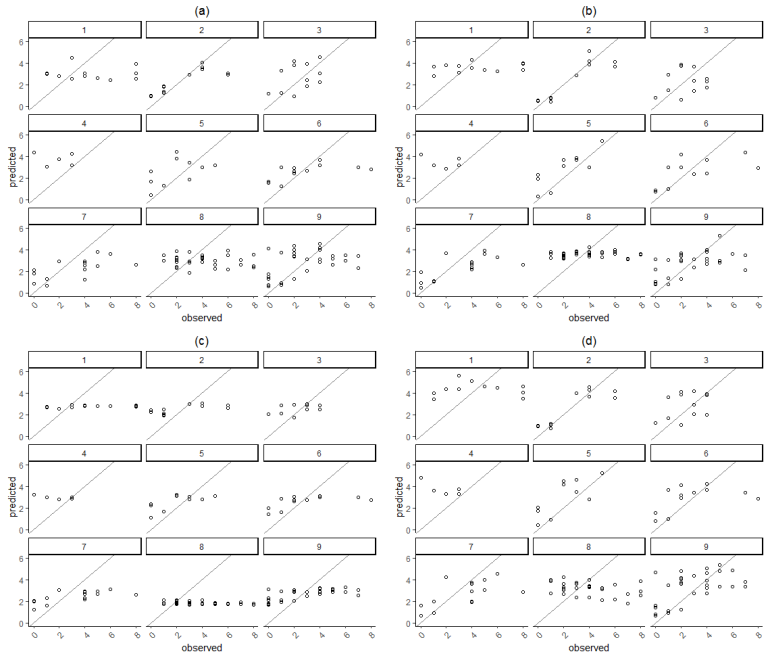

Figura 2.8. Valores retirados dos cenários *versus* preditos da imputação (a) Hotdeck múltiplo; (b) MICE; (c) Média e (d) SGLMM no banco de dados real. (1) Perda de pontos na borda; (2) perda de blocos na borda; (3) perda de blocos no centro; (4) perda de 5% de pontos aleatórios; (5) perda de 10% de pontos aleatórios; (6) perda de 12% de pontos aleatórios; (7) perda de 15% de pontos aleatórios; (8) perda de 30% na borda; (9) perda de 30% de pontos aleatórios

2.4 Conclusão

Para o estudo dentro da amostra, sem levar em consideração os valores retirados, o método MICE apresenta, em geral, resultados mais próximos do SGLMM. Todos os métodos apresentaram certa dificuldade de predição, principalmente quando se trata de perdas mais expressivas, porém o SGLMM se apresentou melhor para extrapolação. Vale destacar que nos dados reais, o Hotdeck múltiplo apresentou melhor desempenho na predição de perdas limítrofes (cenários 1, 2 e 8). É importante destacar ainda que o MICE pode não ter apresentado resultados agradáveis por não estar sendo usando o contexto multivariado.

2.5 Agradecimentos

Esse estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Bergamo, G. C., Dias, C. T. d. S., e Krzanowski, W. J. (2008). Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, 65(4):422–427.

Bonat, W. H. e Ribeiro Jr, P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, 27(2):83–89.

Brown, P. E. et al. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software*, 63(12):1–24.

Christensen, O. F. (2004). Monte carlo maximum likelihood in model-based geostatistics. *Journal of computational and graphical statistics*, 13(3):702–718.

Christensen, O. F. e Ribeiro Jr, P. J. (2002). georglm-a package for generalised linear spatial models. *R News*, 2(2):26–28.

Christensen, O. F. e Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58(2):280–286.

Cranmer, S., Gill, J., Jackson, N., Murr, A., e Armstrong, D. (2016). *hot.deck: Multiple Hot-Deck Imputation*. R package version 1.1.

Cranmer, S. J. e Gill, J. (2013). We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43(2):425–449.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Curran, P. J. (1988). The semivariogram in remote sensing: an introduction. *Remote sensing of Environment*, 24(3):493–507.

Dai, G. e Müller, U. U. (2019). Efficient estimators for expectations in nonlinear parametric regression models with responses missing at random. *Electron. J. Statist.*, 13(2):3985–4014.

Diggle, P. e Ribeiro, P. (2007). *Model-based Statistics*. Springer series in statistics.

Diggle, P. J., Ribeiro, P. J., e Christensen, O. F. (2003). An introduction to model-based geostatistics. In *Spatial statistics and computational methods*, pages 43–86. Springer.

Diggle, P. J., Tawn, J. A., e Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Lang, K. M. e Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, 19(3):284–294.

Lindsey, J. K. (2000). *Applying generalized linear models*. Springer Science & Business Media.

Liu, L., Qiu, Y., Natarajan, L., Messer, K., et al. (2019). Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines. *The Annals of Applied Statistics*, 13(3):1370–1396.

Naegele, R., Granke, L., Fry, J., Hill, T., Ashrafi, H., Van Deynze, A., e Hausbeck, M. (2017). Disease resistance to multiple fungal and oomycete pathogens evaluated using a recombinant inbred line population in pepper. *Phytopathology*, 107(12):1522–1531.

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

Silveira, E. M. d. O., Mello, J. M. d., Acerbi Júnior, F. W., Reis, A. A. d., Withey, K. D., e Ruiz, L. A. (2017). Characterizing landscape spatial heterogeneity using semivariogram parameters derived from ndvi images. *Cerne*, 23(4):413–422.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

Van Buuren, S. e Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Van Buuren, S. e Oudshoorn, C. (2000). Multivariate imputation by chained equations: Mice v1. 0 users's manual: Tno prevention and health. *Public Health*.

Zhang, S., Han, P., Wu, C., et al. (2019). Empirical likelihood inference for non-randomized pretest-posttest studies with missing data. *Electronic Journal of Statistics*, 13(1):2012–2042.

3 ANÁLISE ESPAÇO-TEMPORAL DA DISTRIBUIÇÃO DE PRAGAS E PREDADORES NA CULTURA DO MILHO

Resumo

Este trabalho propõe uma modelagem geoestatística espaço-temporal para confrontar os dados de moscas sirfídeos adultas e colônias de pulgões na cultura do milho. A utilização de um modelo geoestatístico que admite a variação espaço-tempo torna a abordagem mais interessante por ser um modelo mais completo. Foi utilizada a regressão múltipla para modelar a componente de tendência para variável resposta contagem de moscas sirfídeos adultas e colônia de pulgões, com as coordenadas servindo de covariáveis e as variações espaço temporais ao redor do desvio são descritas por um campo aleatório residual espaço-tempo. Por fim, o mapa de predição obtido pela krigagem pode ser um indicador biológico para possíveis colônias de pulgões na cultura do milho.

Palavras-chave: Moscas sirfídeos; pulgão-do-milho; variabilidade espaço-temporal; indicadores biológicos.

3.1 Introdução

As espécies invasoras estão entre os impactos mais importantes na humanidade, menos controlados e menos reversíveis nos ecossistemas do mundo, com consequências negativas, afetando sua sustentabilidade, biodiversidade e sistemas socioeconômicos (Dutra Silva et al., 2017). Na cultura do milho, por exemplo, o ataque às espigas e colmos pode gerar perdas no rendimento de grãos na ordem de 15 a 34% (Lima et al., 2010). Assim, conhecer como pragas e seus inimigos naturais se distribuem, no espaço e no tempo, é relevante para aplicações de manejo integrado de pragas.

O combate a essas pragas geralmente é feito com o uso de agroquímicos, porém algumas pragas como o pulgão-do-milho (*Rhopalosiphum maidis*, Fitch., 1856) apresentam inimigos naturais. Bortolotto et al. (2016) relatam que a ocorrência de populações de moscas sirfídeos acompanha à de colônias de pulgões, aumentando ou diminuindo proporcionalmente a sua incidência. Isso se dá porque as moscas da família Syrphidae costumam depositar seus ovos próximos das colônias de pulgões, já que se alimentam de afídeos na fase larval, enquanto que na fase adulta de pólen e néctar. Portanto, a presença de moscas sirfídeos adultas na lavoura poderá ser um provável indicador biológico da presença de colônia de pulgões. Park e Obrycki (2004) cita que uma maneira de investigar a sincronia espaço-temporal das distribuições de predadores e presas é gerar e comparar mapas de distribuição em sequências temporais.

Ferramentas que possam descrever o comportamento dos predadores são de muita valia para indicação e monitoramento de infestações de pragas. Dentre essas ferramentas estão a amostragem e a análise espacial. Trabalhos de amostragem de pragas pode ser visto em Farias et al. (2001b). A utilização de modelos espaciais já é comumente usada na literatura como pode ser visto em Farias et al. (2001a), Mehrjardi et al. (2008), Duarte et al. (2015), Garcia et al. (2017) entre outros. Uma abordagem mais completa e mais complexa é a análise espaço temporal, que utiliza modelos robustos para descrever e prever a incidência de pragas, estudando conjuntamente a dinâmica espaço-tempo.

O espaço/tempo é visto como um arranjo espacial contínuo combinado com uma ordem temporal dos eventos. Essa união de espaço e tempo é definida em termos de seu produto cartesiano (Christakos e Vyas, 1998). Essa extensão dos modelos espaciais é uma evolução lógica dos modelos estatísticos para o mapeamento de dados espacial e temporalmente correlacionados, além de facilitar a distinção de componentes de variabilidade puramente temporais, puramente espaciais ou espaço-temporais (Kilibarda et al., 2014).

Este trabalho tem por objetivo modelar a variabilidade espaço-temporal de colônias de pulgões-do-milho e moscas da família Syrphidae na cultura do milho para investigar e monitorar a dinâmica de

infestação espaço-temporal de presa e predador em diferentes estádios fenológicos, servindo como um possível indicador biológico no campo, a partir de mapas temporais.

3.2 Materiais e métodos

3.2.1 Área de estudo e amostragem

Esse estudo foi realizado durante os anos de 2015 e 2016 na Fazenda Experimental de Igarapé-açu - FEIGA no município de Igarapé-açu, estado do Pará, Brasil (01º07'33''S; 47º37'27''W; altitude 39 m). A área era de 10.000 m^2 e foi dividida em 100 parcelas de 10 $\times$ 10 m^2, cada. A cultivar de milho utilizada foi híbrida. A semeadura foi realizada em 30/03/2015 com espaçamento entre linhas de 0,90 m e 0,15 m por planta e sem intervenção química. A amostragem do pulgão-do-milho foi realizada de 28/05/2016 à 25/06/2016 e das moscas sirfídeos foi realizada de 06/05/2016 à 11/06/2016, ambas em intervalos semanais. Em cada parcela, 10 plantas foram escolhidas aleatoriamente, totalizando 1.000 plantas por amostragem. A ocorrência de inimigos naturais foi avaliada visualmente e o número de indivíduos, por planta, foi determinado e, para a ocorrência de colônias de pulgões-do-milho, todas as partes aéreas da planta foram analisadas visualmente. A colônia foi considerada quando mais de 15 indivíduos foram encontrados em uma determinada parte da planta.

3.2.2 Modelos espaço-temporais

Seja um processo espaço-temporal gaussiano Z definido sobre um domínio espacial S e domínio temporal T, $\{Z(s,t) : (s,t) \in (S \times T\}$, em que $S \subseteq \Re^d$ e $T \subseteq \Re$, um processo espaço-temporal estatístico, observações são modeladas como uma realização parcial de uma função aleatória espaço-temporal. (Xu e Shu, 2015). A variação espaço-temporal de Z pode ser decomposta pelos componentes da tendência $m(s,t)$ e de um resíduo estocástico $\varepsilon(s,t)$ (Yang et al., 2015), ou seja

$$Z(s,t) = m(s,t) + \varepsilon(s,t). \tag{3.1}$$

Na equação 3.1 assumiu-se que Z possui os momentos de primeira a segunda ordem. O componente de média $m(s,t) = E[Z(s,t)]$ que é a esperança da variável Z. O resíduo estocástico incorpora os três componentes: espacial, temporal e da interação (De Iaco et al., 2015) e é dado por

$$m(s,t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \tag{3.2}$$

em que x_1, x_2, x_3 e x_4 são as variáveis preditoras do modelo de regressão e β_0, β_1, β_2, β_3 e β_4 são os parâmetros a serem estimados.

Para a modelagem foi declarado que esses componentes são estacionários de segunda ordem, mutuamente independentes e que é espacialmente isotrópico. Como função de ligação, temos $g[m(s,t)] = log[m(s,t)]$. Devido ao alto índice de zeros na variável resposta, o componente de tendência foi modelado por uma distribuição Poisson Inflacionada de Zeros (ZIP). Foram consideradas como covariáveis as próprias coordenadas geográficas (latitude e longitude) e um índice temporal linear e quadrático. O modelo de tendência ajustado é dado por

$$\hat{m}(s,t) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4, \tag{3.3}$$

em que, as variáveis x_1, x_2 representam a latitude e longitude, respectivamente. As variáveis x_3 e x_4 representam o efeito temporal linear e quadrático, respectivamente. Após a modelagem da tendência, o próximo passo consistiu na modelagem do variograma espaço-temporal sobre os resíduos (Martínez et al., 2017). Essa variação espaço-temporal não explicada pelas covariáveis é então modelada pelo componente

residual $\varepsilon(s,t) = Z(s,t) - m(s,t)$ usando o variograma espaço-temporal ($\gamma_{s,t}$) dado por,

$$\gamma_{s,t}(h_s, h_t) = \frac{1}{2}E[\varepsilon(s,t) - \varepsilon(s + h_s, t + h_t)]^2, \tag{3.4}$$

sendo $h_s = s - s^{'}$ e $h_t = t - t^{'}$ para quaisquer (s,t) e $(s^{'}, t^{'})$ no domínio espaço-temporal. O próximo passo da análise consiste em determinar um modelo teórico, baseado nas funções de covariâncias, que se ajuste aos dados do variograma amostral espaço-temporal. Para o ajuste de um modelo teórico ao variograma empírico apresentado em 3.4 foi utilizado o modelo produto-soma generalizado (Pebesma e Heuvelink, 2016). O variograma, metade da diferença da variância, é geralmente mais útil do que a função de covariância por causa de suas suposições mais fracas (Xu e Shu, 2015). Assim, temos a relação de covariância e semivariograma em processos estacionários de segunda ordem (média constante e função de covariância estacionária) dado por

$$\gamma_{st}(h_s, h_t) = C_{st}(0,0) - C_{st}(h_s, h_t),$$

em que $C_{st}(0,0)$ e $C_{st}(h_s, h_t)$ são o valor do sill global e a função de covariância estacionária, respectivamente. De Cesare et al. (2001) define o variograma produto soma generalizado como

$$\gamma_{st}(h_s, h_t) = \gamma_{st}(h_s, 0) + \gamma_{st}(0, h_t) - k\gamma_{st}(h_s, 0)\gamma_{st}(0, h_t), \tag{3.5}$$

em que $\gamma_{st}(h_s, 0)$ e $\gamma_{st}(0, h_t)$ são variogramas marginais espaciais e temporais válidos, respectivamente. O parâmetro k, na equação 3.5 está relacionado com os patamares espaciais e temporais.

3.2.3 Krigagem espaço-temporal

Para Heuvelink et al. (2012) as fórmulas de krigagem no domínio espaço-temporal não diferem fundamentalmente, em sentido matemático ou estatístico, daquelas da krigagem espacial. De acordo com Kilibarda et al. (2014) o modelo de variograma é crucial na krigagem espaço-temporal para calcular o melhor preditor linear não-viesado, que é dado por

$$\hat{\varepsilon}(s_0, t_0) = c_0^T C_n^{-1} \overline{\varepsilon}, \tag{3.6}$$

em que $\hat{\varepsilon}(s_0, t_0)$ é o preditor linear, obtido pela krigagem, para a localização (s_0, t_0), C_n é a matriz de covariâncias de ordem $n \times n$ dos resíduos para os n pontos observados no espaço-tempo, derivada do variograma espaço-temporal, c_0 é um vetor de covariâncias dos resíduos para os pontos observados e preditos e ε é um vetor dos resíduos nos n pontos observados. Kilibarda et al. (2014) definiram a variância do erro de predição na krigagem espaço-temporal como

$$\begin{aligned} \sigma^2(s_0, t_0) &= Var[\varepsilon(s_0, t_0) - \hat{\varepsilon}(s_0, t_0)] \\ &= C(0,0) - c_0^T C_n^{-1} c_0 + (m_0 - M^T C_n^{-1} c_0)^T (M^T C_n^{-1} M)^{-1} (m_0 - M^T C_n^{-1} c_0). \end{aligned} \tag{3.7}$$

A Equação 3.7 é semelhante à krigagem espacial uma vez que a formulação matemática é a mesma (Heuvelink et al. (2012), Kilibarda et al. (2014)). Tem-se que, $C(0,0)$ é a variância para Z, M é a matriz do delineamento de ordem $n \times p$ das variáveis preditoras nos pontos observados e m_0 é o vetor de preditores no ponto de predição. Kilibarda et al. (2014) define o preditor final $\hat{z}(s_0, t_0)$ para a variável Z, na localização (s_0, t_0), da forma

$$\hat{z}(s_0, t_0) = \hat{m}(s_0, t_0) + \hat{\varepsilon}(s_0, t_0), \tag{3.8}$$

em que $\hat{m}(s_0, t_0)$ é o valor estimado para a localização (s_0, t_0) obtido por meio da Equação 3.3 e $\hat{\varepsilon}(s_0, t_0)$ como definido na Equação 3.6. Para determinar o melhor modelo ajustado ao variograma amostral teórico, a medida RMSE (Root Mean Square Error) foi calculada, obtido pela diferença nas nuvens de pontos entre o variograma amostral e o variograma teórico.

3.3 Resultados

Um aumento gradativo da quantidade de colônias de pulgões pode ser observado com o decorrer das semanas, tendo seu ápice na última amostragem (Figura 3.1 (a)). Na Figura 3.1 (b), mapas de moscas sirfídeos adultas apresentam um aumento da quantidade de moscas com o decorrer do tempo, o que é um possível indício do aumento da população de pulgões. Observa-se também que a amostra de 04/06/2016 possui baixa quantidade de moscas. Isso se deve à ocorrência de chuva forte no dia da amostragem, o que pode ter causado interferência.

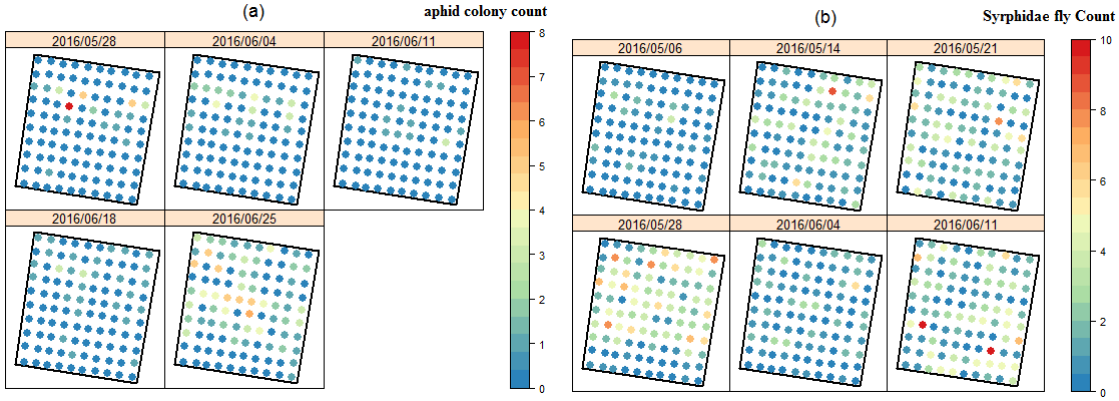

Figura 3.1. Distribuição espaço-temporal: (a) colônia de pulgões e (b) contagem de moscas sirfídeos adultas na área experimental localizada na região nordeste paraense por amostragem semanal

A Figura 3.2 apresenta boxplots como gráficos exploratórios da (a) contagem de moscas sirfídeos adultas e (b) para colônia (> 15 indivíduos) de pulgões, por intervalos semanais de amostragem em um hectare localizado no nordeste Paraense. É possível verificar a grande existência de zeros em ambas, além de uma possível variabilidade alta em cada amostragem. Isto fornece a ideia que para modelar a tendência, deve-se utilizar modelos que possam comportar tais fenômenos, como a ZIP.

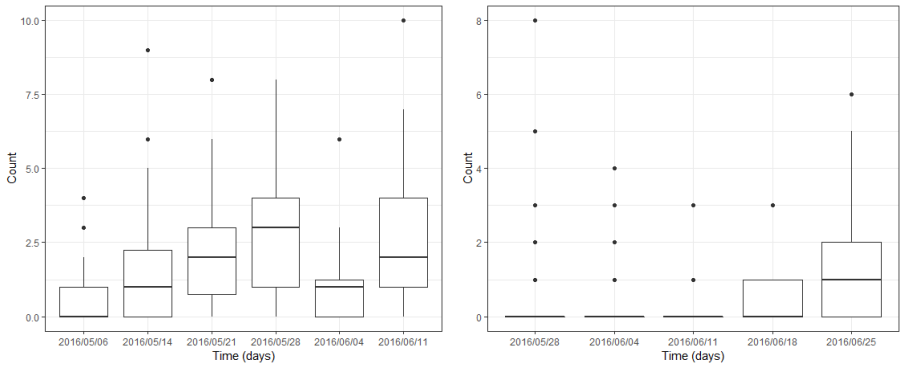

Figura 3.2. Boxplot para (a) contagem de moscas sirfídeos adultas e (b) para colônias de pulgões por intervalos semanais de amostragem em um hectare localizado no nordeste Paraense

Para modelar a tendência dos dados foi utilizado o modelo de regressão como indicado na equação 3.3. A Tabela 3.1 mostra as estimativas dos parâmetros, erro padrão, estatística t e o valor p para colônia de pulgões e moscas sirfídeos adultas. É possível verificar que, ao nível de 0.05 de significância, as covariáveis latitude e tempo estão associadas à contagem de colônias de pulgões e latitude, longitude e o efeito temporal linear e quadrático estão associadas à contagem de moscas sirfídeos adultas. Os resíduos produzidos por este modelo serão utilizados para a análise da depedência espacial.

Tabela 3.1. Estimativa dos parâmetros do modelo descrito em 3.3 para contagens de moscas sirfídeos adultas e contagem de colônias de pulgões por amostragem semanal em um hectare localizado no nordeste Paraense.

Sirfídeos				
Variável	Estimação	Erro Padrão	valor de t	valor de p
Intercepto	-39126.13	9337	-4.190	< 0.01
Latitude	0.0039	0.0009	4.131	< 0.01
Longitude	0.0027	0.0009	2.905	0.0038
Tempo	0.9575	0.1045	9.165	< 0.01
Tempo^2	-0.1079	0.0134	-8.274	< 0.01
Pulgões				
Intercepto	-135800	29600	-4.588	< 0.01
Latitude	0.0138	0.0030	4.588	< 0.01
Tempo	0.3227	0.0447	7.22	< 0.01

Após removida a tendência espaço-temporal da contagem de moscas adultas de sirfídeos e de contagem de colônias de pulgões, é calculado o variograma sobre os resíduos. Na Figura 3.3 mostra-se o modelo de variograma espaço-temporal empírico (a) e (c) e o modelo produto-soma ajustado (b) e (d). Como pode ser observado, os resíduos apresentam uma clara correlação tanto no espaço quanto no tempo e a variação total desses resíduos é explicada por componentes espaciais e também temporais. Nota-se que a estrutura espacial torna-se mais fraca à medida que as diferenças de tempo aumentam e a estrutura temporal torna-se mais fraca à medida em que se aumentam as diferenças espaciais. O variograma amostral (a) e (c) podem ser visualizados em termos de seus variogramas marginais, um para o espaço e outro para o tempo. A tendência crescente na dimensão espacial e temporal no variograma amostral indica a presença de uma forte correlação espaço-temporal em ambos.

Na Tabela 3.2 são apresentadas as estimativas dos parâmetros do modelo produto-soma generalizado considerando as diferentes estruturas para os componentes espacial e temporal de contagem de moscas sirfídeos adultas e contagem de colônias de pulgões. Com o intuito de comparação dos modelos foi utilizada a RMSE. Trabalhos como de Hu et al. (2015), De Iaco et al. (2015), Kilibarda et al. (2014) e Jovein e Hosseini (2017) utilizaram apenas uma única estrutura de correlação espaço-temporal.

Neste trabalho, foi proposto uma análise mais abrangente utilizando diferentes estruturas na formação do modelo. Para o modelo de variograma produto-soma generalizado, foi considerado o modelo Exponencial para o componente espacial e o modelo Linear para o componente temporal para sirfídeos e Gaussiano e linear para pulgões, respectivamente. A amplitude espacial de sirfídeos foi de 46 metros, sugerindo que a correlação espacial se tornou insignificante após este limiar. Por outro lado, o intervalo temporal de 23 dias indicou que a correlação temporal tornou-se insignificante após este intervalo. Já para a colônia de pulgões temos que a amplitude espacial foi de 45 metros e o intervalo temporal foi de 20 dias.

Tabela 3.2. Estimativa dos parâmetros do modelo produto soma generalizado para contagem de moscas sírfídeos adultas e contagem de colônias de pulgões

Sirfídeos						
Modelo		Sill	Range	Nugget	k	RMSE
Espaço	Exponencial	0.06921283	45.65721 m	0.05899567	118.540903	0.05062
Tempo	Linear	0.1250645	23 dias	0.1250645		
Pulgão						
Espaço	Gaussiano	0.05545781	44.75414 m	0.05545781	116.67090137949	0.03871882
Tempo	Linear	0.241602	20 dias	0.12080100		

É possível verificar, na Figura 3.4 (a), um aumento das colônias de pulgões semanalmente. Nas amostragens temos o surgimento do pulgão no estádio de pendoamento do milho até o último dia de amostragem com o estádio de formação dos dentes. Já em 3.4 (b) é possível observar que existe

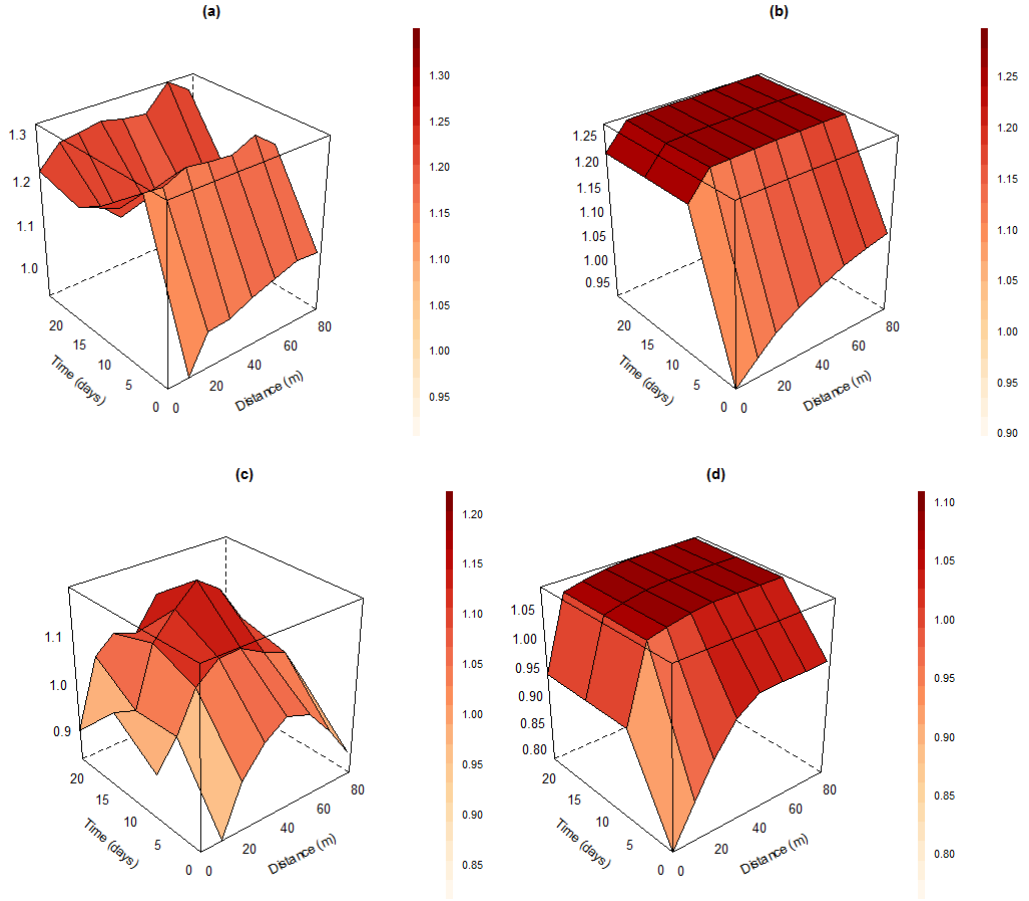

Figura 3.3. Variograma espaço-temporal: (a) amostral para contagem de moscas sirfídeos adultas; (b) modelo produto soma generalizado para contagem de moscas sirfideos adultas obtido a partir dos resíduos da regressão linear múltipla; (c) amostral para contagem de colônias de pulgões e (d) modelo produto soma generalizado para contagem de colônias de pulgões obtido a partir dos resíduos da regressão linear múltipla

uma tendência de infestação que começa do oeste sentido leste. Na primeira amostragem (06/05/2016) há quantidades bem pequenas de moscas. A partir da segunda amostragem (14/05/2016) já é nítido o aumento de moscas na plantação. Isso possivelmente acontece por conta da oferta de alimentos que deve estar crescendo (colônia de pulgões). A amostragem do dia 28/05/2016 apresenta alto índice de infestação. Esta fase é a fase de pendoamento do milho e segundo Weismann (2008) é neste estádio que a planta está mais sucetível a ataques devido à maior exposição do pendão e de todas as folhas. Nas duas últimas amostragens já é possível verificar uma pequena diminuição da infestação de moscas. A incerteza é representada nos mapas Figura 3.5 (a) e (b) e evidencia maiores estimativas de erros para as bordas.

3.4 Discussão

Infestações naturais de pulgões pode produzir uma gama de densidades populacionais por meio de diferentes estágios. Al-Eryan e El-Tabbakh (2004) relatam esse fato nos estágios de crescimento V10, pendoamento (VT) e amadurecimento. Relatam ainda uma perda de rendimento de 28.14% nos estádios de crescimento V10 - VT, enquanto a infestação pelos estádios de crescimento de maturação R2 a R4 causou perda de rendimento de 16,28%. Cruz (2004) comentam que a infestação inicia-se em plantas isoladas, principalmente no perído vegetativo, próximo ao estádio VT. Este estudo mostrou o surgimento do pulgão no estádio VT, confirmando a hipótese do autor e maior ocorrência no estádio reprodutivo de

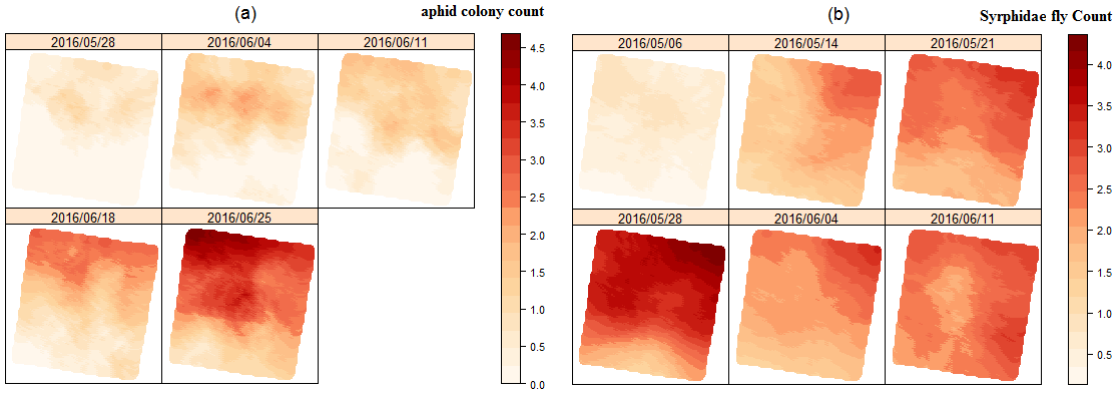

Figura 3.4. Mapas de prediçao: (a) contagem de colônias de pulgões e (b) contagem de moscas sirfideos adultas por amostragem semanal em um hectare localizado no nordeste paraense

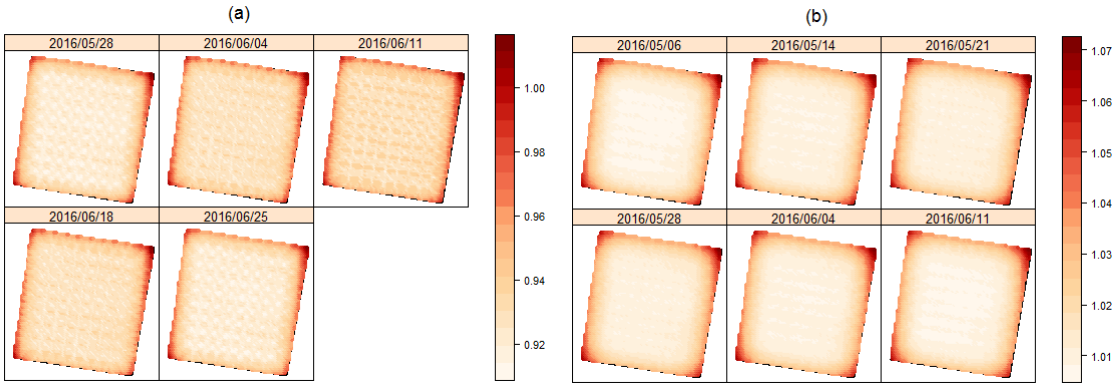

Figura 3.5. Mapas de erros padrão da predição: (a) contagem de colônias de pulgões e (b) contagem de moscas sírfídeos adultas por amostragem semanal em um hectare localizado no nordeste paraense

formação dos dentes (R5).

Estudos como Yang et al. (2019) mostraram que habitats não cultivados, como lotes de floresta e vegetação em torno de habitações ou áreas úmidas na paisagem, todos tiveram correlações positivas com a abundância de pulgões nos campos de trigo porque eles forneceram fontes para a colonização de pulgões. Ao lado do experimento existia uma área de mangueira, pastagem e fragmento florestal. Estas mostraram ter influência sobre a colonização de pulgões e moscas sirfídeos adultas também no milho, como pôde ser visto nos mapas.

Pesquisas como Klingauf (1987) e Damicone et al. (2007) relatam que o vento pode influenciar a distribuição de outros tipos de pulgões com asas. O surgimento de pulgões se deu próximo a área de pastagem, sendo o vento um possível fator para seu deslocamento. Conforme colônias de pulgões começaram a aparecer e a aumentar, houve um expressivo aumento da quantidade de moscas adultas na área (28/05/2016) que saíram das áreas adjacentes de fragmento florestal e adentraram na cultura do milho.

Provavelmente, fragmento florestal servia de abrigo às moscas e forneciam alimentação, pois na fase adulta as moscas se alimentam de pólen e néctar (Müller (1883); Holloway (1976)). Esses habitats semi-naturais favorecem as populações de inimigos naturais e melhoram sua eficiência como agentes de controle (Alignier et al., 2014).

Moscas adultas depositam seus ovos próximos às colônias de pulgões (White et al., 1995) para as larvas se alimentarem dos mesmos (Belliure e Michaud, 2001), então, a observação de moscas adultas

é um possível indício de ocorrência de colônias. Apesar de não termos amostragens de moscas nas datas finais, podemos inferir que a diminuição das moscas propiciou o aumento das colônias ou que a ocorrência somente deste predador não foi suficiente para diminuir a população de pulgões. Uma provável barreira natural para controle de colônias é implantar experimentos com áreas adjacentes à flores para contribuir com a rápida colonização de moscas adultas na presença de pulgões.

Muitos estudos como Medeiros et al. (2019), Gasch et al. (2015) entre outros, mostram que inserir mais covariáveis pode melhorar a acurácia do modelo. Talvez a inserção de uma ou mais covariáveis pudesse melhorar a predição no nosso estudo e melhor explicar a influência da relação praga versus predador. Algumas vantagens na utilização dessa abordagem é que a mesma propicia uma flexibilidade para estimação do componente de tendência, podendo considerar tanto variáveis contínuas, como categóricas, além de aplicar diferentes funções de ligação no preditor, não sendo necessário uma transformação Box-Cox como em Hussain et al. (2010), além de modelar conjuntamente espaço e tempo, nem sempre utilizado em pesquisas aplicadas. Alguns exemplos de análises marginais desses dois componentes podem ser verificadas nos trabalhos de Rojo e Pérez-Badia (2015), Pelissari et al. (2017) e Sciarretta et al. (2018). Um bom ajuste pode fornecer erros Gaussianos que podem ser facilmente interpolados mantendo suposições de krigagem e ferramentas de diagnóstico disponíveis para avaliar o ajuste do modelo (Poggio et al., 2012). Porém a utilização dessa abordagem oferece como desvantagem o fato de que para modelar a tendência temos que supor independência e os dados possuem conhecida dependência espaço-temporal.

3.5 Conclusão

Uma abordagem atual e flexível para estimar a distribuição espaço-temporal de pragas e inimigos naturais com um modelo que admite a utilização da distrbuição Poisson Inflacionada de Zeros e uma amostragem temporal semanal foi apresentada neste artigo. A relação de pragas e predadores pode ser um indicador biológico no controle em culturas para amenizar o uso de agroquímicos nas lavouras de produção. Habitats semi-naturais podem favorecer as populações de inimigos naturais. Foi possível verificar que a ocorrência da praga propiciou um aumento significativo de predadores adultos. Por fim, pudemos concluir que barreiras naturais podem ser utilizadas para controle de colônias de pulgões, favorecendo seus inimigos naturais.

3.6 Agradecimentos

Esse estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por financiar em parte a pesquisa.

Referências

Al-Eryan, M. e El-Tabbakh, S. S. (2004). Forecasting yield of corn, zea mays infested with corn leaf aphid, rhopalosiphum maidis. *Journal of applied entomology*, 128(4):312–315.

Alignier, A., Raymond, L., Deconchat, M., Menozzi, P., Monteil, C., Sarthou, J.-P., Vialatte, A., e Ouin, A. (2014). The effect of semi-natural habitats on aphids and their natural enemies across spatial and temporal scales. *Biological control*, 77:76–82.

Belliure, B. e Michaud, J. (2001). Biology and behavior of pseudodorus clavatus (diptera: Syrphidae), an important predator of citrus aphids. *Annals of the Entomological Society of America*, 94(1):91–96.

Bortolotto, O. C., de Oliveira Menezes Jr, A., e Hoshino, A. T. (2016). Abundância de inimigos naturais de pulgões do trigo em diferentes distâncias da borda da mata. *Pesquisa Agropecuária Brasileira*, 51(2):187–191.

Christakos, G. e Vyas, V. M. (1998). A novel method for studying population health impacts of spatio-temporal ozone distribution. *Social Science & Medicine*, 47(8):1051–1066.

Cruz, I. (2004). Manejo de pragas da cultura do milho. In *Tecnologias de produção de milho*, pages 311–366. UFV.

Damicone, J., Edelson, J., Sherwood, J., Myers, L., e Motes, J. (2007). Effects of border crops and intercrops on control of cucurbit virus diseases. *Plant disease*, 91(5):509–516.

De Cesare, L., Myers, D., e Posa, D. (2001). Estimating and modeling space–time correlation structures. *Statistics & Probability Letters*, 51(1):9–14.

De Iaco, S., Palma, M., e Posa, D. (2015). Spatio-temporal geostatistical modeling for french fertility predictions. *Spatial Statistics*, 14:546–562.

Duarte, F., Calvo, M., Borges, A., e Scatoni, I. (2015). Geostatistics and geographic information systems to study the spatial distribution of grapholita molesta (busck)(lepidoptera: Tortricidae) in peach fields. *Neotropical entomology*, 44(4):319–327.

Dutra Silva, L., Brito de Azevedo, E., Bento Elias, R., e Silva, L. (2017). Species distribution modeling: Comparison of fixed and mixed effects models using inla. *ISPRS International Journal of Geo-Information*, 6(12):391.

Farias, P. R., Barbosa, J. C., e Busoli, A. C. (2001a). Distribuição espacial da lagarta-do-cartucho, spodoptera frugiperda (je smith)(lepidoptera: Noctuidae), na cultura do milho. *Neotropical Entomology*, pages 681–689.

Farias, P. R. S., Barbosa, J. C., e Busoli, A. C. (2001b). Amostragem seqüencial com base na lei de taylor para levantamento de spodoptera frugiperda na cultura do milho. *Scientia Agricola*, 58(2):395–399.

Garcia, A., Araujo, M., Uramoto, K., Walder, J., e Zucchi, R. (2017). Geostatistics and geographic information system to analyze the spatial distribution of the diversity of anastrepha species (diptera: Tephritidae): the effect of forest fragments in an urban area. *Environmental entomology*, 46(6):1189–1194.

Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., e Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3d+ t: The cook agronomy farm data set. *Spatial Statistics*, 14:70–90.

Heuvelink, G. B., Griffith, D. A., Hengl, T., e Melles, S. J. (2012). Sampling design optimization for space-time kriging. In *Spatio-Temporal Design*, pages 207–230. Wiley Online Library.

Holloway, B. A. (1976). Pollen-feeding in hover-flies (diptera: Syrphidae). *New Zealand journal of zoology*, 3(4):339–350.

Hu, Y., Li, R., Bergquist, R., Lynn, H., Gao, F., Wang, Q., Zhang, S., Sun, L., Zhang, Z., e Jiang, Q. (2015). Spatio-temporal transmission and environmental determinants of schistosomiasis japonica in anhui province, china. *PLoS neglected tropical diseases*, 9(2):e0003470.

Hussain, I., Spöck, G., Pilz, J., e Yu, H.-L. (2010). Spatio-temporal interpolation of precipitation during monsoon periods in pakistan. *Advances in water resources*, 33(8):880–886.

Jovein, E. B. e Hosseini, S. M. (2017). Predicting saltwater intrusion into aquifers in vicinity of deserts using spatio-temporal kriging. *Environmental monitoring and assessment*, 189(2):81.

Kilibarda, M., Hengl, T., Heuvelink, G. B., Gräler, B., Pebesma, E., Perčec Tadić, M., e Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, 119(5):2294–2313.

Klingauf, F. (1987). Host plant finding and acceptance. In *Aphids: their biology, natural enemies and control*, pages 209–223. Elsevier: Amsterdam.

Lima, M., Silva, P., Oliveira, O., Silva, K., e Freitas, F. (2010). Corn yield response to weed and fall armyworm controls. *Planta Daninha*, 28(1):103–111.

Martínez, W. A., Melo, C. E., e Melo, O. O. (2017). Median polish kriging for space–time analysis of precipitation. *Spatial Statistics*, 19:1–20.

Medeiros, E. S. d., Lima, R. R. d., Olinda, R. A. d., e Santos, C. A. C. d. (2019). Modeling spatiotemporal rainfall variability in paraíba, brazil. *Water*, 11(9):1843.

Mehrjardi, R. T., Jahromi, M. Z., e Heidari, A. (2008). Spatial distribution of groundwater quality with geostatistics (case study: Yazd-ardakan plain) 1.

Müller, H. (1883). Diptera and thysanoptera. *The fertilization of flowers. Macmillan, London*, pages 36–45.

Park, Y.-L. e Obrycki, J. J. (2004). Spatio-temporal distribution of corn leaf aphids (homoptera: Aphididae) and lady beetles (coleoptera: Coccinellidae) in iowa cornfields. *Biological control*, 31(2):210–217.

Pebesma, E. e Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *RFID Journal*, 8(1):204–218.

Pelissari, A. L., Figueiredo Filho, A., Netto, S. P., Ebling, A. A., Roveda, M., e Sanquetta, C. R. (2017). Geostatistical modeling applied to spatiotemporal dynamics of successional tree species groups in a natural mixed tropical forest. *Ecological indicators*, 78:1–7.

Poggio, L., Gimona, A., e Brown, I. (2012). Spatio-temporal modis evi gap filling under cloud cover: An example in scotland. *ISPRS journal of photogrammetry and remote sensing*, 72:56–72.

Rojo, J. e Pérez-Badia, R. (2015). Spatiotemporal analysis of olive flowering using geostatistical techniques. *Science of the total Environment*, 505:860–869.

Sciarretta, A., Tabilio, M. R., Lampazzi, E., Ceccaroli, C., Colacci, M., e Trematerra, P. (2018). Analysis of the mediterranean fruit fly [ceratitis capitata (wiedemann)] spatio-temporal distribution in relation to sex and female mating status for precision ipm. *PloS one*, 13(4):e0195097.

Weismann, M. (2008). Fases de desenvolvimento da cultura do milho. *Tecnologias e produção: Milho safrinha e culturas de inverno. Maracaju: Fundação MS*, pages 31–38.

White, A. J., Wratten, S. D., Berry, N. A., e Weigmann, U. (1995). Habitat manipulation to enhance biological control of brassica pests by hover flies (diptera: Syrphidae). *Journal of Economic Entomology*, 88(5):1171–1176.

Xu, J. e Shu, H. (2015). Spatio-temporal kriging based on the product-sum model: some computational aspects. *Earth Science Informatics*, 8(3):639–648.

Yang, L., Liu, B., Zhang, Q., Zeng, Y., Pan, Y., Li, M., e Lu, Y. (2019). Landscape structure alters the abundance and species composition of early-season aphid populations in wheat fields. *Agriculture, Ecosystems & Environment*, 269:167–173.

Yang, Y., Wu, J., e Christakos, G. (2015). Prediction of soil heavy metal distribution using spatiotemporal kriging with trend model. *Ecological Indicators*, 56:125–133.

4 CONSIDERAÇÕES FINAIS

Esta tese apresentou a aplicação de modelos generalizados mistos com efeitos espaciais e modelagem espaço-temporal. Imputação e análises espaço temporais não são ferramentas usuais neste tipo de abordagem, portanto a ideia foi mostrar comparações e aplicabilidade.

No primeiro artigo, um estudo de comparação de métodos de imputação simples e múltipla, junto ao modelos SGLMM sem imputação foi realizado para verificar como se comportavam as imputações e se é vantajoso realizá-las em dados espacialmente correlacionados. Um estudo de simulação foi realizado para comparar os diferentes métodos e posteriomente uma aplicação em dados reais foi realizada. Pôde-se constatar que para o estudo dentro da amostra, sem levar em consideração os valores retirados, o método MICE apresenta, em geral, resultados mais próximos do SGLMM. Todos os métodos apresentaram certa dificuldade de predição, principalmente quando se trata de perdas mais expressivas. Vale destacar que o método Hotdeck obteve melhor desempenho em perdas limítrofes e a média obteve o pior desempenho dentre os métodos estudados.

No segundo artigo, uma modelagem espaço temporal foi realizada em dados reais, o que não é comum na área da entomologia. Muitos estudos fixam o tempo e analisam somente modelos espaciais de forma separada, então a proposta foi inserir o componente temporal para extrair as informações de forma mais completa e coerente, já que esses modelos consideram componentes espaciais, temporais e a interação. Pôde-se verificar que habitats semi naturais podem favorecer as populações de inimigos naturais das pragas que atacam a plantação, sendo uma boa alternativa de controle colônias de pulgões.

APÊNDICES

Apêndice I: Material suplementar para o capítulo 2

Esquema representativo a partir do diagrama para a imputação múltipla.

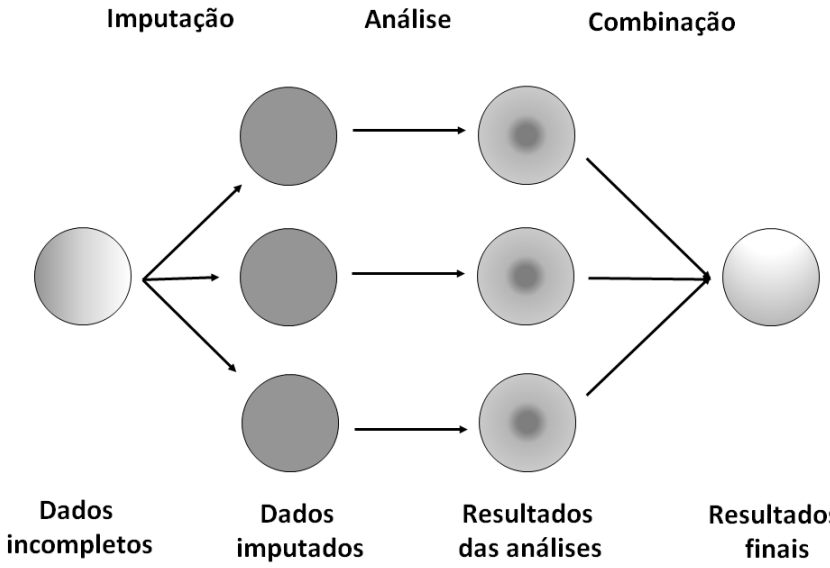

Figura A.1. Diagrama representativo da imputação múltipla

Na seção 2.2.4 do capítulo 2, um estudo de simulação foi apresentado. Nesta seção, é fornecido resultados complementares com valor dos parâmetros $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\phi}$ e $\hat{\tau}^2$.

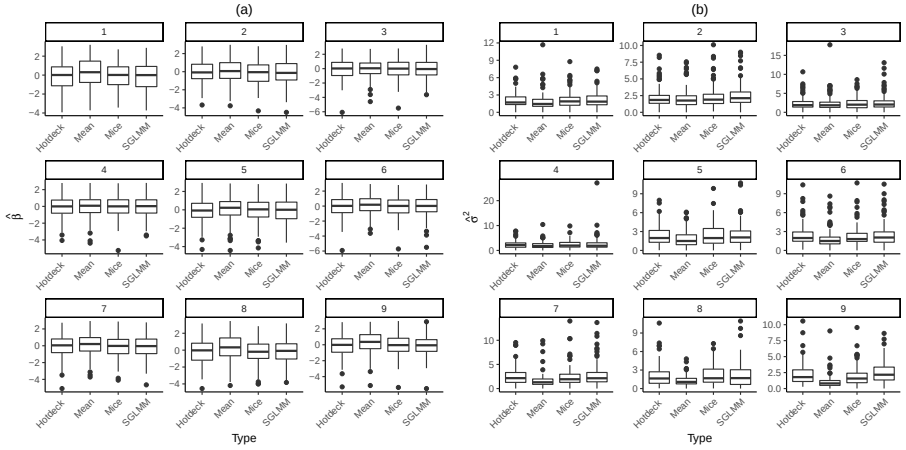

Figura A.2. Valores estimados: (a) intercepto ($\hat{\beta}$) e (b) parâmetro $\hat{\sigma}^2$ para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

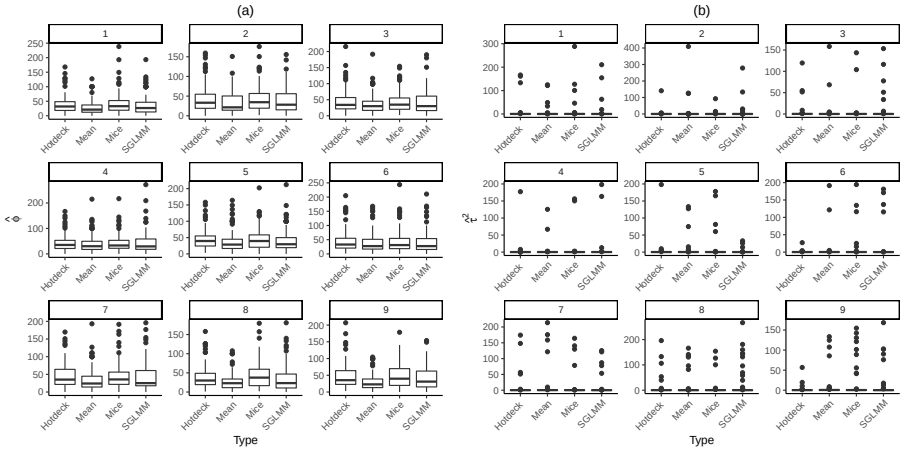

Figura A.3. Valores estimados: (a) parâmetro $\hat{\phi}$ e (b) parâmetro $\hat{\tau}^2$ para os nove cenários simulados e diferentes tipos de imputações e modelo SGLMM sem imputação

ANEXOS

Anexo A: Distribuição Poisson inflacionada de Zeros - ZIP

Seja $Z = 0$ com probabilidade p e $Z \sim Poisson(\mu)$ com probabilidade $(1 - p)$, então Z têm distribuição ZIP dado por

$$f(z) = p + (1 - p)e^{-\mu},$$

se $z = 0$

$$f(z) = (1 - p)\frac{e^{-\mu}\mu^z}{z!},$$

se $z > 0$ for $z = 0, 1, \ldots$

Anexo B: Modelagem espaço-temporal

A abordagem da Geoestatística espaço-temporal visa modelar o componente de tendência, este que admite diversos tipos de distribuições flexíveis como ZIP, Splines (cúbico e penalizado), além da abordagem de modelos GAMLSS. Após obtido os parâmetros do componente de tendência, o resíduo ε é obtido e realiza-se a krigagem.

modelo linear

O modelo linear é composto pela adição de covariâncias espacial e temporal e geralmente é insatisfatória para predições por ser positiva semidefinida e singular em algumas configurações dos dados espaço-temporais[1]. A equação é dada por

$$C_{st}(h_s, h_t) = C_s(h_s) + C_t(h_t),$$

em que C_{st} é a covariância espaço-temporal, C_s a covariância espacial e C_t a covariância temporal.

modelo produto

No modelo produto[2] a covariância espaço temporal pode ser dada por

$$C_{st}(h_s, h_t) = C_s(h_s) \times C_t(h_t),$$

tendo os componentes espaço e tempo separados. Em termos de variograma, o modelo produto pode ser expresso por

$$\gamma_{st}(h_s, h_t) = C_s(0)\gamma_s(h_s) + C_t(0)\gamma_t(h_t) - \gamma_s(h_s)\gamma_t(h_t),$$

em que γ_{st}, γ_s e γ_t são os variogramas espaço temporal, espacial e temporal, respectivamente.

modelo produto soma

O modelo do tipo produto soma é uma extensão dos modelos linear e modelo produto e pode ser obtido por

$$C_{st}(h_s, h_t) = k_1 C_s(h_s)C_t(h_t) + k_2 C_s(h_s) + k_3 C_t(h_t),$$

[1] Myers, D.E., Journel, A.G., 1990. Variograms with Zonal Anisotropies and Non-Invertible Kriging Systems. Math. Geology 22, 779-785

[2] Rodriguez-Iturbe, I., Meija, J.M., 1974. The design of rainfall networks in time and space. Water Resources Res. 10, 713-728

ou em termos de variograma

$$\gamma_{st}(h_s, h_t) = [k_2 + k_1 C_t(0)]\gamma_s(h_s) + [k_3 + k_1 C_s(0)]\gamma_t(h_t) - k_1\gamma_s(h_s)\gamma_t(h_t),$$

em que C_s e C_t são funções de covariâncias. Note que $C_s(0)$, $C_s(0)$ e $C_s(0)$ são o sill dos respectivos variogramas[3] e com as restrições $k_1 > 0$, $k_1 \geq 0$ e $k_3 \geq 0$.

modelo produto soma generalizado

Uma extensão do modelo produto soma, o produto soma generalizado[4] foi proposto para modelar a covariância espaço-temporal. O semivariograma é dado por

$$\gamma_{st}(h_s, h_t) = \gamma_{st}(h_s, 0) + \gamma_{st}(0, h_t) - k\gamma_{st}(h_s, 0)\gamma_{st}(0, h_t),$$

em que $\gamma_{st}(h_s, 0)$ e $\gamma_{st}(0, h_t)$ são os variogramas marginais espacial e temporal, respectivamente. O k pode ser obtido por

$$k = \frac{sill\gamma_{st}(h_s, 0) + \gamma_{st}(0, h_t) - sill\gamma_{st}(h_s, h_t)}{sill\gamma_{st}(h_s, 0)sill\gamma_{st}(0, h_t)}.$$

[3]De Cesare, L., Myers, D., e Posa, D. (2001). Estimating and modeling space–time correlation structures. Statistics & Probability Letters, 51(1):9–14

[4]De Iaco, S., Myers, D., e Posa, D. (2002). Nonseparable space-time covariance models: some parametric families. Mathematical Geology, 34(1):23–42