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

Modelos para análise de dados espaciais binários: uma abordagem Bayesiana

Ana Cristina Alves Garcêz

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

Piracicaba
2021

Ana Cristina Alves Garcêz
Bacharel em Estatística

Modelos para análise de dados espaciais binários: uma abordagem Bayesiana

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2021

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

Garcêz, Ana Cristina Alves
Modelos para análise de dados espaciais binários: uma abordagem Bayesiana / Ana Cristina Alves Garcêz. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011 – – Piracicaba, 2021 .
39 p.

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

1. Estatística bayesiana 2. Dados binários 3. Modelos hierárquicos 4. Análise espacial 5. Imputação múltipla . I. Título.

DEDICATÓRIA

Aos meus pais, irmãs e sobrinhos.

AGRADECIMENTOS

À Deus todo poderoso, que me concedeu a força e a inspiração necessária para chegar ao final deste trabalho;

Aos meus pais, Jurandi Garcêz e Letícia Garcêz, e a minha avó Leopoldina Lima, pelo carinho, amor, paciência e confiança depositada à minha pessoa. Fatores estes, que foram fundamentais para o cumprimento de mais essa etapa em minha vida;

Às minhas amadas irmãs Cinthia Garcêz e Sílvia Garcêz, pelo companheirismo, apoio e incentivo que sempre me deram;

Áos meus sobrinhos Bettina Ferro, Lorenzo Ferro e Guilherme Watanabe, razões da minha felicidade diária;

Ao meu orientador, Prof. Dr. Carlos Tadeu dos Santos Dias, pelo apoio, pela paciência, motivação e por toda confiança;

Ao Prof. Dr. Danilo Sarti, pela atenção e incentivo no desenvolvimento deste trabalho;

Ao Prof. Dr. Daniel Andrade da Faculdade de Ciências Agrárias e Veterinárias da Universidade Paulista Júlio de Mesquita Filho, campus Jaboticabal-SP, pelo conjunto de dados fornecidos para a realização deste trabalho;

Ao meu amigo, Prof. Dr. Djair Durand por toda atenção, paciência, ensinamentos prestados e grande contribuição a este trabalho;

Aos meus amigos Marcello Neiva, Pedro Amoedo, Diego Lima, Lorena Moreira, Fabiani Santos e Magda Santos por todo carinho, amizade, companheirismo e diversão que tivemos em Piracicaba;

Aos amigos Ana Paula Silva, Guilherme Nacata e Thiago Viana por todos os momentos especiais que vocês me proporcionaram no meu último ano em Piracicaba;

À todos os professores do Departamento de Ciências Exatas da ESALQ / USP;

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

À Universidade Federal do Oeste do Pará (UFOPA), em especial ao Centro de Formação Interdisciplinar, por todo apoio a minha qualificação como Docente desta Universidade;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior(CAPES) e a Fundação da Amazônia de Amparo a Estudo e Pesquisa (FAPESPA) pelo apoio financeiro a este estudo a partir do programa de formação doutoral docente(Prodoutoral).

Finalmente, a todos, que de alguma forma contribuíram para a realização deste trabalho.

EPÍGRAFE

"Talvez não tenha conseguido fazer o melhor,
mas lutei para que o melhor fosse feito.
Não sou o que deveria ser, mas Graças a Deus,
não sou o que era antes".

Marthin Luther King

SUMÁRIO

Resumo . . . 7
Abstract . . . 8
1 Introdução . . . 9
Referências . . . 10
2 Análise Espacial da Leprose dos Citros via estimação INLA . . . 13
Resumo . . . 13
Abstract . . . 13
2.1 Introdução . . . 13
2.2 Metodologia . . . 14
2.2.1 O experimento . . . 14
2.2.2 Definição dos modelos . . . 15
2.2.2.1 Modelos Hierárquicos Bayesianos . . . 16
2.2.3 Integrated Nested Laplace Approximation - INLA . . . 17
2.3 Seleção dos Modelos . . . 19
2.3.1 Critério de informação deviance (DIC) . . . 19
2.3.2 Condicionais preditivas ordenadas (CPO) . . . 19
2.4 Resultados e Discussão . . . 20
2.5 Conclusão . . . 23
2.6 Agradecimentos . . . 23
Referências . . . 23
3 Imputação Múltipla via Método da Média Preditiva: uma comparação entre o algoritmo MICE e o método INLA. . . . 25
Resumo . . . 25
Abstract . . . 25
3.1 Introdução . . . 25
3.2 Metodologia . . . 26
3.2.1 Imputação Múltipla . . . 26
3.2.1.1 Multivariate Imputation by Chained Equation - MICE . . . 27
3.2.1.2 Integrated Nested Laplace Approximation - INLA . . . 28
3.2.1.3 Método da Média Preditiva (PMM – Predictive Mean Meating) . . . 30
3.2.2 Validação do Modelo . . . 31
3.2.3 Estudo de Simulação . . . 31
3.3 Aplicação em dados de leprose do citros . . . 33
3.3.1 Área de Estudo . . . 33
3.3.2 Resultados e Discussão . . . 33
3.3.2.1 Critério de informação deviance (DIC) . . . 34
3.3.2.2 Condicionais preditivas ordenadas (CPO) . . . 34
3.4 Conclusão . . . 36
3.5 Agradecimentos . . . 36
Referências . . . 36
4 Considerações finais . . . 39

RESUMO

Modelos para análise de dados espaciais binários: uma abordagem Bayesiana

A relação da geoestatística e agricultura de precisão podem resultar em tomada de decisões e intervenções específicas. A geoestatística e a agricultura de precisão possuem ferramentas que são complementares para o entendimento das variações que ocorrem no campo. Em várias situações a geoestatística oferece recursos ímpares a agricultura de precisão e, reciprocamente, a agricultura de precisão proporciona recursos que permitem determinar a estrutura de dependência espacial e juntas permitem construir cenários para tomada de decisões e ações práticas de manejo específico. A estatística bayesiana vem sendo utilizada como método de inferência na análise estatística de dados espaciais. Dada a importância atual e aplicações de modelagem de fenômenos espaciais com dados binários não serem tão abordadas na literatura, este trabalho tem por objetivo apresentar um estudo de modelos propostos na literatura para analisar dados espaciais binários utilizando modelo hierárquicos bayesianos, que envolvem o uso de modelos autorregressivos condicionais e modelos lineares generalizados para capturar a estrutura espacial dos dados. Foi verificado no primeiro artigo uma modelagem geoestatística espacial para analisar os dados de leprose dos citros em plantações de laranja. O uso de uma classe flexível de modelos, associada ao método de inferência computacionalmente eficiente através da aproximação de Laplace aninhada integrada (INLA), tornou a abordagem mais interessante e possibilitou a incorporação desta metodologia em sistemas da citricultura brasileira. Sabe-se que tanto no aspecto espacial quanto no espaço-temporal, perdas de observações ocorrem por diversos fatores. E, técnicas que possibilitem tratar tais tipos de dados tem sido propostas, dentre elas a imputação. No segundo artigo, utiliza-se o método de imputação múltipla da média preditiva para comparar o ajuste feito pelo método INLA com a utilização da imputaçã MICE e imputação INLA. A idéia foi fornecer diferentes tamanhos e perdas amostrais para comparar os dois procedimentos de imputação. Foi verificado na imputação INLA que a medida que aumentavam-se o tamanho e as perdas amostrais o valor da raiz normalizada do erro quadrático médio (NRMSE) ia diminuindo e para a imputação MICE esse valor aumentava. Porém, apesar de aumentar o valor do NRMSE neste método, o mesmo ainda apresentou valores imputados muito mais próximos aos valores originais. Mostrando assim, que o método de imputação da média preditiva utilizado na imputação MICE teve um melhor desempenho para dados espaciais binários.

Palavras-chave: Estatística Bayesiana, Dados Binários, Modelos Hierárquicos, Análise Espacial, Imputação Múltipla

ABSTRACT

Models for spatial analysisbof binary data: a Bayesian approach

The relationship of geostatistics and precision agriculture can result in specific decision making and interventions. Geostatistics and precision agriculture have tools that are complementary to understanding the variations that occur in the field. In many situations, geostatistics offers unique resources to precision agriculture and, conversely, precision agriculture provides resources that allow the determination of the spatial dependence structure and together allow the construction of scenarios for decision-making and practical actions for specific management. Bayesian statistics has been used as an inference method in the statistical analysis of spatial data. Given the current importance and applications of modeling spatial phenomena with binary data are not so covered in the literature, this paper aims to present a study of models proposed in the literature to analyze binary spatial data using hierarchical Bayesian models, which involve the use of models conditional autoregressives and generalized linear models to capture the spatial structure of the data. A geostatistical spatial modeling to analyze citrus leprosis data in orange plantations was verified in the first article. The use of a flexible class of models, associated with the computationally efficient inference method through the integrated nested Laplace approximation (INLA), made the approach more interesting and enabled the incorporation of this methodology in Brazilian citrus farming systems. It is known that both in the spatial and in the spatiotemporal aspect, losses of observations occur due to several factors. And, techniques that make it possible to treat such types of data have been proposed, including imputation. In the second article, the predictive mean multiple imputation method is used to compare the adjustment made by the INLA method with the use of MICE imputation and INLA imputation. The idea was to provide different sample sizes and losses to compare the two imputation procedures. It was verified in the INLA imputation that as the sample size and losses increased, the value of the normalized root mean squared error (NRMSE) decreased and for the MICE imputation this value increased. However, despite increasing the NRMSE value in this method, it still presented imputed values much closer to the original values. Thus, showing that the predictive mean imputation method used in MICE imputation had a better performance for binary spatial data.

Keywords: Bayesian Statistics, Binary Data, Hierarchical Models, Spatial Analysis, Multiple Imputation

1 INTRODUÇÃO

A relação da geoestatística e agricultura de precisão podem resultar em tomada de decisões e intervenções específicas. Entretanto, a área está em pleno desenvolvimento e muitas aplicações e soluções na agricultura já foram resolvidas por análises geoestatísticas e muitas outras poderão ser criadas trazendo benefícios para a agricultura. A agricultura de precisão é uma maneira de gerenciar a agricultura levando em consideração a variabilidade temporal e espacial da produtividade das culturas e dos fatores a ela inerentes. Porém, há necessidade de profissionais qualificados para interpretar e modelar corretamente a variabilidade existente e entender a necessidade de tomada de decisões e intervenções localizadas, realizadas com base no retorno econômico, social e ambiental das atividades. Gebbers e Adamchuk (2010), definem a agricultura de precisão como um ciclo que inicia com a coleta, passa por análise e interpretação e finaliza com a geração de protocolos que serão implantados no campo. A geoestatística e a agricultura de precisão possuem ferramentas que são complementares para o entendimento das variações que ocorrem no campo. Em várias situações a geoestatística oferece recursos ímpares a agricultura de precisão e, reciprocamente, a agricultura de precisão proporciona recursos que permitem determinar a estrutura de dependência espacial e juntas permitem construir cenários para tomada de decisões e ações práticas de manejo específico.

O estudo da variabilidade espacial de algumas variáveis vem sendo objeto de preocupações de pesquisadores desde o principio do século passado. Segundo Vieira (2000), os estudos sobre variabilidade espacial foram iniciados por Smith (1938), estudando a disposição de parcelas no campo em experimentos de rendimento de variedades de milho, numa tentativa de eliminar o efeito de variações do solo. Para Queiroz (1995), variabilidade espacial é uma característica da maioria dos fenômenos naturais, podendo-se incluir, dentre eles, os parâmetros relacionados com o movimento da água no solo. Os procedimentos adotados na época utilizavam a estatística clássica e utilizavam grandes quantidades de dados amostrais, visando caracterizar ou descrever a distribuição espacial da característica em estudo. Por estatística clássica, entende-se aquela na qual se utilizam parâmetros como média e desvio padrão para representar um fenômeno, e se baseia na hipótese principal de que as variações de um local para outro são aleatórias e independentes. Krige (1951), trabalhando com dados de concentração de ouro, concluiu que somente a informação dada pela variância seria insuficiente para explicar a variabilidade do fenômeno para tal, seria necessário levar em consideração a distância entre as observações. A partir daí, surge o conceito da geoestatística, que leva em consideração a localização geográfica e a dependência espacial. Matheron (1963, 1971), baseado nas observações de Krige, desenvolveu a Teoria das Variáveis Regionalizadas ou metodologia geoestatística. Uma variável regionalizada é uma função numérica com distribuição espacial, que varia de um ponto a outro com continuidade aparente, a cujas variações devem ser ajustadas funções teóricas. Segundo Woodbury e Sudicky (1991), a estimação dos parâmetros para descrever a variação espacial de uma propriedade é sempre um problema de difícil resolução. Quando muitos dados são avaliados, a ênfase deve ser em como os vários modelos utilizados podem afetar o valor do parâmetro inferido.

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*. Essa classe de modelos viabiliza acomodar a dependência relacionada as medidas espaciais, e as estruturas complexas referentes às variabilidades inerentes ao estudo observacional. Modelagem de dados de área tem sido tema de

diversas pesquisas em Estatística nas últimas décadas. Modelos espaço-temporais têm sido utilizados para lidar com esse tipo de dados de um modo natural, uma vez que envolvem processos ambientais, epidemiológicos, ecológicos, entre outros que têm, em geral, transições no tempo e no espaço. O avanço da tecnologia e, simultaneamente, de métodos estatísticos, têm permitido a elaboração de modelos cada vez mais estruturados para a descrição de fenômenos aleatórios complexos. A estatística bayesiana pode ser utilizada como método de inferência na análise estatística de dados espaciais. Em geral, o objetivo da inferência estatística é estimar parâmetros de uma certa população e/ou testar hipóteses sobre eles, por meio de uma amostra extraída desta população. A ideia é descrever, de forma realista, a estrutura de correlação presente nos dados, o que pode ser feito através do uso de modelos hierárquicos bayesianos (Silva, 2014).

Ao se trabalhar com um conjunto de dados com estrutura espacial, pode-se assumir que existe um efeito aleatório relacionado a cada região em estudo. Esse efeito pode apresentar algum tipo de estrutura, em que cada região terá um nível de dependência com suas regiões vizinhas. Uma forma de incorporar estruturas de dependência espacial para o efeito aleatório é utilizar o modelo Conditional Autoregressive (Besag, 1974). Neste modelo a estrutura de vizinhança é dada por um grafo não direcionado representado por uma matriz composta de zeros e uns, em que o número um significa que as regiões são vizinhas e zero significa que as regiões não são vizinhas. A partir da matriz de vizinhança deriva-se a estrutura de dependência entre as regiões em estudo. Dada a importância atual e aplicações de modelagem de fenômenos espaciais com dados binários não serem tão abordadas na literatura, este trabalho tem por objetivo o estudo de modelos propostos na literatura para analisar dados espaciais de área utilizando modelo hierárquicos bayesianos, que envolvem o uso de modelos autorregressivos condicionais e modelos lineares generalizados para capturar a estrutura espacial dos dados.

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 danificadas, 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, 2007).

Neste contexto, no capítulo 2 será apresentado uma modelagem geoestatística espacial para analisar os dados de leprose dos citros em plantações de laranja. O uso de uma classe flexível de modelos, associada ao método de inferência computacionalmente eficiente através da aproximação de Laplace aninhada integrada (INLA), torna a abordagem mais interessante e possibilita a incorporação desta metodologia em sistemas da citricultura brasileira. No capitulo 3, será realizado um estudo de comparação entre o algoritmo MICE(Multivariate Imputation by Chained Equation) e a abordagem INLA no uso do método de imputação da média preditiva. E, no capitulo 4 será apresentado as considerações finais.

Referências

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society (Series B)*, 36:192–236.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Diggle, P. J., Tawn, J. A., e Moyeed, R. A. (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.

Gebbers, R. e Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*, 327:828–831.

Krige, D. G. (1951). *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. Thesis (magister science engineering), University of the Witwatersrand, Johannesburg, 1951.

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

Matheron, G. (1971). *The theory of regionalized variables and its applications*. Les cahiers du Centre de Morphologie Mathématique de Fontainebleau. École Nationale Supérieure des Mines.

Queiroz, J. E. (1995). *Parâmetros hidro-dinâmicos de um solo de várzea para fins de drenagem subterrânea*. Tese (doutorado em irrigação e drenagem), Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1995.

Silva, S. U. (2014). *Modelo dinâmico bayesiano multivariado para análise espaço-temporal de dados de área*. Dissertação (mestrado em matemática aplicada e computacional), Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Ciências e Tecnologia, Presidente Prudente, 2014.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28:1–23.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–243.

Vieira, S. R. (2000). Geoestatística em estudos de variabilidade espacial do solo. in: Novais, r. f et al (eds). *Tópicos Em Ciência Do Solo, 1. Sociedade Brasileira De Ciência Do Solo*, pages 1–53.

Woodbury, A. D. e Sudicky, E. A. (1991). The geostatistical characteristics of the borden aquifer. *Water Resources Research*, 27(4):533–546.

2 ANÁLISE ESPACIAL DA LEPROSE DOS CITROS VIA ESTIMAÇÃO INLA

Resumo

Este trabalho propõe uma modelagem geoestatística espacial para analisar os dados de leprose dos citros em plantações de laranja. O uso de uma classe flexível de modelos, associada ao método de inferência computacionalmente eficiente através da aproximação de Laplace aninhada integrada (INLA), torna a abordagem mais interessante e possibilita a incorporação desta metodologia em sistemas da citricultura brasileira. Assim, para analisar o padrão espacial da doença leprose dos citros em um talhão com 594 plantas, foram construídos modelos hierárquicos bayesianos com estrutura gaussiana latente para representar os efeitos espaciais.

Palavras chave: Leprose do Citros; Laplace aninhada integrada; Bayesiana; Distribuição no espaço.

Abstract

This work proposes a spatial geostatistical model to analyze citrus leprosis data in orange plantations. The use of a flexible class of models, associated with the computationally efficient inference method through the integrated nested Laplace approximation (INLA), makes the approach more interesting and enables the incorporation of this methodology in Brazilian citrus systems. Thus, to analyze the spatial pattern of citrus leprosis disease in a stand with 594 plants, Bayesian hierarchical models with latent Gaussian structure were built to represent the spatial effects.

Keywords: Citrus Leprosis; Integrated nested Laplace; Bayesian; Distribution in space.

2.1 Introdução

Um pomar cítrico, por ser formado de plantas perenes, se constitui num ambiente bastante complexo onde vivem milhares de espécies de insetos. Tratos culturais adequados são de fundamental importância para o equilíbrio entre os insetos pragas e os inimigos naturais (insetos úteis). As plantas novas de até quatro anos são as mais afetadas e sofrem mais com o ataque das pragas. Nessa faixa de idade, é praticamente impossível se dispensar o controle químico, no entanto não se deve abusar dos agrotóxicos. Só aplicar o inseticida nas plantas com ataque intenso. A inspeção periódica do pomar é de fundamental importância, pois permite detectar a presença das pragas e o seu grau de infestação. O conhecimento da distribuição no espaço ou espaço-temporal de parasitas e doenças usando ferramentas adequadas pode ajudar no desenvolvimento de planos de amostragem de maior número de representantes para cenários agrícolas. Esses exames podem ajudar na verificação de práticas de manejo com menor impacto ambiental e maior viabilidade econômica.

A leprose dos citros (*Citrus leprosis virus* - CiLV) é uma das doenças mais importantes na cultura da laranja doce no Brasil, cujos custos para seu controle chegam a US$ 80 milhões por ano (Bassanezi e Laranjeira, 2007). A doença acarreta desde a depreciação do fruto até queda prematura das folhas e frutos, podendo reduzir a vida útil da planta. Os sintomas da leprose são localizados e incluem descamação intensa nos ramos, manchas anelares cloróticas nas folhas e manchas circulares necróticas deprimidas na região central nos frutos (Bitancourt, 1955). Sua disseminação no pomar é facilitada pela presença do ácaro*Brevipalpus yothersi Baker* (Acari: *Tenuipalpidae*), conhecido como ácaro da leprose dos citros (CiLV) ou ácaro plano (Rodrigues et al., 2008). O CiLV é considerado não-sistêmico, portanto a presença de vetores e a existência de plântulas assintomáticas são necessárias para sua disseminação, comprometendo a produção de frutos e levando plantas jovens mais susceptíveis à morte (Locali et al., 2003).

Segundo Bastianel et al. (2010), a leprose dos citros ocorre em 22 Estados brasileiros, onde há o cultivo dessa cultura, com maior importância econômica para o Estado de São Paulo que é responsável por cerca de 80 % da produção nacional de laranja, distribuída em 322 municípios. A atividade econômica é essencial nesses municípios, pois o setor emprega diretamente cerca de 400 mil pessoas (Borges e Almeida, 2000). Entretanto, a produção citrícola poderia ser muito maior se não fosse a ação de inúmeras pragas e doenças que acometem a cultura. Somente para o controle do ácaro da leprose, os custos chegam a variar de US$ 60 a 75 milhões por ano, um impacto da ordem de 35% no total de insumos e 15% nos custos finais da produção. Desta forma, a melhoria no manejo dos ácaros e da doença pode contribuir decisivamente para a diminuição dos custos de produção e no valor dessa "commoditie"(Bassanezi e Laranjeira, 2007; Rodrigues et al., 2008).

Devido à importância do controle dessa doença para a citricultura brasileira, muitos trabalhos vêm sendo desenvolvidos nos últimos anos enfatizando medidas de controle do ácaro vetor e da doença (Amorim et al., 2006; Silva et al., 2012). Segundo Gyenge e Edelstein (1999), para a escolha do melhor critério para o controle de uma praga é necessário o conhecimento da distribuição espacial da população. A geoestatística é uma ferramenta que tem sido, nos últimos anos, bastante aplicada para modelar o padrão espaço-temporal e formular hipóteses a respeito do aspecto epidemiológico de doenças de plantas (Charest et al., 2002; Farias et al., 2008).

2.2 Metodologia

2.2.1 O experimento

A área de estudo localiza-se no município de Taquaral, no estado de São Paulo, Brasil. O clima da região segundo a classificação de Koppen (1948) é do tipo Aw, com verão quente e úmido e inverno com estação seca. A precipitação pluvial média anual é de 1.629 mm, concentrada no período de setembro a março. A temperatura média anual é de 23 °C. Os experimentos foram instalados em duas áreas com incidência de leprose dos citros e ácaros *Brevipalpus*, onde todas as plantas foram georreferenciadas. Em ambas as áreas, as linhas de plantio seguiram a direção leste-oeste. Os dados climáticos (precipitação, temperatura, velocidade e direção do vento) foram coletados a partir da estação automática do Instituto Nacional de Meteorologia (INMET) e de uma estação agrometeorológica próxima às áreas experimentais. O estudo nessas áreas foi desenvolvido de março de 2013 a dezembro de 2014. O presente trabalho fez o estudo apenas na Área 1.

O talhão desse estudo é composto por 594 laranjeiras "Pera" (*C. sinensis*) enxertadas sobre limoeiro "Cravo" (*C. limonia*) com espaçamento de 6,5 x 4,5 metros e com 14 anos de idade. O talhão apresentava uma área total de 17.374,5 m^2 e encontrava-se margeado ao norte por cultura de soja [*Glycine max* (L.) Merrill], ao sul por solo descoberto, ao leste por pomar cítrico convencional e a oeste por estrada de terra e pastagem. Este talhão vinha sendo tratado com produtos fitossanitários comumente utilizados no controle de pragas, doenças e plantas infestantes de pomares cítricos. Após a seleção deste talhão, o experimento foi instalado em março de 2013 e todas as plantas foram georreferenciadas (Figura 2.1).

A incidência da leprose dos citros desta área foram avaliadas em março de 2013, setembro de 2013 e janeiro de 2014. Neste talhão, foram demarcadas 594 plantas distribuídas uniformemente no talhão, cada planta demarcada obteve um alcance de 27 metros por planta na mesma linha e 6,5 metros por plantas entre linhas. O número mínimo de amostras foi definido usando a variabilidade inferida pelos coeficientes de variação (Chen e Peace, 2011).

Figura 2.1. A) Mapa da área localizada o município de Taquaral-SP e B) Área de produção do citros em estudo.

2.2.2 Definição dos modelos

No geral, os estudos para descrever o padrão espacial de plantas doentes são usadas técnicas predominantemente descritivas que visam detectar aglomerações espaciais de incidência da doença. Exemplos de técnicas são: a análise por quadrantes, variogramas, e o uso de diversos índices de associação (Lima et al., 2006). Entretanto, a necessidade de quantificar padrões e estabelecer relações entre fatores que afetam a incidência, levou a proposição de modelos que permitem ir além da simples detecção da presença de padrões espaciais nos dados (Franciscon et al., 2008). Uma classe inicial de modelos que atende a essa especificação é a classe dos modelos lineares generalizados, originalmente introduzida por Nelder e Wedderburn (1972) que será descrita com mais detalhes abaixo. A notação utilizada segue a do artigo (Rue et al., 2009), onde os autores não fazem distinção entre y_i para a variável aleatória e dados observados.

Considere o Modelo Linear Generalizado (GLM) o qual assume que, dado um conjunto de covariáveis $\boldsymbol{z}$ e parâmetros desconhecidos $\boldsymbol{\beta}$, a distribuição da variável resposta y pertence a família exponencial, ou seja,

$$P(\boldsymbol{y}|\mathbf{z}) = \exp\left(\frac{\boldsymbol{y\theta} - \mathbf{b}(\boldsymbol{\theta})}{\boldsymbol{\phi}}\right)\mathbf{c}(\boldsymbol{y}, \boldsymbol{\phi}) \tag{2.1}$$

em que $b(\cdot)$, $c(\cdot)$ são funções que dependem da distribuição atribuída a variável resposta, $\boldsymbol{\phi}$ é um parâmetro de escala ou dispersão comum para todas as observações, e $\boldsymbol{\theta}$ é o parâmetro natural da família exponencial. A média de cada variável aleatória $E(y_i|\boldsymbol{z}, \boldsymbol{\beta})$ é ligada ao preditor linear da forma

$$g(\mu_i) = \eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i, \tag{2.2}$$

em que $g(\cdot)$ é uma função conhecida chamada de função de ligação, $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ são parâmetros de regressão desconhecidos, e $\boldsymbol{\varepsilon}_i$ é um termo não estruturado.

É importante notar que no preditor assume-se que todas as covariáveis $\boldsymbol{z}$ têm um relacionamento estritamente linear com a resposta. Esta restrição impossibilita uma análise adequada do conjunto de dados, principalmente quando deseja-se incluir efeitos espaciais e temporais que variem de forma *suave* ao longo das coordenadas e valores de tempo, bem como, efeitos não-lineares de covariáveis contínuas. Para

superar estas dificuldades, pode-se trocar o preditor estritamente linear por um preditor aditivamente estruturado

$$g(\mu_i) = \boldsymbol{\eta}_i = \boldsymbol{\alpha} + \sum_{j=1}^{n_f} f^{(j)}(\boldsymbol{u}_{ji}) + \sum_{k=1}^{n_\beta} \beta_k \boldsymbol{z}_{ki} + \boldsymbol{\varepsilon}_i \tag{2.3}$$

A diferença de 2.2 para 2.3 são as funções $f(\cdot)$ desconhecidas das covariáveis u, e é justamente pelas diferentes formas que estas funções podem tomar, que esta classe de modelos é extremamente flexível, para incorporar efeitos aleatórios estruturados ou não no modelo.

2.2.2.1 Modelos Hierárquicos Bayesianos

Os modelos hierárquicos Bayesianos são apropriados para modelar problemas com estruturas complexas de dependência.Podem ser generalizados como modelos em que os parâmetros do modelo que descreve a variabilidade da variável de interesse são considerados quantidades aleatórias oriundas de uma distribuição de probabilidade cujos parâmetros são chamados de hiperparâmetros. E, a distribuição dos parâmetros *a priori* possui hiperparâmetros, que são estimados pelos dados.

Um modelo hierárquico bayesiano pode ser representado considerando, pelo menos, três níveis de hierarquia:

- 1º nível: modelo das observações: de Y com os parâmetros;
- 2º nível: modelos dos parâmetros com seus hiperparâmetros;
- 3º nível: modelo dos hiperparâmetros.

Quando não se tem uma idéia *a priori* sobre o tipo de heterogeneidade dos efeitos de área, têm-se uma alternativa sugerida por (Besag et al., 1991), e amplamente utilizada na literatura epidemiológica, representa o risco de cada área como a soma de dois componentes aleatórios independentes, um deles (ϕ_i), seguindo um padrão de heterogeneidade não estruturada que representa efeitos aleatórios sem dependência espacial na área i, enquanto que o segundo componente (φ_i), segue um padrão de heterogeneidade estruturada e representa efeitos aleatórios espacialmente dependentes da área i, isto é,

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\phi}_i + \boldsymbol{\varphi}_i \tag{2.4}$$

Segundo Assunção (2001), a primeira parte do modelo não possui uma estrutura espacial e representa efeitos de pequena escala, ou diferenças entre uma área e outra que não ultrapassam suas fronteiras geográficas. Estes efeitos poderiam, por exemplo, ser ações de saúde pública ou características socio-econômicas locais não compartilhadas pelas áreas vizinhas. Assim, por serem efeitos específicos de cada área, podem ser modelados como efeitos aleatórios independentes. A segunda parte do modelo possui uma estrutura espacial e é definida de modo a considerar que áreas vizinhas são mais semelhantes em termos de seu risco relativo do que duas áreas arbitrariamente escolhidas na região. Portanto, este componente representa efeitos de longa escala, de alcance geográfico maior do que as fronteiras das áreas que estão sendo consideradas. Muitas variáveis observadas, ou não, que afetam o risco de uma doença em particular, possuem esta característica. Por exemplo, características ambientais, em geral, mudam suavemente no espaço e assim, áreas próximas tendem a ter valores similares para estas variáveis.

Em termos de modelagem, é comum assumir que os elementos do primeiro componente da equação (2.4) possuem, *a priori*, distribuições normais independentes com média zero e variância σ^2_ϕ.

Assim, $\phi = (\phi_1, .., \phi_I)^T$ tem uma distribuição normal multivariada composta de termos independentes. Esta distribuição depende de um parâmetro de precisão $\tau_\phi = 1/\sigma_\phi^2$, que em geral é desconhecido. Este parâmetro controla o grau de dispersão dos efeitos aleatórios sem estrutura espacial. A distribuição *a priori* escolhida para o segundo componente $\varphi = (\varphi_1, ..., \varphi_I)^T$ que tem uma estrutura espacial é tipicamente um campo aleatório Markoviano Gaussiano. Esta distribuição, além de ter uma estrutura de vizinhança espacial, depende de um parâmetro desconhecido adicional τ_φ inversamente relacionado à variabilidade dos $\varphi_i's$ e que mede o grau de similaridade espacial entre os componentes.(Assunção, 2001)

Os efeitos aleatórios espacialmente estruturados de um modelo hierárquico Bayesiano seguem uma distribuição contínua. O modelo *a priori* mais popular para este tipo de efeitos é aquele em que a distribuição conjunta é normal multivariada. Em particular, as auto-regressões condicionais também conhecidas como modelos CAR (Besag, 1974), têm sido amplamente usadas na literatura.

Considere n áreas de uma variável aleatória $\boldsymbol{x} = (x_1, ..., x_n)$, cada uma caracterizada por um grupo de vizinhos N_i, assume-se que x_i segue a seguinte distribuição:

$$x|\boldsymbol{x}_{-i} \sim Normal\left(\mu_i + \sum_{j=1}^{n} r_{ij}(x_j - \mu_j), s_i^2\right) \tag{2.5}$$

em que μ_j é a média da área j e $s_i^2 = (\sigma_x^2)/N_i$ é a variância da área i, que depende do número de vizinhos $N_i = \#N_{(i)}$ o que faz com que à medida que o número de vizinhos da área i aumente, sua variância diminua. Esta estrutura de variância estabelece que na presença de alta correlação espacial, quanto mais vizinhos uma área tiver, mais informação terá para estimar o seu efeito aleatório, enquanto que o parâmetro de variância σ_x^2 controla a variação dos efeitos aleatórios estruturados espacialmente. A quantidade r_{ij} indica a proximidade espacial e é calculada como $\phi \times W_{ij}$, em que $W_{ij} = a_{ij}/N_i$, a_{ij} é 1 se as áreas i e j são vizinhas e 0 caso contrário, enquanto que o parâmetro ϕ controla a adequação da distribuição.

Para completar a especificação Bayesiana do modelo, designa-se as distribuições *a priori* para os hiperparâmetros dos campos Gaussianos. Os hiperparâmetros para o modelo são τ_ϕ e τ_φ, eles determinam a variação de cada bloco, e precisam ser estimados através dos dados. Seguindo a proposta de Knorr-Held (2000) assume-se para todos estes parâmetros prioris Gamma próprias, $\tau \sim G(a; b)$, para evitar problemas com hiperprioris impróprias. Prioris Gamma são computacionalmente convenientes, porque a condicional completa de τ vai também ser Gamma. Para todas as aplicações neste trabalho, hiperprioris Gamma de alta dispersão foram escolhidas para todos os blocos com valores a = 0,01 e b = 0,01.

Devido a complexidade dos modelos em estudo e o uso da abordagem Bayesiana como procedimento de inferência, será utilizada a abordagem INLA (Integrated Nested Laplace Approximation) na estimação dos parâmetros de interesse.

2.2.3 Integrated Nested Laplace Approximation - INLA

A classe de modelos que a metodologia INLA trabalha é denominada modelos Gaussianos Latentes onde a variável dependente y_i pertence a uma família de distribuições onde a sua média μ_j é conectada a um preditor linear aditivo η_i a partir de uma função de ligação $g(\mu_i) = \eta_i$ tal que

$$g(\mu_i) = \boldsymbol{\eta}_i = \boldsymbol{\alpha} + \sum_{j=1}^{n_f} f^{(j)}(\boldsymbol{u}_{ji}) + \sum_{k=1}^{n_\beta} \beta_k \boldsymbol{z}_{ki} + \boldsymbol{\varepsilon}_i,$$

em que β_k representa o efeito linear de covariáveis $\boldsymbol{z}$, $f^{(h)}(\cdot)$ são funções não conhecidas de covariáveis $\boldsymbol{u}$ e $\boldsymbol{\varepsilon}_i$ são os termos não estruturados, utilizados para acomodar a correlação espacial, temporal e/ou espaço-temporal, caso exista. Esta classe de modelos também pode ser escrita como uma classe de

modelos hierárquicos Bayesianos em três estágios: a verossimilhança, o campo latente e *a priori* dos hiperparâmetros.

Segundo Rue et al. (2009), a definição dos modelos gaussianos latentes (MGL) é dada como sendo modelos que assumem a forma aditiva apresentada em 2.6 e assume-se *priori* gaussiana para todos os parâmetros que compõe η, o vetor desses parâmetros é chamado de campo aleatório gaussiano (CAG). Esta classe de modelos também pode ser escrita como uma classe de modelos hierárquicos Bayesianos em três estágios: a verossimilhança, o campo latente e *a priori* dos hiperparâmetros.

De acordo com Dey et al. (2000), o preditor linear dos modelos lineares generalizados pode ser expresso por,

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k \boldsymbol{z}_{ki}. \tag{2.6}$$

Segundo Martins et al. (2013), o modelo descrito em 2.6 pode ser escrito utilizando uma estrutura hierárquica, em que o primeiro estágio é formado pela função de verossimilhança com propriedades de independência condicional, dada pelo campo latente gaussiano $\boldsymbol{x} = (\boldsymbol{\eta}, \alpha, \boldsymbol{\beta})$ e possíveis hiperparâmetros $\boldsymbol{\theta}$ (associados aos componentes de variância). Assim, o primeiro estágio da hierarquia é dado por,

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) = L(\boldsymbol{x}, \boldsymbol{\theta}_1|\boldsymbol{y}) = \prod_{i=1}^{n_d} \pi(y_i|x_i, \boldsymbol{\theta}_1).$$

A distribuição condicional do campo latente $\boldsymbol{x}$, dados alguns possíveis hiperparâmetros $\boldsymbol{\theta}_2$ forma o segundo estágio do modelo e tem uma distribuição Gaussiana conjunta.

$$\boldsymbol{x}|\boldsymbol{\theta}_2, \boldsymbol{y} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}_2, \boldsymbol{y}) = N(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)),$$

em que $N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ denota a distribuição gaussiana multivariada com vetor de média $\boldsymbol{\mu}$ e matriz de precisão $\boldsymbol{Q}$. Na maioria das aplicações, os campos Gaussianos latentes possuem propriedades de independência condicional, o que se traduz em uma matriz de precisão esparsa $\boldsymbol{Q}(\boldsymbol{\theta}_2)$, que é de extrema importância para reduzir o custo computacional. Uma distribuição Gaussiana multivariada com uma matriz de precisão esparsa é conhecida como um campo aleatório de Markov de Gauss (GMRF) (Rue e Held, 2005).

O terceiro e último estágio do modelo hierárquico é dado pela especificação de uma distribuição a *priori* apropriada para cada um dos hiperparâmetros do modelo,

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

Para o modelo descrito anteriormente, a distribuição conjunta a *posteriori* para os parâmetros do modelo linear generalizado, pode ser expressa por,

$$\begin{aligned} \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n_d} \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{n/2} \exp\left[-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n_d} \log \pi(y_i|x_i, \boldsymbol{\theta})\right]. \end{aligned} \tag{2.7}$$

Segundo Rue et al. (2009), as distribuições marginais a *posteriori* de interesse podem ser expressas por,

$$\begin{aligned} \pi(x_i|\boldsymbol{y}) &= \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}, \\ \pi(\theta_j|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \boldsymbol{\theta}_{-j}, \end{aligned}$$

em que $i = 1, \ldots, I$, sendo I o número de parâmetros de efeito fixo e efeito aleatório do modelo, e $j = 1, \ldots, J$, sendo J o número de componentes de variância. Rue et al. (2009) utilizam as equações

anteriores para construir as aproximações aninhadas,

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}, \tag{2.8}$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}_{-j}. \tag{2.9}$$

Para as equações (2.8) e (2.9), Rue et al. (2009) sugerem dividir o método de aproximação INLA em três etapas: Primeiramente propor uma aproximação $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ para a distribuição conjunta a *posteriori* dos hiperparâmetros $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. O segundo passo consiste em propor uma aproximação $\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ para as marginais da distribuição condicional do campo latente dado as observações e os hiperparâmetros $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$, e finalmente explorar $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ em uma grade de valores e usá-los para integrar $\boldsymbol{\theta}$ na Equação (2.8) e $\boldsymbol{\theta}_{-j}$ na Equação (2.9).

2.3 Seleção dos Modelos

Foram ajustados cinco modelos e para todos estes, a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. O modelo nulo (apenas com intercepto), um modelo que considera que não existe nenhum tipo de relação espacial entre as observações, outro que considera que existe relação espacial entre as observações e por ultimo um modelo com os dois componentes espaciais (não estruturado e estruturado).

Para comparação destes modelos e consequente escolha do modelo que melhor descreve a resposta, foram utilizados três critérios em conjuntos: o critério de informação deviance (DIC), o número efetivo de parâmetros e a ordenada preditiva condicional (*Conditional Predictive Ordinate*, CPO).

2.3.1 Critério de informação deviance (DIC)

Para o cálculo do DIC considera-se a medida *deviance* dada por $D = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, em que $\boldsymbol{y}$ é o vetor de dados, $\boldsymbol{\theta}$ é o vetor de parâmetros de interesse sob determinado modelo e $p(\boldsymbol{y}|\boldsymbol{\theta})$ a função de verossimilhança. O DIC é calculada como $DIC = \hat{D}+2p_D$, em que $\hat{D} = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D]$ é uma medida que indica a qualidade do ajuste, e p_D o número efetivo de parâmetros, representando a complexidade do modelo. Segundo Spiegelhalter et al. (2000), a quantidade efetiva de parâmetros do modelo é calculado como a média a *posteriori* da *deviance* menos a *deviance* avaliada na média a *posteriori*, ou seja, $p_D = \hat{D} - D(\hat{\boldsymbol{\theta}})$.

A ideia é que os modelos com menor DIC devem ser preferidos a modelos com maiores valores para essa estatística. Os modelos são penalizados tanto pelo valor de $\hat{D}$, o que favorece um bom ajuste, como também pelo número efetivo de parâmetros p_D. É importante ressaltar que para o cálculo do DIC assume-se que a média a *posteriori* é uma boa estimativa pontual para os parâmetros e, portanto, não representa uma boa medida para a avaliação do ajuste na presença de distribuições a *posteriori* assimétricas para os parâmetros do modelo.

2.3.2 Condicionais preditivas ordenadas (CPO)

A estatística CPO é um método de avaliação cruzada utilizado para avaliar a performance do modelo medindo o seu poder preditivo (Geisser e E., 1979). A distribuição preditiva a *posteriori* para a i-ésima observação é dada por

$$p(\boldsymbol{y}_i|\boldsymbol{y}_{-i}) = \int p(\boldsymbol{y}_i|\boldsymbol{\theta}) \prod(\boldsymbol{\theta}|\boldsymbol{y}_{-i})\, d\boldsymbol{\theta},$$

em que $\boldsymbol{y}_i$ é a i-ésima observação de $\boldsymbol{y}$ e $\boldsymbol{y}_{-i}$ é $\boldsymbol{y}$ sem a observação i. Então, o CPO é a probabilidade a *posteriori* de observar y_i quando o modelo é ajustado sem y_i. Quanto maior o valor da estatística CPO

melhor é o ajuste do modelo aos dados. Ao invés de ajustar o modelo a cada vez que é retirada uma observação, uma aproximação do CPO pode ser obtida por amostragem da distribuição a *posteriori*:

$$CPO_i = \left(\frac{1}{T \sum_{t=1}^{T} p\left(\boldsymbol{y}_i | \boldsymbol{\theta}^t\right)} \right)^{-1},$$

em que $\boldsymbol{\theta}^t$ é a amostra de $\prod(\boldsymbol{\theta}|y_i)$ e T indica o número de amostras a *posteriori*.

2.4 Resultados e Discussão

Para todos os modelos ajustados a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. As condições espaciais em que o estudo foi conduzido são levados em consideração nos modelos de diversas formas, a fim de encontrar aquela que melhor explica a variabilidade da variável resposta.

A Tabela 2.1 apresenta o critério de informação da Deviance (DIC), número estimado de parâmetros (NP) e as condicionais preditivas ordenadas (CPO) dos modelos espaciais ajustados que foram calculados para possibilitar uma comparação e consequente escolha do modelo que melhor descreve a resposta. Na notação usada para apresentar os modelos, Y sempre representará a variável resposta e os efeitos espaciais seguem a notação introduzida na seção 2.2.2. Para todos os modelos o intercepto está presente.

O modelo 1 foi ajustado apenas com o intercepto para servir como comparação. Os modelos 2 e 3 levam em consideração o efeito espacial, sendo que o 2 assume a priori que o efeito não tem nenhuma estrutura no espaço, e o 3 assume uma estrutura condicional autoregressiva. O modelo 4 também assume uma estrutura condicional autoregressiva mas com a priori Gamma(0,01;0,01) do hiperparâmetro $\tau_{\varphi i}$. Finalizando, o modelo 5 condensa os modelos 2 e 3, dizendo que os efeitos espaciais podem ser divididos em uma parte estruturada no espaço, e outra parte é não estruturada.

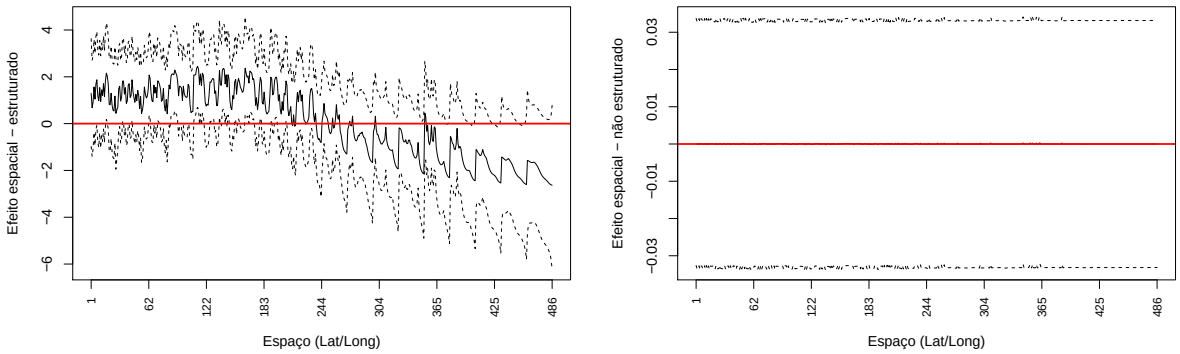
Tabela 2.1. Critério de informação da deviance (DIC), número estimado de parâmetros (NP), ordenada preditiva condicional (CPO) para os diferentes modelos espaciais ajustados.

Modelos	Preditor Linear	Março de 2013			Setembro de 2013			Janeiro de 2014		
		DIC	NP	CPO	DIC	NP	CPO	DIC	NP	CPO
1	$y \sim 1$	583,78	1,00	0,601	664,43	1,00	0,684	391,83	1,00	0,403
2	$y \sim \phi_i$	583,73	1,04	0,601	664,37	1,06	0,684	391,81	1,03	0,403
3	$y \sim \varphi_i$	461,90	64,59	0,470	495,75	88,53	0,505	322,25	45,45	0,326
4	$y \sim \varphi_i{}^*$	462,68	69,95	0,471	496,94	96,78	0,508	324,60	50,87	0,328
5	$y \sim \phi_i + \varphi_i$	461,87	64,63	0,470	496,25	88,61	0,505	322,34	45,47	0,326

Obs: $*\tau_{\varphi_i} \sim \Gamma(0,01; 0,01)$

Ao compararmos os modelos 2 e 3 para todos os meses do estudo, verifica-se que o modelo com efeito espacial estruturado apresenta um melhor ajuste em termos de DIC. Em termos de número de parâmetros e CPO o modelo com efeito espacial não estruturado apresenta o melhor ajuste. Porém, ao analisarmos em termos gerais pode-se verificar que para o mês de Março de 2013 o modelo 5 é o que apresenta menor DIC, o que indica como o melhor entre os modelos ajustados para descrever o comportamento da variável resposta. Ainda em relação ao mês de Março de 2013, o número de parâmetros do modelo 3 é mais parcimonioso que o modelo 4 e em termos de CPO os modelos 3 e 5 são equivalentes.

A Figura 2.2 apresenta um gráfico do ajuste de cada bloco dos parâmetros do modelo 5 para o mês de Março de 2013. O painel A é o ajuste do efeito espacial estruturado φ_i e o painel B é o ajuste do efeito espacial não estruturado ϕ_i.

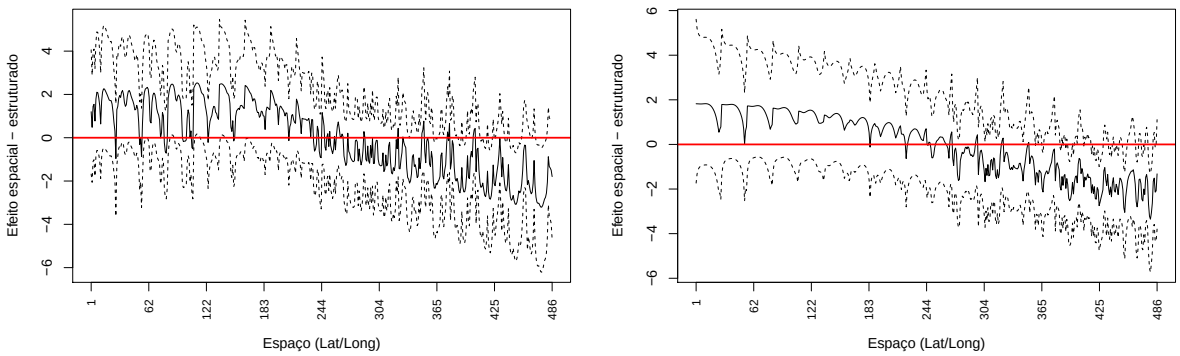


(A) (B)

Figura 2.2. Ajuste de cada bloco de parâmetros do modelo 5 - Março/2013.

A Figura 2.2 mostra claramente que o efeito não estruturados não ajuda a descrever o comportamento da variável resposta, já que, não apresenta diferença entre os pontos no espaço. Nota-se que o efeito estruturado capta toda a variabilidade da variável resposta, mostrando que apenas este efeito é responsável pela qualidade do ajuste. Desta forma, opta-se pela escolha do modelo 3 como o modelo que melhor descreve o padrão espacial da leprose dos citros no mês de Março de 2013.

Ao analisar os meses de Setembro/2013 e Janeiro/2014, observa-se que o modelo 3 é o que apresenta um melhor ajuste em termos de DIC, número de parâmetros e CPO, indicando assim o modelo 3, como o melhor modelo para descrever o padrão espacial da leprose dos citros nesses meses(ver Tabela 2.1). A Figura 2.3 apresenta o gráfico de ajuste do parâmetro estruturado do modelo 3 para o meses de Setembro/2013 e Janeiro/2014. O painel C é o ajuste do efeito espacial estruturado φ_i em Setembro/2013 e o painel D é o ajuste do efeito espacial estruturado φ_i para Janeiro/2014.



(C) (D)

Figura 2.3. Ajuste do parâmetro espacial do modelo 3 - Setembro/2013 e Janeiro/2014.

A Figura 2.3 mostra claramente que o efeito estruturado capta toda a variabilidade da variável resposta, mostrando que este efeito é responsável pela qualidade do ajuste, como observou-se as medidas de ajuste anteriormente. Logo, confirma-se que esse modelo é o mais adequado para descrever o padrão espacial da leprose dos citros nos meses de setembro/2013 e janeiro/2014. Levando em consideração os resultados obtidos, o modelo 3 foi escolhido como o modelo para mapear os riscos de infestação no espaço durante os três meses do estudo, cujas médias e probabilidades a *posteriori* são representadas

graficamente na figura abaixo.

A Figura 2.4 mostra as médias *posteriores* (esquerda) e a probabilidade *posteriori*(direita) do risco de infestação para a leprose dos citros para o modelo 3 (CAR) durante os meses de Março/2013, Setembro/2013 e Janeiro/2014. Nela pode-se observar que em todos os meses avaliados houve algumas plantas não infestadas pela leprose, porém na primeira avaliação realizada em Março de 2013, portanto, antes da aplicação do acaricida, têm-se que as estimativas dos parâmetros para este período diferem das estimativas para os meses de Setembro/13 e Janeiro/14, período em que ocorre segunda e terceira avaliação do talhão, respectivamente. Observa-se, que após a segunda avaliação (Setembro/13) de incidência da leprose, a mesma se disseminou consideravelmente neste talhão, mesmo após a aplicação de acaricida para o combate ao ácaro-vetor. Já na terceira avaliação realizada (Janeiro/2014), nota-se que a leprose não estava presente apenas em uma parte da área.

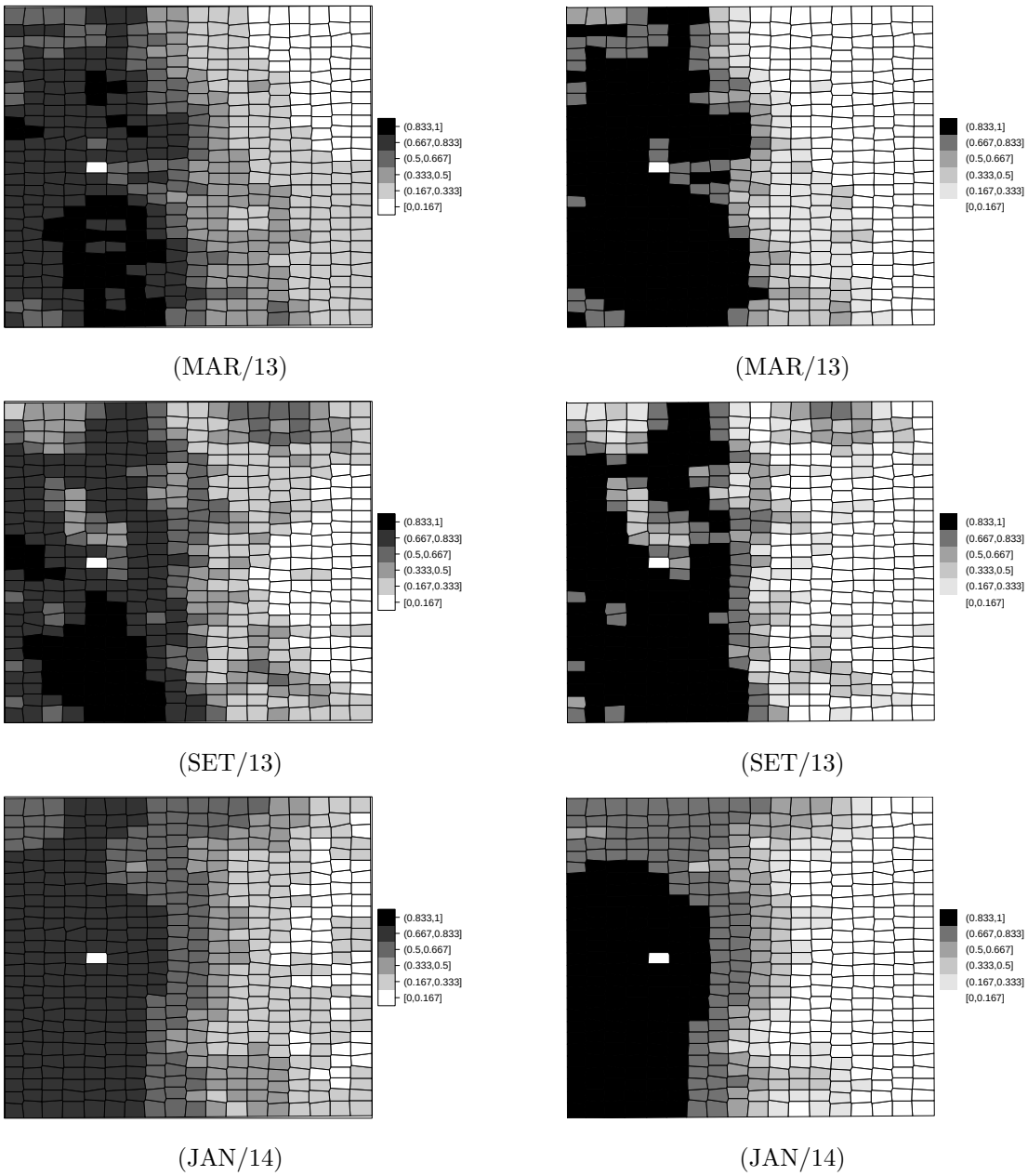


Figura 2.4. Médias *posteriores* (esquerda) e probabilidade *posteriori* (direita) do risco de infestação para a leprose dos citros para o modelo 3 (CAR) durante os meses de Março/2013, Setembro/2013 e Janeiro/2014.

2.5 Conclusão

Para todos os modelos ajustados a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. A partir da análise de comparação dos modelos, notou-se que para os meses em estudo o modelo com efeito espacial estruturado teve melhor desempenho quando comparado ao modelo com efeito espacial estruturado e não estruturado. Ao analisar as médias e probabilidades *posteriores* do risco de infestação para a leprose dos citros para o modelo estruturado espacialmente durante os meses de estudo verificou-se que a aplicação de acaricida não foi suficiente para contenção da leprose dos citros na área experimental e que o progresso da doença foi intenso durante a realização do experimento.

2.6 Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior(CAPES) e a Fundação da Amazônia de Amparo a Estudo e Pesquisa (FAPESPA) pelo apoio financeiro a este estudo a partir do programa de formação doutoral docente(Prodoutoral).

Referências

Amorim, L. C. S., Silva, J. L., Gravena, S., Bevenga, S. R., e Junior, N. A. (2006). Efeito de acaricidas sobre ovos do ácaro-da-leprose dos citros, em diferentes idades. *Citrus R & T*, 27(2):0.

Assunção, R. M. (2001). Estatística espacial com aplicações em epidemiologia, economia, sociologia. *In: Escola de Modelos de Regressão, 7., Minicurso.*

Bassanezi, R. B. e Laranjeira, F. F. (2007). Spatial patterns of leprosis and its mite vector in commercial citrus groves in brazil. *Plant Pathology*, 56:97–106.

Bastianel, M., Novelli, V. M., Kitajima, E. W., Kubo, K. S., Bassanezi, R. B., Machado, M. A., e Freitas-Astúa, J. (2010). Citrus Leprosis: Centennial of an Unusual Mite–Virus Pathosystem. *Plant Disease*, 94(3):284–292.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society (Series B)*, 36:192–236.

Besag, J. E., York, J., e Molli'e, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of Institute of Statistical Mathematics*, 43:1–59.

Bitancourt, A. A. (1955). A leprosi do citrus. *O Biólogo*, 6(1):39–45.

Borges, R. S. e Almeida, F. J. (2000). Câmbios en la producción de plantones cítricos em Brasil. *Todos Citrus*, 3:5–12.

Charest, J., Dewdney, M., Paulitz, T., Philion, V., e Carisse, O. (2002). Spatial Distribution of Venturia inaequalis Airborne Ascospores in Orchards. *Phytopathology®*, 92(7):769–779.

Chen, D. G. e Peace, K. E. (2011). Clinical Trial Data Analysis using R. *Chapman & Hall/CRC Biostatistics Series.*, page 387p.

Dey, D. K., Ghosh, S. K., e Mallick, B. K. (2000). *Generalized linear models: A Bayesian Perspective.* London: Chapman & Hall/CRC, London, 2000.

Farias, P. L., Miranda, V. S., e Ribeiro, S. M. (2008). Spatial analysis of the distribution of Spodoptera frugiperda (J.E. Smith) (Lepidoptera: Noctuidae) and losses in maize crop productivity using geostatistics. *Neotropical Entomology*, 37:321 – 327.

Franciscon, L., Ribeiro Jr., P. J.; Krainski, E. T., Bassanezi, R., e Czermainski, A. B. (2008). Modelo autologístico espaço-temporal com aplicação à análise de padrões espaciais da leprose-dos-citros. *Pesquisa Agropecuária Brasileira.*, 43(12):1677–1682.

Geisser, S. e E., W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gyenge, J. E., T. E. V. e Edelstein, J. D. (1999). Diseño de planes de muestreo con niveles fijos de precisión del pulgón manchado de la alfalfa, Therioaphis trifolii Monell (Homoptera: Aphididae) en alfalfa (Medicago sativa L.). *Anais da Sociedade Entomológica do Brasil*, 28:729 – 737.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.

Koppen, W. (1948). Climatologia: com un estudio de lós climas de la tierra. *Fundo de Cultura Econômica*, page 478p.

Lima, R. R., Demétrio, C. G. B., Ribeiro Jr., P. J., e Ridout, M. (2006). Uma comparação de técnicas baseadas em quadrats para caracterização de padrões espaciais em doenças de plantas. *Revista Brasileira de Biometria*, 24:7–26.

Locali, E. C., Freitas-Astua, J., de Souza, A. A., Takita, M. A., Astua-Monge, G., Antonioli, R., Kitajima, E. W., e Machado, M. A. (2003). Development of a Molecular Tool for the Diagnosis of Leprosis, a Major Threat to Citrus Production in the Americas. *Plant Disease*, 87(11):1317–1321.

Martins, T. G., Simpson, D., Lindgren, F., e Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society Series A*, 135:370–384.

Rodrigues, J. C. V., Antony, L. M., Salaroli, R. B., e Kitajima, E. W. (2008). Brevipalpus-associated viruses in the central Amazon Basin. *Tropical Plant Pathology*, 33:12–19.

Rue, H. e Held, L. E. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: Chapman & Hall/CRC, London, 2005.

Rue, H., Martino, S., e Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Silva, M. Z., Sato, M. E., Oliveira, C. A., e Leite, D. (2012). Diversidade e dinâmica populacional de ácaros em pomar cítrico. *Bragantia*, 71:210–218.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., e Linde, A. V. D. (2000). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4):583–639.

3 IMPUTAÇÃO MÚLTIPLA VIA MÉTODO DA MÉDIA PREDITIVA: UMA COMPARAÇÃO ENTRE O ALGORITMO MICE E O MÉTODO INLA.

Resumo

Tanto no aspecto espacial quanto no espaço-temporal, perdas de observações podem ocorrer por diversos fatores. Técnicas que possibilitem tratar tais tipos de dados foram propostas, dentre elas a imputação. O objetivo deste trabalho é realizar um estudo de comparação entre o algoritmo MICE e a abordagem INLA utilizando o método de imputação da média preditiva no uso de dados binários. Para tal foram feitas simulações com diferentes tamanhos amostrais e percentuais de perdas e a qualidade do método de imputação é avaliada por meio da raiz normalizada do erro quadrático médio (NRMSE). Finalmente, é realizada uma aplicação em um experimento de leprose do citros, realizado no município de Taquaral-São Paulo e a construção de modelos hierárquicos bayesianos com estrutura gaussiana latente para representar os efeitos espaciais, que será utilizada para contribuir com a metodologia que envolve o problema de dados incompletos.

Palavras - chave: Imputação Múltipla; Média Preditiva; MICE; INLA.

Abstract

Both in the spatial and in the spatiotemporal aspect, loss of observations can occur due to several factors. Techniques that make it possible to handle such types of data have been proposed, including imputation. The aim of this work is to carry out a comparison study between the MICE algorithm and the INLA approach using the predictive mean imputation method for binary data. For this purpose, simulations were carried out with different sample sizes and percentages of losses and the quality of the imputation method is evaluated through the normalized root mean square error (NRMSE). Finally, an application is carried out in a citrus leprosis experiment, carried out in the city of Taquaral-São Paulo and the construction of hierarchical Bayesian models with latent Gaussian structure to represent the spatial effects, which will be used to contribute to the methodology involving the Incomplete data problem.

Keywords: Multiple Imputation; Predictive Average; MICE; INLA.

3.1 Introdução

Observações ausentes são comuns em conjuntos de dados da vida real. Os valores podem não ser registrados por vários motivos. Por exemplo, os dados ambientais muitas vezes não são registrados devido a falhas nos dispositivos ou instrumentos de medição. Em ensaios clínicos, os tempos de sobrevivência e outras covariáeis podem estar ausentes devido ao abandono do paciente. 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.

A literatura sobre valores ausentes é ampla. Para mencionar alguns, Little e Rubin (2002) fornece uma boa visão geral dos valores ausentes e imputação. Em geral, os valores ausentes raramente podem ser ignorados. Modelos que incluem uma maneira de contabilizar dados ausentes devem ser preferidos a simplesmente ignorar as observações ausentes. Quando os valores ausentes podem ser modelados a partir dos dados observados, os modelos de imputação podem ser usados para fornecer estimativas das observações ausentes. Os modelos podem ser estendidos para incorporar um submodelo para a imputação. Esta técnica trata 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, 2007).

Nas ciências sociais, é quase inevitável que alguns entrevistados se recusem a participar ou a responder a certas perguntas. Em estudos médicos, o desgaste de pacientes é muito comum. A teoria, metodologia e software para lidar com problemas de dados incompletos foram amplamente expandidos e refinados nas últimas décadas. Em situações com dados faltantes, uma abordagem bastante comum é restringir a análise aos sujeitos com dados completos nas variáveis envolvidas. Entretanto, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que são incluídos na análise forem sistematicamente diferentes daqueles que foram excluídos em termos de uma ou mais variáveis.

Como os métodos de análises estatísticas e aplicativos computacionais foram e são desenvolvidos, em sua maioria, para dados completos, mesmo uma pequena quantidade de dados faltantes pode causar problemas nas estimativas (viés, ineficiência), justificando, então, necessidade de ser considerado nas análises o problema de dados faltantes. Os principais pacotes de análise estatística agora possuem recursos para realizar as análises apropriadas. Neste sentido, o presente trabalho tem como objetivo comparar o método de imputação da média preditiva utilizando o algoritmo "*Multivariate Imputation by Chained Equation*- MICE e o método "*Integrated Nested Laplace Approximational*- INLA, que será utilizando para fazer a estimação dos paraâmetros na aplicação de dados espaciais binários. E, assim contribuir com a metodologia para uma melhor compreensão das questões envolvidas com problemas de dados incompletos para fornecer e subsidiar um melhor resultado quanto à utilização desse tipo de técnica.

3.2 Metodologia

3.2.1 Imputação Múltipla

A imputação de dados faltantes tem sido uma estratégia comum para a análise de dados com esse problema. Entende-se por imputação a técnica de preencher os dados faltantes com valores plausíveis. Um atrativo para a utilização de técnicas de imputação é o fato de, após a imputação dos dados, o investigador poder utilizar técnicas tradicionais de análise estatística para dados completos. Métodos simples como imputação pela média ou pela mediana, também conhecidos como métodos de imputação única, têm sido bastante usados devido à sua facilidade de implementação. Entretanto, existem desvantagens na utilização desses métodos, tais como a subestimação da variabilidade da variável imputada que gerará intervalos de confiança mais estreitos do que o esperado e a impossibilidade de levar em consideração a variabilidade que possa existir entre diferentes imputações.

Com o intuito de resolver problemas relacionados a dados faltantes, alguns métodos mais elaborados têm sido desenvolvidos. O objetivo desses métodos é preencher os dados faltantes, tornando possível a realização da análise com a base de dados completa, isto é, com todos os indivíduos e variáveis. Como alternativa à imputação única e com o objetivo de corrigir suas desvantagens, surgiu na década de 80 a imputação múltipla (IM) proposta por Donald Rubin. Porém, apenas recentemente esta técnica vem sendo mais utilizada devido aos desenvolvimentos computacionais para a implementação da técnica. Essa técnica possibilita, além da estimativa pontual dos parâmetros, a inclusão da incerteza da imputação dos dados na variância dos resultados estimados, corrigindo o maior problema associados à imputação única (Schafer, 1999; Rubin, 1987).

A idéia da imputação múltipla é a de que cada dado ausente é imputado m vezes, gerando m bancos de dados completos. Os m bancos são analisados separadamente por uma técnica tradicional de análise estatística e finalmente os m resultados obtidos são combinados de maneira simples para a análise final. A primeira parte é a mais fundamental da IM, pois as técnicas de imputação utilizadas têm que preservar a relação das observações faltantes e observadas, levar em conta o mecanismo de ausência e o

padrão dos dados faltantes. Na seção a seguir trataremos das abordagens que serão utilizadas no estudo do método da imputação da múltipla.

3.2.1.1 Multivariate Imputation by Chained Equation - MICE

Segundo Van Buuren e Groothuis-Oudshoorn (2011), MICE é um método de cadeia de Markov Monte Carlo (MCMC), onde o espaço de estado é o conjunto de todos os valores imputados. Mais especificamente, se as condicionais são compatíveis, o algoritmo MICE é um amostrador de Gibbs, uma técnica de simulação Bayesiana. Implementado no *software* R, o pacote mice, é de tal forma que o usuário pode especificar um método de imputação para cada coluna de dados incompletos. O método de imputação leva um conjunto de preditores completos, e retorna uma única imputação para cada entrada em falta na coluna incompleta. Várias imputações são criadas por chamadas repetidas para a função (Van Buuren e Groothuis-Oudshoorn, 2000). Uma vantagem da abordagem variável por variável é que, para cada uma das variáveis, um modelo de imputação diferente pode ser utilizado. Portanto, um conjunto de dados pode ter variáveis tanto contínuas, binárias ou categóricas (Oudshoorn et al., 1999).

O algoritmo MICE é uma maneira conceitualmente simples, flexível e prática para gerar imputações múltiplas. Para cada variável incompleta o usuário pode escolher um conjunto de preditores que serão utilizados para a imputação. Isto é útil para a imputação de grandes conjuntos de dados (Pirdawd, 2007). A idéia por trás do pacote MICE é a seguinte: começa com um sorteio a partir dos dados observados, e imputa os dados incompletos variável por variável. Uma iteração consiste de um ciclo passando por todo Y_j. O número de iterações T frequentemente é baixo, 5 ou 10. As observações faltantes são substituídas por m conjuntos de valores simulados independentes, extraídos da distribuição preditiva posteriori das observações faltantes condicionado aos dados observados. Os passos do algoritmo MICE para a imputação são:

1. Especifica um modelo de imputação $P(Y_j^{imp}|Y_j^{obs}, Y_{-j}, R)$ para a variável Y_j com $j = 1, ..., p$;

2. Para cada j começa o preenchimento das imputações por Y_j^0 por aleatórias retiradas de Y_j^{obs};

3. Repete para $t = 1, ..., T$;

4. Repete para $j = 1, ..., p$;

5. Define $Y_{-j}^t = (Y_1^t, ..., Y_{j-1}^t, Y_{j+1}^{t-1}, ..., Y_p^{t-1})$ como os dados atualmente completos exceto Y_j;

6.Retira $\phi_j^t \sim P(\phi_j^t|Y_j obs, Y_{-j}^t, R)$;

7. Retira as imputações $Y_j^t \sim P(Y_j^{imp}|Y_j^{obs}, Y_{-j}^t, R, \phi_j^t)$;

8. Repete j;

9. E repete t.

Em cada iteração, um modelo especificado é ajustado para cada variável com informações faltantes usando observações para essa variável, que podem incluir valores imputados para outras variáveis. Esse modelo resultante é usado para imputar valores faltantes da variável imputada.

3.2.1.2 Integrated Nested Laplace Approximation - INLA

Segundo Fong et al. (2010), no caso de pequenas amostras, a estimação dos parâmetros como é usualmente feita, na abordagem clássica, pode apresentar problemas, especialmente na estimação dos componentes de variância, pois a inferência baseia-se na distribuição assintótica dos estimadores de máxima verossimilhança. È comum propor para solucionar esses problemas a utilização da abordagem bayesiana, a qual é baseada nos dados observados e permite a atualização de informação. Todavia, essa abordagem tem sido prejudicada pela necessidade de programação computacional intensiva. Dada a necessidade de ajuste de modelos hierárquicos, além dos modelos lineares generalizados mistos, de maneira eficiente e rápida, Rue et al. (2009) propuseram a estimação INLA, do inglês, "Integrated Nested Laplace Approximation". O método INLA é uma aproximação determinística para a inferência bayesiana para a classe do modelos gaussianos latentes. Nesses modelos, como descrito por Rue et al. (2009), o campo latente é gaussiano, controlado por hiperparâmetros e com variáveis respostas não gaussianas, de modo que as distribuições marginais a posteriori não apresentam forma fechada.

Segundo Rue et al. (2009), a definição dos modelos gaussianos latentes (MGL) é dada como sendo modelos que assumem a forma aditiva apresentada em 3.1 e assume-se *priori* gaussiana para todos os parâmetros que compõe η, o vetor desses parâmetros é chamado de campo aleatório gaussiano (CAG). Neste trabalho será considerada apenas aplicações em modelos lineares generalizados. Segundo Dey et al. (2000), o preditor linear dos modelos lineares generalizados pode ser expresso por,

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k \boldsymbol{z}_{ki}. \tag{3.1}$$

Segundo Martins et al. (2013), o modelo descrito em 3.1 pode ser escrito utilizando uma estrutura hierárquica, em que o primeiro estágio é formado pela função de verossimilhança com propriedades de independência condicional, dada pelo campo latente gaussiano $\boldsymbol{x} = (\boldsymbol{\eta}, \alpha, \boldsymbol{\beta})$ e possíveis hiperparâmetros $\boldsymbol{\theta}$ (associados aos componentes de variância). Assim, o primeiro estágio da hierarquia é dado por,

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) = L(\boldsymbol{x}, \boldsymbol{\theta}_1|\boldsymbol{y}) = \prod_{i=1}^{n_d} \pi(y_i|x_i, \boldsymbol{\theta}_1).$$

A distribuição condicional do campo latente $\boldsymbol{x}$, dados alguns possíveis hiperparâmetros $\boldsymbol{\theta}_2$ forma o segundo estágio do modelo e tem uma distribuição Gaussiana conjunta.

$$\boldsymbol{x}|\boldsymbol{\theta}_2, \boldsymbol{y} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}_2, \boldsymbol{y}) = N(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)),$$

em que $N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ denota a distribuição gaussiana multivariada com vetor de média $\boldsymbol{\mu}$ e matriz de precisão $\boldsymbol{Q}$. Na maioria das aplicações, os campos Gaussianos latentes possuem propriedades de independência condicional, o que se traduz em uma matriz de precisão esparsa $\boldsymbol{Q}(\boldsymbol{\theta}_2)$, que é de extrema importância para reduzir o custo computacional. Uma distribuição Gaussiana multivariada com uma matriz de precisão esparsa é conhecida como um campo aleatório de Markov de Gauss (GMRF) (Rue e Held, 2005).

O terceiro e último estágio do modelo hierárquico é dado pela especificação de uma distribuição a *priori* apropriada para cada um dos hiperparâmetros do modelo,

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

Para o modelo descrito anteriormente, a distribuição conjunta a *posteriori* para os parâmetros do modelo linear generalizado, pode ser expressa por,

$$\begin{aligned}\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i=1}^{n_d}\pi(y_i|x_i,\boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{n/2}\exp\left[-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}+\sum_{i=1}^{n_d}\log\pi(y_i|x_i,\boldsymbol{\theta})\right]. \qquad (3.2)\end{aligned}$$

Segundo Rue et al. (2009), as distribuições marginais a *posteriori* de interesse podem ser expressas por,

$$\begin{aligned}\pi(x_i|\boldsymbol{y}) &= \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta}, \\ \pi(\theta_j|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\,\boldsymbol{\theta}_{-j},\end{aligned}$$

em que $i = 1,\ldots,I$, sendo I o número de parâmetros de efeito fixo e efeito aleatório do modelo, e $j = 1,\ldots,J$, sendo J o número de componentes de variância. Rue et al. (2009) utilizam as equações anteriores para construir as aproximações aninhadas,

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta}, \qquad (3.3)$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta}_{-j}. \qquad (3.4)$$

Para as equações (3.3) e (3.4), Rue et al. (2009) sugerem dividir o método de aproximação INLA em três etapas; Primeiramente propor uma aproximação $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ para a distribuição conjunta a *posteriori* dos hiperparâmetros $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. O segundo passo consiste em propor uma aproximação $\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ para as marginais da distribuição condicional do campo latente dado as observações e os hiperparâmetros $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$, e finalmente explorar $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ em uma grade de valores e usá-los para integrar $\boldsymbol{\theta}$ na Equação (3.3) e $\boldsymbol{\theta}_{-j}$ na Equação (3.4).

Modelos que incluem uma maneira de contabilizar dados ausentes devem ser preferidos a simplesmente ignorar as observações ausentes. Quando os valores ausentes podem ser modelados a partir dos dados observados, os modelos de imputação podem ser usados para fornecer estimativas das observações ausentes. Os modelos podem ser estendidos para incorporar um submodelo para a imputação.

O INLA pode ajustar modelos com observações ausentes na resposta computando sua distribuição preditiva. Dado que a distribuição da variável resposta faz parte do modelo, é possível prever os valores perdidos calculando sua distribuição preditiva. Dada uma resposta ausente sua distribuição preditiva é:

$$\pi(y_m|\mathbf{y}_{obs}) = \int \pi(y_m,\boldsymbol{\theta}|\mathbf{y_{obs}})\,d\boldsymbol{\theta} = \int \pi(y_m|\mathbf{y}_{obs},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_{obs})\,d\boldsymbol{\theta}. \qquad (3.5)$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo usado para estimar os dados ausentes y_m.

Assim, as imputações podem ser criadas mostrando-se primeiro da densidade *a posteriori* de θ, para obter θ^* e depois mostrando da distribuição preditiva $\pi(y_m,\boldsymbol{\theta}^* \mid \mathbf{y}_{obs})$.

A partir das m imputações realizadas, usando-se métodos padrões fornecem estimativas diferentes de θ, ou seja, θ_i para $i = 1,\cdots,m$. Segundo Schafer (1999), θ pode ser qualquer medida escalar a ser estimada, tal como média, correlação, coeficiente de regressão ou razão de chances. O passo seguinte é combinar essas estimativas para obter inferências finais que reflitam apropriadamente a incerteza associada aos dados ausentes, sob o modelo especificado. A estimativa combinada para cada θ é dada por

$$\overline{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \tag{3.6}$$

Para a variância combinada, calcula-se a variância dentro das imputações: $\overline{U} = \frac{1}{m}\sum_{i=1}^{m} U_i$ e a variância entre as imputações: $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \overline{\theta})$. Então, a variância total, que é a variância combinada, será:

$$T = \overline{U} + \frac{m+1}{m}B, \tag{3.7}$$

em que $(m+1)/m$ é uma correção pelo fato de m ser finito.

O INLA será utilizado para calcular automaticamente as distribuições preditivas para todos os valores ausentes na resposta, que deve receber um *NA* valor ao definir os dados.

3.2.1.3 Método da Média Preditiva (PMM – Predictive Mean Meating)

O *PMM* é um método de imputação múltipla desenvolvida por Little (1988), tendo como objetivo imputar valores ausentes por meio do método vizinho mais próximo(método do tipo *hot-deck*, com a distância baseada nos valores esperados das variáveis ausentes condicional às covariáveis observadas, combinando elementos de regressão linear. A idéia básica do método de média preditiva é usar diferentes modelos de regressão para calcular a observação faltante. Cada conjunto de estimativas dos modelos será usado para gerar um único conjunto de dados imputados Heitjan e Little (1991). As estimativas dos parâmetros devem ser extraídas da distribuição posteriori dos dados.

Assume-se Y como uma variável com valores faltantes, na qual deseja-se imputar a partir de outras variáveis completas. Seja X_k, com k=1,K,l, com sendo l variáveis completas, em que $\mathbf{X}$=(X_1, K, X_k). Define-se por Y_i os dados observados, Y_{mis} os dados faltantes e n_{obs} o número de observações com dados observados de Y.

A imputação múltipla dos Y_{mis} por meio de PMM é realizado pelo seguinte algoritmo Little (1988); Vink (2014):

1. Usa-se a regressão linear de $Y|X, \beta$. Então tem-se:

$$Y \sim N(X\beta, \sigma^2) \tag{3.8}$$

Assim, a especificação para $f(Y_i/X_i, \theta)$, $\theta = (\beta, log\sigma)$, β um vetor de q componentes, onde q é o número de preditores, e σ um escalar. Admitindo-se uma distribuição a priori não-informativa para θ, $P(\theta) \propto 1$, para evitar grandes complexidades assume-se $n_1 > q$, onde n_1 é o número de respondentes. Assim de acordo com Rubin (1987) a distribuição a posteriori de θ envolve apenas os Y_i observados. De modo que,

$$\hat{\sigma}_1{}^2 = \frac{\sum_{obs}(Y_i - X_i\hat{\beta}_1)^2}{n_1 - q}, \tag{3.9}$$

sendo:

$$\hat{\beta}_1 = V\left[\sum_{obs} X_i^T Y_i\right], \tag{3.10}$$

e

$$V = \left[\sum_{obs} X_i^T X_i\right], \tag{3.11}$$

Logo a distribuição *à posteriori* de θ descrita em termos de distribuições padrão pode-se estimar os parâmetros a serem usados na imputação. 2. Simular uma variável aleatória σ_*^2 qui quadrado, χ_{n_1-q}, g e seja:

$$\hat{\sigma}_*^2 = \frac{\sigma_1^2(n_1 - q)}{g}. \tag{3.12}$$

3. Simular q variáveis independentes$N(0,1)$ para criar um vetor Z de q componentes e seja:

$$\beta^* = \hat{\beta}_1 + \sigma_*[V]^{1/2}Z, \tag{3.13}$$

em que $[V]^1/2$ é a raiz quadrada de V tal como a raiz quadrada triangular obtida pela fatoração de Cholesky.

4. Simular os n_0 valores dos Y_{mis} como:

$$Y_i^* = X_i\beta^*. \tag{3.14}$$

onde i $\in$ dados em falta e n_0 desvios normais.

5. Para cada Y_i^*, i $\in$ faltantes, encontra-se o respondente cujo Y_i(i $\in$ observados) de modo que esteja o mais próximo de Y_i^*. O valor de Y_i será usado para a próxima imputação.

6. Repita os passos 1 a 5 m vezes, cada vez é salvo o conjunto de dados concluído.

O ajuste do modelo é realizado pelo INLA, com imputação MICE e imputação INLA utilizando o método PMM, ambos implementados no *software* R. Em seguida, será realizada uma comparação entre essas abordagens para identificar qual melhor ajuste do modelo adotado.

3.2.2 Validação do Modelo

O "normalized root mean squared error"(NRMSE) será a abordagem aplicada para avaliar a precisão das previsões feitas pelas imputações e pelo modelo binomial. O NRMSE é comparado por meio das médias ajustadas, ou seja, os valores imputados são comparados com os correspondentes valores observados no conjunto dos dados originais e é considerado como melhor método ou método com melhor desempenho, aquele que apresentar o menor valor da estatística NRMSE.

$$NRMSE = \sqrt{\frac{media(x_{imp} - x_{obs})^2}{media(x_{obs})}} \tag{3.15}$$

em que x_{imp} e x_{obs} são vetores contendo os respectivos valores médios ajustados (imputados) e os valores verdadeiros das observações ausentes simuladas e média(x_{obs}) é a média dos valores contidos no vetor x_{obs}. Quanto menor for o valor da estatística NRMSE, melhor será o método de imputação.

3.2.3 Estudo de Simulação

O principal objetivo do estudo de simulação é comparar, por meio do NRMSE, os valores simulados com os resultados obtidos pela imputação MICE e imputação INLA nos diferentes tamanhos amostrais e perdas aleatórias. Como área de estudo utilizou-se um grid regular com 100 amostras de tamanho 100, 400, 900 e 2500 e, percentuais com perdas aleatórias de 10%, 20%, 30% e 40%. Os dados

serão simulados seguindo o modelo $Y(x)|S \sim Binomial(n, p(x))$ com n = 1 e $p(x) = exp(\mu + S(x))/[1 + exp(\mu + S(x))]$, onde $S(x)$ é um processo gaussiano com média $\mu = 0$ e funcão de correlação Matern com $\phi = 0.3$ e $\sigma^2 = 4$. Para criar as imputações é utilizado o método PMM a partir do algoritmo MICE e do método INLA, o que possibilitará fazer previsões aos pontos retirados e utilizá-lo como referência.

A Figura 3.1 apresenta a média da estatística NRMSE para os diferentes tamanhos amostrais e percentuais de perdas utilizando imputação MICE e INLA. Nela pode-se observar, que para os tamanhos de amostras 100, 900 e 2500 a média da estatística NRMSE na imputação INLA diminui a medida que aumenta-se o percentual de retiradas. Porém, quando o tamanho da amostra é 400 nota-se que a medida que aumenta-se o percentual de retirada há um aumento na média dessa estatística. Ao analisar a imputação MICE, observa-se que para o tamanho de amostra 100 a média da estatística NRMSE é bem baixa e permanece constante a medida que aumenta-se os percentuais de retiradas dessa amostra. Já para as amostras de tamanho 400, 900 e 2500 percebe-se que a média da NRMSE da imputação MICE vai aumentando a medida que o tamanho da amostra e percentuais de retiradas aumentam.

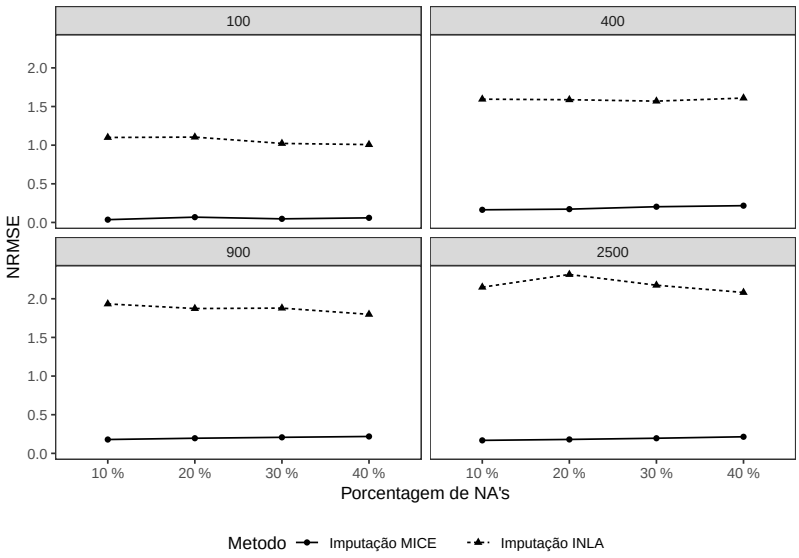


Figura 3.1. Média da estatística NRMSE para os diferentes tamanhos amostrais e percentuais de perdas utilizando imputação INLA e imputação MICE.

A Figura 3.2 apresenta o boxplot da estatística NRMSE para os diferentes tamanhos amostrais e percentuais de perdas utilizando imputação INLA e MICE. Nela observa-se que os valores da estatística NRMSE para imputação INLA são baixos, porém a medida que aumentam-se o tamanho amostral e o os percentuais de perdas há um aumento na variabilidade e também nos valores medianos e médios da estatística NRMSE. Já para imputação MICE têm-se valores altos da NRMSE porém os valores médios, medianos e a variabilidade vão diminuindo quando aumentam-se os percentuais de perdas e tamanhos da amostra.

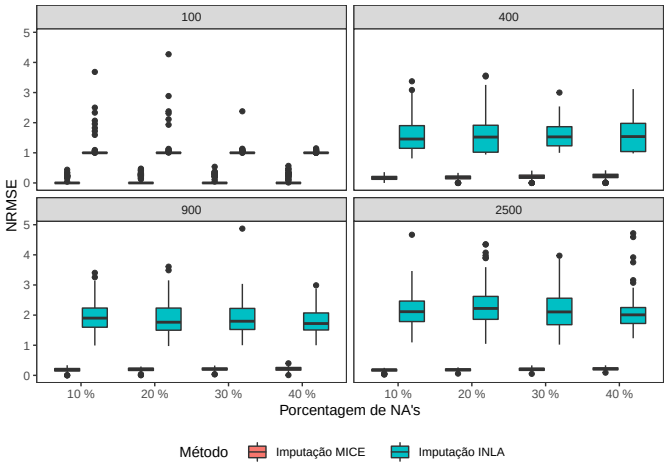


Figura 3.2. Boxplot da estatística NRMSE para os diferentes tamanhos amostrais e percentuais de perdas utilizando imputação INLA e imputação MICE.

3.3 Aplicação em dados de leprose do citros

3.3.1 Área de Estudo

A área de estudo localiza-se no município de Taquaral, no estado de São Paulo, Brasil. O clima da região segundo a classificação de Koppen (1948) é do tipo Aw, com verão quente e úmido e inverno com estação seca. A precipitação pluvial média anual é de 1.629 mm, concentrada no período de setembro a março. A temperatura média anual é de 23 °C. Os experimentos foram instalados em duas áreas com incidência de leprose dos citros e ácaros *Brevipalpus*, onde todas as plantas foram georreferenciadas. A área desse estudo, continha inicialmente 2.686 plantas de laranja "Valência"(*C.sinensis*) enxertadas sobre limoeiro "Volkamericano"(*Citrus volkamericano* Tem & Pasq.), com espaçamento de 6,0 x 2,5 metros e com 4 anos de idade. O talhão selecionado possuía área total de 38.610 m^2 e encontra-se margeado ao norte e ao sul por cerca-viva *Syzygium jambolanum* (Lam.), ao leste por cultura de cana-de-açúcar *Saccharum officinarum* L. e a oeste por pomar cítrico convencional. O talhão também apresenta aplicações de produtos fitossanitários. Após a seleção deste talhão todas as plantas foram georreferenciadas e o experimento foi instalado em julho de 2013, porém para este estudo utiliza-se apenas o mês de Dezembro de 2014 que é o mês que ocorreu o estudo na área 2 onde se teve células em branco (NA). O número mínimo de amostras foi definido usando a variabilidade inferida pelos coeficientes de variação (Chen e Peace, 2011). E para os dados em braco utiliza-se a imputação MICE.

3.3.2 Resultados e Discussão

Foram ajustados cinco modelos e para todos estes, a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. O modelo nulo (apenas com intercepto), um modelo que considera que não existe nenhum tipo de relação espacial entre as observações, outro que considera que existe relação espacial entre as observações e por ultimo um modelo com os dois componentes espaciais (não estruturado e estruturado). As condições espaciais em que o estudo foi conduzido são levados em consideração nos modelos de diversas formas, a fim de encontrar aquela que melhor explica a variabilidade da variável resposta. Para comparação destes modelos e consequente escolha do modelo que melhor descreve a resposta, foram utilizados três critérios em conjuntos: o critério de informação deviance (DIC), o número efetivo de parâmetros e a ordenada preditiva condicional (*Conditional Predictive Ordinate*, CPO).

3.3.2.1 Critério de informação deviance (DIC)

Para o cálculo do DIC considera-se a medida *deviance* dada por $D = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, em que $\boldsymbol{y}$ é o vetor de dados, $\boldsymbol{\theta}$ é o vetor de parâmetros de interesse sob determinado modelo e $p(\boldsymbol{y}|\boldsymbol{\theta})$ a função de verossimilhança. O DIC é calculada como $DIC = \hat{D}+2p_D$, em que $\hat{D} = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D]$ é uma medida que indica a qualidade do ajuste, e p_D o número efetivo de parâmetros, representando a complexidade do modelo. Segundo Spiegelhalter et al. (2000), a quantidade efetiva de parâmetros do modelo é calculado como a média a *posteriori* da *deviance* menos a *deviance* avaliada na média a *posteriori*, ou seja, $p_D = \hat{D} - D(\hat{\boldsymbol{\theta}})$.

A idéia é que os modelos com menor DIC devem ser preferidos a modelos com maiores valores para essa estatística. Os modelos são penalizados tanto pelo valor de $\hat{D}$, o que favorece um bom ajuste, como também pelo número efetivo de parâmetros p_D. É importante ressaltar que para o cálculo do DIC assume-se que a média a *posteriori* é uma boa estimativa pontual para os parâmetros e, portanto, não representa uma boa medida para a avaliação do ajuste na presença de distribuições a *posteriori* assimétricas para os parâmetros do modelo.

3.3.2.2 Condicionais preditivas ordenadas (CPO)

A estatística CPO é um método de avaliação cruzada utilizado para avaliar a performance do modelo medindo o seu poder preditivo (Geisser e E., 1979). A distribuição preditiva a *posteriori* para a i-ésima observação é dada por

$$p(\boldsymbol{y}_i|\boldsymbol{y}_{-i}) = \int p(\boldsymbol{y}_i|\boldsymbol{\theta}) \prod(\boldsymbol{\theta}|\boldsymbol{y}_{-i})\, d\boldsymbol{\theta},$$

em que $\boldsymbol{y}_i$ é a i-ésima observação de $\boldsymbol{y}$ e $\boldsymbol{y}_{-i}$ é $\boldsymbol{y}$ sem a observação i. Então, o CPO é a probabilidade a *posteriori* de observar y_i quando o modelo é ajustado sem y_i. Quanto maior o valor da estatística CPO melhor é o ajuste do modelo aos dados. Ao invés de ajustar o modelo a cada vez que é retirada uma observação, uma aproximação do CPO pode ser obtida por amostragem da distribuição a *posteriori*:

$$CPO_i = \left(\frac{1}{T \sum_{t=1}^{T} p\left(\boldsymbol{y}_i|\boldsymbol{\theta}^t\right)} \right)^{-1},$$

em que $\boldsymbol{\theta}^t$ é a amostra de $\prod(\boldsymbol{\theta}|y_i)$ e T indica o número de amostras a *posteriori*.

A Tabela 3.1 apresenta o critério de informação da Deviance (DIC), número estimado de parâmetros (NP) e as condicionais preditivas ordenadas (CPO) dos modelos espaciais ajustados que foram calculados para possibilitar uma comparação e consequente escolha do modelo que melhor descreve a resposta. Na notação usada para apresentar os modelos, Y sempre representará a variável resposta e os efeitos espaciais seguem a notação: φ_i - efeito espacial estruturado e ϕ_i - efeito espacial não estruturado. Para todos os modelos o intercepto está presente.

O modelo 1 foi ajustado apenas com o intercepto para servir como comparação. Os modelos 2 e 3 levam em consideração o efeito espacial, sendo que o 2 assume a priori que o efeito não tem nenhuma estrutura no espaço, e o 3 assume uma estrutura condicional autorregressiva. O modelo 4 também assume uma estrutura condicional autoregresiva mas com a priori Gamma(0,01;0,01) do hiperparâmetro τ_{φ_i}. Finalizando, o modelo 5 condensa os modelos 2 e 3, dizendo que os efeitos espaciais podem ser divididos em uma parte estruturada no espaço, e outra parte é não estruturada.

Ao compararmos os modelos 2 e 3, verifica-se que o modelo com efeito espacial estruturado apresenta um melhor ajuste em termos de DIC. Em termos de número de parâmetros e CPO o modelo com efeito espacial não estruturado apresenta o melhor ajuste. Porém, ao analisarmos em termos gerais

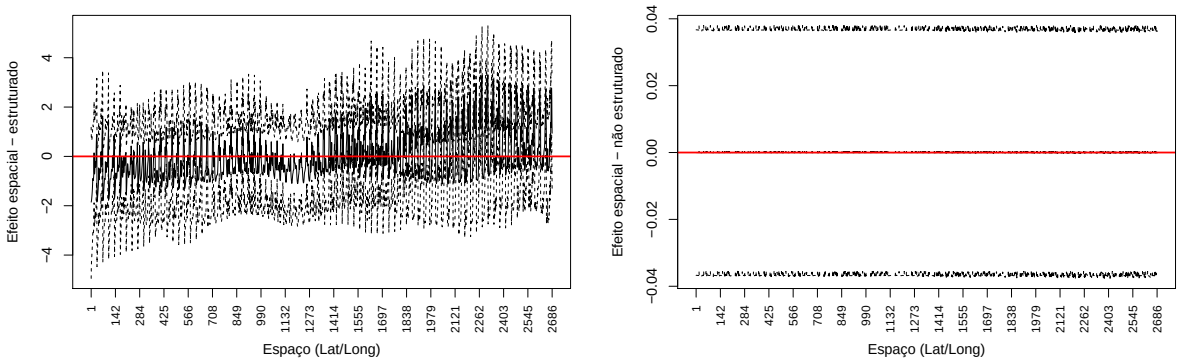
Tabela 3.1. Critério de informação da deviance (DIC), número estimado de parâmetros (NP), ordenada preditiva condicional (CPO) para os diferentes modelos espaciais ajustados.

Modelos	Preditor Linear	Dezembro de 2014		
		DIC	NP	CPO
1	$y \sim 1$	2535,902	1,00	0,4721
2	$y \sim \phi_i$	2535,730	1,19	0,4721
3	$y \sim \varphi_i$	2167,784	223,82	0,3998
4	$y \sim \varphi_i$*	2169,182	231,91	0,3999
5	$y \sim \phi_i + \varphi_i$	2166,674	224,03	0,3998

Obs: * $\tau_{\varphi_i} \sim \Gamma(0,01; 0,01)$

pode-se verificar que o modelo 5 é o que apresenta menor DIC, o que indica como o melhor entre os modelos ajustados para descrever o comportamento da variável resposta. Porém, olhando o ajuste do efeito não estruturado verifica-se que este não ajuda a explicar a variável resposta, ou seja, não apresenta significância. Em relação ao número de parâmetros o modelo 3 é mais parcimonioso que o 4 e em termos de CPO os modelos 3 e 5 são equivalentes.

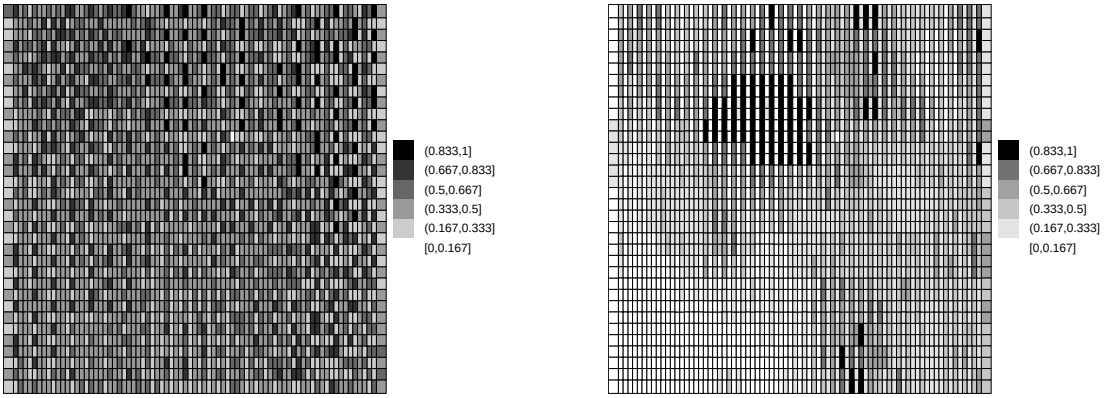
A Figura 3.3 apresenta um gráfico do ajuste de cada bloco dos parâmetros do modelo 5 para o mês de Dezembro/2014. O painel A é o ajuste do efeito espacial estruturado φ_i e o painel B é o ajuste do efeito espacial não estruturado ϕ_i. Esta mostra claramente que o efeito não estruturados não ajuda a descrever o comportamento da variável resposta, já que, não apresenta diferença entre os pontos no espaço. Nota-se que o efeito estruturado capta toda a variabilidade da variável resposta, mostrando que apenas este efeito é responsável pela qualidade do ajuste. Desta forma, opta-se pela escolha do modelo 3 como o modelo que melhor descreve o padrão espacial da leprose dos citros no mês de Dezembro de 2014.



(A) (B)

Figura 3.3. Ajuste de cada bloco de parâmetros do modelo 5 - Dezembro/2014.

A Figura 3.4 mostra as médias *posteriores* (A) e a probabilidade *posteriori*(B) do risco de infestação para a leprose dos citros para o modelo 3 (CAR) durante o mês de Dezembro de 2014. Nela pode-se observar que no mês do estudo há um aumento expressivo no risco de infestação da leprose, mesmo já ocorrendo o controle químico do ácaro meses antes. Porém, ao analisarmos a probabilidade do risco, observa-se que há uma diminuição na proliferação da doença, porém constata-se que existe uma probabilidade de mais de 20% dessa área ainda apresentar plantas com sintomas da leprose do citros.



(A) (B)

Figura 3.4. Médias *posteriores* (A) e probabilidade *posteriori* (B) do risco de infestação para a leprose dos citros para o modelo 3 (CAR) durante o mês Dezembro/2014.

3.4 Conclusão

No presente trabalho, foi analisada a imputação de dados pelo método PMM via método INLA e algoritmo MICE. Os valores da estatística NRMSE para estes dois procedimentos permitiram constatar vantagem a favor da imputação MICE pois verificou-se que o procedimento está imputando em média valores muito mais próximos dos valores verdadeiros independente do tamanho amostral e percentuais de perdas simulados. Na aplicação de dados reais foi utilizado o método PMM para criar as imputações e o procedimento utilizado foi imputação MICE, em seguida utiliza-se o INLA para realizar o ajuste dos modelos.

3.5 Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Fundação da Amazônia de Amparo a Estudo e Pesquisa (FAPESPA) pelo apoio financeiro a este estudo a partir do programa de formação doutoral docente (Prodoutoral).

Referências

Chen, D. G. e Peace, K. E. (2011). Clinical Trial Data Analysis using R. *Chapman & Hall/CRC Biostatistics Series.*, page 387p.

Dey, D. K., Ghosh, S. K., e Mallick, B. K. (2000). *Generalized linear models: A Bayesian Perspective.* London: Chapman & Hall/CRC, London, 2000.

Fong, Y., Rue, H., e Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics, Oxford*, 11:397–412.

Geisser, S. e E., W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, 74(365):153–160.

Heitjan, L. F. e Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society*, 40(1):13–29.

Koppen, W. (1948). Climatologia: com un estudio de lós climas de la tierra. *Fundo de Cultura Econômica*, page 478p.

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys Missing-Data Ad ustments in Large Surveys. *Journal of Business & Economic Statistics*, 6:287–296.

Little, R. J. A. e Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley Series in Probability and Statistics. New York: Wiley, 2002.

Martins, T. G., Simpson, D., Lindgren, F., e Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83.

Oudshoorn, C. G. M., Van Buuren, S., e Van Rijckevorsel, J. L. A. (1999). Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey. *TNO Prevention and Health.*

Pirdawd, H. Q. (2007). *Multiple Imputation Method for Missing Data in (RCBD) Factorial Experiments.* Tese (doutorado), University of Sulaimani.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley Series in Probability and Statistics. New York: Wiley, 1987.

Rue, H. e Held, L. E. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: Chapman & Hall/CRC, London, 2005.

Rue, H., Martino, S., e Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., e Linde, A. V. D. (2000). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4):583–639.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–243.

Van Buuren, S. e Groothuis-Oudshoorn, C. G. M. (2000). Multivariate Imputation by Chained Equations: MICE V1.0 User manual. *TNO Prevention and Health.*

Van Buuren, S. e Groothuis-Oudshoorn, C. G. M. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45.

Vink, G. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90.

4 CONSIDERAÇÕES FINAIS

Esta tese apresentou a aplicação de modelos para análise de dados espaciais binários utilizando abordagem bayesiana. Aplicações de modelagem de fenômenos espaciais com dados binários e imputação de dados espaciais não são ferramentas tão abordadas, portanto a idéia foi mostrar uma modelagem útil, mas ainda pouco explorada pela agricultura de precisão.

No primeiro artigo, uma classe de modelos espaciais foi adotada para descrever a distribuição espacial da leprose dos citros coletada em uma fazenda no município de Taquaral/SP. O uso desta classe de modelos na agricultura de precisão ainda é restrito por tratar-se de modelos altamente estruturados que requerem procedimentos especializados para inferência. Adotou-se a metodologia INLA para inferência bayesiana com ajuste e comparação de modelos com diferentes estruturas. Para todos os modelos ajustados a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. As condições espaciais em que o estudo foi conduzido foram levadas em consideração nos modelos de diversas formas, a fim de encontrar aquela que melhor explica a variabilidade da variável resposta. A partir da análise de comparação dos modelos, notou-se que para todos os meses de estudo o modelo com efeito espacial estruturado teve melhor desempenho quando comparado ao modelo com efeito espacial estruturado e não estruturado. Ao analisar o mapa de médias e probabilidades *posteriores* do risco de infestação para a leprose dos citros mostrou que após a segunda avaliação (Setembro/13) de incidência da leprose, a mesma se disseminou consideravelmente neste talhão, mesmo após a aplicação de acaricida para o combate ao ácaro-vetor. Já na terceira avaliação realizada (Janeiro/2014), nota-se que a leprose não estava presente apenas em uma parte da área. Neste sentido, estes resultados apontaram a necessidade de integração de táticas de controle no manejo da leprose, e não somente o controle químico do ácaro com acaricidas.

No segundo artigo, têm-se um estudo para comparar imputação MICE e INLA na utilização do método PMM. Um estudo de simulação foi realizado para comparar os diferentes métodos e posteriormente uma aplicação em dados reais foi realizada. Pôde-se constatar que a medida que o tamanho amostral e os percentuais de perdas aumentam a imputação INLA apresenta valores baixos em relação a variabilidade, média e mediana. Fato este que ocorre de forma ao contrário imputação MICE. Porém, vale destacar que independente do valor da NRMSE aumentar a medida que o tamanho amostral e percentuais de perdas aumentam, a imputação MICE apresenta, em média, seus valores imputados muito mais próximos dos valores verdadeiros. Sendo assim, considera-se este método como uma boa alternativa na aplicação de modelos espaciais binários. E assim, trazer mais alternativas para subsidiar a citricultura brasileira.