WAGNER HUGO BONAT

APLICAÇÕES DE INFERÊNCIA BAYESIANA APROXIMADA PARA MODELOS GAUSSIANOS LATENTES ESPAÇO TEMPORAIS

CURITIBA

FEVEREIRO 2010

WAGNER HUGO BONAT

APLICAÇÕES DE INFERÊNCIA BAYESIANA APROXIMADA PARA MODELOS GAUSSIANOS LATENTES ESPAÇO TEMPORAIS

Dissertação apresentada ao Curso de Pós-graduação em Métodos Numéricos em Engenharia do Setor de Tecnologia do Centro de Estudos de Engenharia Civil Professor Inaldo Ayres Vieira da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Ciências.

Orientador: Prof. PhD. Paulo Justiniano Ribeiro Jr.

CURITIBA

FEVEREIRO 2010

TERMO DE APROVAÇÃO

WAGNER HUGO BONAT

APLICAÇÕES DE INFERÊNCIA BAYESIANA APROXIMADA PARA MODELOS GAUSSIANOS LATENTES ESPAÇO TEMPORAIS

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Ciências, pelo Programa de Pós-Graduação em Métodos Numéricos em Engenharia do Setor de Tecnologia do Centro de Estudos de Engenharia Civil Professor Inaldo Ayres Vieira da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. PhD. Paulo Justiniano Ribeiro Jr.
Universidade Federal do Paraná

Profa. PhD Marilia Sá Carvalho
Fundação Oswaldo Cruz - FIOCRUZ

Profa. PhD Silvia Emiko Shimakura
Universidade Federal do Paraná

Dr. Ramiro Ruiz Cárdenas
Universidade Federal de Minas Gerais

Curitiba, 22 de fevereiro de 2010.

A meus pais, Hugo e Leoni Bonat.

AGRADECIMENTOS

Em primeiro lugar agradeço à Deus e a Nossa Senhora do Carmo, por serem os pontos de apoio nos momentos difíceis de minha vida.

Agradeço a meus pais Hugo e Leoni Bonat por sempre acreditarem na minha capacidade e por me darem o maior presente que um filho pode ganhar a EDUCAÇÃO. A minha irmã Fabiana Bonat e minha sobrinha e afiliada Gabriele Bonat Michtal pelos momentos de descontração que tanto ajudaram na elaboração desta dissertação.

A meus amigos de tantos anos Rodrigo Baumann, Daniel Fernandez e Aparecido Correa, pela convivência durante os últimos 10 anos.

Aos estatísticos que entraram junto comigo no Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Silvio, Marcos e Antonio.

A Ana Beatriz Tozzo Martins por momentos de verdadeiras vitórias em estudos conjuntos na fase de elaboração de sua tese de doutorado, o meu crescimento pessoal e teórico na sua presença foi enorme, merece um grande agradecimento.

A Henrique Silva Dallazuanna e Gledson Picharski por sempre me ajudarem quando os conhecimentos em R e LaTeXme faltaram.

A minha namorada, amiga e companheira Rodriane Moreno pelo apoio e carinho durante esta etapa da minha vida.

A todo o pessoal do PPGMNE principalmente a Maristela pela sua eterna alegria em resolver todos os problemas burocrático durante o mestrado.

Em nome de Nicole Machuca Brassac agradeço a equipe do Lactec pela cessão dos dados do exemplo de avaliações da qualidade da água. Em nome da Dra Leda Régis agradeço à equipe de projeto SAUDAVEL pela cessão do conjunto de dados sobre fatores associados a ocorrência de Aedes aegypti em Recife/PE. E em nome do pesquisador Dr Renato Bassanezi agradeço ao FUNDECITRUS a cessão dos dados de leprose dos citrus.

Finalmente, um grande agradecimento ao Prof. Paulo Justiniano Ribeiro Júnior pela confiança e perfeita orientação durante todo o desenvolvimento da dissertação.

Sumário

Lista de Figuras ..... **vi**

Lista de Tabelas ..... **ix**

Resumo ..... **x**

Abstract ..... **xii**

1 INTRODUÇÃO **1**

2 APLICAÇÕES **5**

2.1 Qualidade da água em reservatórios operados pela COPEL no estado do Paraná. 5

2.1.1 Motivação e objetivos 5

2.1.2 O experimento 7

2.2 Investigando fatores associados a ocorrência de ovos de *Aedes aegypti* coletados em ovitrampas, em Recife/PE. 8

2.2.1 Motivação e objetivos 8

2.2.2 O experimento 10

2.3 Análise do padrão espaço-temporal da leprose-dos-citrus. 13

2.3.1 Motivação e objetivos 13

2.3.2 O experimento 15

3 METODOLOGIA **16**

3.1 Modelos aditivamente estruturados 16

3.2 Modelos Gaussianos latentes para processos espaço-temporais 18

3.2.1 Modelos com efeitos principais 19
3.2.2 Especificando prioris para efeitos de interação 20
3.2.3 Hiperprioris 24
3.3 Inferência Bayesiana Aproximada via INLA para modelos Gaussianos latentes 25
3.4 Aproximações Gaussianas 27
3.5 Integração aproximada aninhada de Laplace 28
3.5.1 Explorando $\widetilde{\pi}(\boldsymbol{\theta}|y)$ 29
3.5.2 Aproximando $\pi(\boldsymbol{\theta}_j|y)$. 31
3.5.3 Aproximando $\pi(x_i|\boldsymbol{\theta},y)$ 31
3.6 Aproximando a Verossimilhança marginal 33
3.7 Critério de informação da *Deviance* 33

4 RESULTADOS 35

4.1 Qualidade da água em reservatórios operados pela COPEL no estado do Paraná 36
4.2 Investigando fatores associados a ocorrência de ovos de *Aedes aegypti* coletados em ovitrampas em Recife/PE 44
4.3 Análise do padrão espaço-temporal da leprose-dos-citrus 59

5 CONCLUSÕES 67

Referências Bibliográficas 71

Anexo A – Apêndice 75

Lista de Figuras

Figura 2.1 Mapa temático das usinas hidrolétricas da COPEL. 7

Figura 2.2 Mapa do bairro Brasilia Teimosa, Recife/PE. Malha de armadilhas classificadas por grupos de coletas. 11

Figura 2.3 Representação do talhão 15

Figura 3.1 Representação simbólica do modelo de efeito principais. Círculos representam independência a priori, retângulos representam dependência a priori. Observações no tempo espaço são indicadas por círculos sólidos. 21

Figura 3.2 Representação simbólica para os quatro tipos de interação. Círculos representam independência a priori, retângulos representam dependência a priori. 22

Figura 3.3 Ilustração da exploração da marginal posteriori para $\boldsymbol{\theta}$. 31

Figura 4.1 Análise descritiva para o Índice de Qualidade da Água. As linhas cheias representam os limites da classe Bom e Ótimo. 37

Figura 4.2 Distribuição a posteriori para o intercepto e efeitos dos locais de coleta. 39

Figura 4.3 Sobreposição do efeito temporal ajustado pelo INLA e GAM. 40

Figura 4.4 Posterioris para os parâmetros de interesse do modelo 3. 41

Figura 4.5 Sobreposição do efeito temporal usando diferentes prioris. 42

Figura 4.6 Sobreposição do efeito das usinas (UHEs) usando diferentes prioris. ... 43

Figura 4.7 Tecelagem de Voronoi com base na malha de armadilhas para a construção da matriz de vizinhança. 45

Figura 4.8 Boxplots do log das contagens de ovos por categorias para cada covariável (X representa a média amostral), bairro Brasilia Teimosa Recife/PE. .. 47

Figura 4.9 Comparação entre as distribuições a posteriori e assintótica das estimativas dos efeitos das covariáveis locais. 50

Figura 4.10 Sobreposição do efeito estimado e respectivas faixas de confiança de cada covariável ambiental pelas abordagens INLA e GAM. 52

Figura 4.11 Sobreposição do efeito estimado de cada covariável conforme o modelo proposto em (BONAT et al., 2009). 54

Figura 4.12 Sobreposição do efeito temporal estimado conforme o modelo proposto em (BONAT et al., 2009) pelas abordagens INLA e GAM. 55

Figura 4.13 Sobreposição do efeito espacial estimado conforme o modelo proposto em (BONAT et al., 2009) pelas abordagens INLA e GAM. 56

Figura 4.14 Mapas das estimativas do efeito espacial pelas abordagens INLA e GAM 57

Figura 4.15 Diagrama de dispersão entre o efeito espacial calculado pelas abordagens INLA e GAM. 58

Figura 4.16 Mapas dos desvios padrão para cada localização espacial pelas abordagens INLA e GAM. 59

Figura 4.17 Ajuste de cada bloco de parâmetros do modelo 8. 61

Figura 4.18 Sobreposição do efeito temporal estimado via INLA e GAM na estrutura do modelo 7. 62

Figura 4.19 Efeito espacial estimado via INLA e GAM na estrutura do modelo 7. . 63

Figura 4.20 Comparação entre o percentual observado e estimado pelas abordagens INLA e GAM por data de coleta. 64

Figura 4.21 Comparação entre o percentual de acertos estimado pelas abordagens INLA e GAM por data de coleta. 65

Figura 4.22 Distribuições a posteriori de acordo com a especificação de diferentes prioris para os parâmetros de precisão dos efeitos espaciais e temporais. .. 66

Lista de Tabelas

Tabela 2.1 Relação de covariáveis locais. 12

Tabela 4.1 Modelos ajustados, critério de informação da *Deviance*, número estimado de parâmetros, verossimilhança marginal e critério de informação de *Akaike*. 38

Tabela 4.2 Resultados do modelo 3 via INLA e GAM. 39

Tabela 4.3 Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados. 43

Tabela 4.4 Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal. 45

Tabela 4.5 Ajustes dos modelos para cada covariável na presença dos efeitos espaciais e temporais, abordagens INLA e GAM. 48

Tabela 4.6 Ajuste do modelo proposto em Bonat et al. (2009) pelas abordagens INLA e GAM. 53

Tabela 4.7 Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados. 58

Tabela 4.8 Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados, verossimilhança marginal e critério de informação de *Akaike*. 60

Resumo

A família dos modelos gaussianos latentes é adaptável a uma grande quantidade de aplicações que requerem modelagem complexa. Em particular, dados espaço-temporais estão entre as mais desafiadoras para modelagem estatística. O objetivo deste trabalho foi revisar algumas possíveis estratégias de modelagem para dados deste tipo, incluindo interações espaço-temporal. A inferência nesta classe de modelos é comumente realizada usando métodos computacionalmente intensivos, tais como, os algoritmos MCMC *Markov Chain Monte Carlo*. Entretanto implementações rotineiras de tais algoritmos em problemas espaciais e/ou temporais não estão livres de problemas associados à dimensão e estrutura de dependências. Assim novos métodos e algoritmos para inferência nesta família de modelos têm sido propostos. Este trabalho revisou a abordagem 'INLA' '*Integrated Nested Laplace Approximations*' proposta por RUE, MARTINO e CHOPIN (2009), que se mostrou eficiente para ajustar modelos altamente estruturados em diversas situações práticas. A nova metodologia de inferência foi aplicada a três problemas com diferentes objetivos e estruturas no conjunto de dados. Sempre que possível os modelos ajustados pelo INLA, foram confrontados com ajustes de modelos aditivos generalizados para verificar a concordância entre as abordagens, principalmente no que diz respeito ao modo como captam os efeitos espaciais e temporais. Os conjuntos de dados foram selecionados de modo a cobrir os modelos mais comumente usados na literatura. O primeiro conjunto refere-se a avaliações da qualidade da água, assumindo normalidade para a variável resposta. O segundo conjunto tem como resposta a contagens de ovos do mosquito *Aedes aegypti* coletados em ovitrampas em Recife/PE, para a qual assume-se a distribuição binomial negativa. O terceiro conjunto corresponde a dados sobre a doença leprose-dos-citros, assumindo a distribuição binomial para a variável resposta de presença ou ausência da doença. Nos três conjuntos de dados analisados foi feita ainda uma comparação entre os resultados obtidos pelas abordagens INLA e GAM (modelos aditivos generalizados). No primeiro problema os resultados produzidos pelas duas abordagens foram semelhantes. Para o segundo conjunto algumas diferenças importantes foram encontradas, covariáveis que pela abordagem GAM eram indicadas como significativas, pela abordagem INLA foram indicadas como não significativas, embora com predições semelhantes para os efeitos espaciais e temporais. O último e mais desafiador exemplo, mostrou uma grande diferença entre as abordagens na forma como captam os efeitos espaciais e temporais. De forma geral a abordagem GAM tende a suavizar demais estes efeitos e fornece intervalos de confiança pouco realísticos, ao passo que a abordagem INLA apresenta melhores resultados e intervalos de credibilidade para predições com melhor cobertura. Neste caso não foi possível obter estimativas confiáveis de interações espaço-temporais. Nos três exemplos, medidas de concordância entre as observações e os modelos foram tomadas, foram elas: erro quadrático médio, erro absoluto médio, correlação entre observados e preditos e taxa de cobertura. Por estas medidas em todos os exemplos analisados a abordagem INLA se mostrou mais flexível e adequada apresentando melhores resultados.

Palavras-chave: Inferência bayesiana aproximada, Campos aleatórios Markovianos, Aproximação de Laplace, Modelos de regressão aditivamente estruturados, Modelos espaço-temporais.

Abstract

The family of latent Gaussian models is flexible and suitable for a wide range of applications requiring complex modelling. In particular, spatial temporal data is among the most challenging structures for statistical models. The present work revises some strategies for modelling such data, including spatio-temporal interactions. Inference for such models usually relies on computational intensive algorithms such as MCMC (Monte Carlo Markov Chain). However, routine implementation of algorithms of this kind for spatial and/or temporal problems typically faces difficulties related to dimensionality and dependence structure. Proposals for new methods and algorithms for this class of models are still investigated by on recent literature. The INLA (integrated nested Laplace approximation) approach recently proposed by RUE, MARTINO and CHOPIN, 2009 has proved efficient and promising for different applications and is investigated here and used in the analysis of three problems with different targets and data structures. The first refers to the assessment of water quality with an assumed Gaussian distribution for the response variable. The second has dengue mosquito (*Aedes aegypti*) egg counts collected at ovitraps in the municipality of Recife/PE as a negative binomial response variable. The final application refers to citrus leprosis disease assuming a binomial distribution for the presence/absence binary data. Close results for the two approaches were obtained in the first case. For the second some relevant differences were found such as significance of coefficients associated to covariates being investigated, although with comparable predictions of spatio-temporal effects. The third problem proved more ch alleging, with no reliable estimates of spatio-temporal interactions and showing greater differences between the two approaches with an apparent over-smoothing of spatial effect obtained by GAM associated with low coverage for prediction intervals. Measures of concordance between predicted and observed values such as mean and absolute squared error, correlation and coverage rates were used when comparing the two approaches within the three applications. Results show that in general INLA provides a more realistic and flexible approach producing more reliable results.

Key-words: Approximate Bayesian inference, Gaussian Markov random Fields, Laplace approximation, Structured additive regression models, models space-time.

1 INTRODUÇÃO

Modelos estocásticos têm sido amplamente utilizados tanto na comunidade científica como no mundo dos negócios em geral. Estudos de mercado usando modelos altamente estruturados, modelos de predição em séries financeiras, análises de componentes de solos para agricultura de precisão, mapeamento de doenças são algumas das áreas de aplicações dos modelos estocásticos.

Nesta diversidade de aplicações é muito fácil encontrar situações de relevância prática, onde os modelos tradicionais, considera-se por tradicionais a classe de Modelos Lineares Generalizados (GLM) deixam de ser adequados, em geral pela existência de pelo menos uma das seguintes características:

- para covariáveis contínuas, a suposição de um efeito estritamente linear no preditor pode não ser apropriada,
- as observações podem ser correlacionadas no espaço,
- as observações podem ser correlacionadas no tempo,
- interações complexas podem ser necessárias para modelar o efeito conjunto de algumas covariáveis,
- heterogeneidade entre indivíduos ou unidades pode não ser suficientemente descrita por covariáveis.

Nestas situações a classe dos modelos de regressão estruturados aditivamente têm sido extensivamente usada, por ser altamente flexível, ver por exemplo (FAHRMEIR; TUTZ, 2001) para uma descrição mais detalhada. Para melhor explorar esta classe de modelos, considere uma variável resposta y_i assuma que a distribuição de probabilidade

desta variável, possa ser escrita na forma da família exponencial, onde a média μ_i é ligada ao preditor estruturado aditivamente η_i através de uma função de ligação $g(.)$, tal que $g(\mu_i) = \eta_i$. O preditor pode acomodar diferentes efeitos de várias covariáveis de forma aditiva:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i$$

Aqui, as $f^{(j)}(.)$'s são funções desconhecidas das covariáveis u, os β_k's representam efeitos lineares das covariáveis z e os ε_i são termos não estruturados. A grande dimensão de aplicações de modelos desta classe, vêm das diferentes formas que as funções $f^{(j)}$ podem tomar. Os modelos Gaussianos Latentes são um subconjunto de todos os modelos Bayesianos estruturados aditivamente, onde se supõe uma priori Gaussiana para α,$f^{(j)}(.)$, β_k e ε_i.

Uma situação bastante complexa e que têm tido grande avanço com o uso de modelos Gaussianos latentes, é a situação de dados espaço-temporais. Dados com estrutura espaço-temporal consistem de observações de uma variável resposta e um conjunto de covariáveis, onde, adicionalmente a localização espacial e temporal de cada unidade da amostra é conhecida. Em geral,o problema específico de dados espaço-temporais é que as observações são correlacionadas no tempo e/ou espaço, podendo ainda ter um efeito de interação espaço-tempo. É claro que as aplicações de modelos Gaussianos latentes não se restrigem apenas a situação espaço-temporal, existe uma lista enorme de aplicações destes modelos, ver por exemplo (GELMAN et al., 2004) e (ROBERT; CASELLA, 1999).

Como é de se esperar a grande flexibilidade destes modelos, vêm acompanhada de grandes problemas para a estimação dos parâmetros de interesse. A abordagem comum para inferência em modelos Gaussianos latentes no contexto completamente Bayesiano, são os métodos de Monte Carlo via Cadeias de Markov, ou MCMC(*Markov Chain Monte Carlo*) na sigla em inglês.

É também conhecido que tais métodos tendem a apresentar um desempenho insatisfatório quando aplicados para tais modelos. Rue, Martino e Chopin (2009) argumentam que vários são os fatores que explicam isto, porém praticamente todos são decorrentes da alta dependência entre os parâmetros de interesse do modelo. Os autores ainda citam algumas possíveis estratégias para contornar este problema, como o algoritmo *one block* (KNORR-HELD; RUE, 2002), no qual são amostrados todos os parâmetros conjuntamente construindo uma proposta conjunta baseada em uma aproximação Gaus-

siana para a distribuição condicional completa dos campos Gaussianos, além de considerar reparametrizações ou usar variáveis auxiliares para simplicar as aproximações Gaussianas. Porém, mesmo com estes desenvolvimentos o ajuste de modelos via MCMC permanece complicada e lento do ponto de vista do usuário final (RUE; MARTINO; CHOPIN, 2009).

Tendo estes problemas em mente, alguns métodos de aproximação determinística vêm sendo propostos. O método *Variational Bayes* (VB) (BISHOP, 2006) ; (HINTON; CAMP, 1993) desenvolvido na literatura de aprendizagem de máquina, bem como, o método *Expectation-Propagation* (EP) (MINKA, 2001) vêm apresentando bons resultados em diversas aplicações, ver (BEAL, 2003) para uma extensa revisão sobre VB e, (KUSS; RASMUSSEN, 2005) para aplicações de EP.

Neste sentido também, Rue, Martino e Chopin (2009) propuseram uma abordagem para fazer inferência Bayesiana aproximada, em modelos Gaussianos latentes, à qual é denominada de INLA - *Integrated Nested Laplace Approximations*. Os autores mostram em seu artigo que a nova abordagem é extremamente rápida, principalmente por fazer uso de algoritmos especialistas para matrizes esparsas, que são inerentes a essa classe de modelos. Além disso, mostram que a nova abordagem supera o MCMC em termos de acurácia e tempo computacional. Os autores também descrevem como usar várias aproximações para derivar ferramentas para testar o erro de aproximação, aproximar posterioris marginais, calcular quantidades como verossimilhança marginal, critério de informação da *Deviance* (DIC) e várias medidas preditivas Bayesianas.

Tendo em vista a contribuição que o artigo de Rue, Martino e Chopin (2009), trouxe para a inferência estatística, principalmente do ponto de vista Bayesiano, os objetivos desta dissertação são os seguintes:

- revisar os fundamentos dos modelos Gaussianos latentes, mostrando diferentes estruturas para acomodar efeitos suaves de covariáveis contínuas, efeitos espaciais, temporais e interações espaço-temporais,

- revisar o artigo (RUE; MARTINO; CHOPIN, 2009) onde é descrita em detalhes a metodologia INLA, mostrando que é possível estimar modelos com interações espaço-temporais, como parte da estrutura geral de modelos que este processo de estimação permite,

- aplicar a metodologia a três conjuntos de dados, o primeiro assumindo para a variável resposta a distribuição Normal, o segundo para dados típicos de contagens e o último assumindo distribuição Binomial. Assim, exemplifica-se os três modelos de regressão

mais usados em aplicações gerais.

A presente dissertação encontra-se dividida da seguinte maneira: este primeiro Capítulo busca motivar o uso de modelos altamente estruturados como os modelos Gaussianos latentes, além de dar uma breve visão sobre as possibilidades de métodos para inferência estatística no contexto Bayesiano.

O segundo Capítulo descreve os três conjuntos de dados que serão utilizados como exemplos de aplicação das metodologias descritas. O primeiro conjunto corresponde a avaliações da Qualidade da Água em reservatórios operados pela COPEL no estado do Paraná. O segundo corresponde a contagens de ovos do mosquito *Aedes aegypti* coletados em ovitrampas em Recife/PE. E o terceiro conjunto corresponde ao mapeamento da doença leprose-dos-citrus.

O terceiro Capítulo revisa a bibliografia utilizada para construir o modelo observacional, apresenta um resumo do artigo que serve de base para a dissertação, mostrando o procedimento de inferência baseado na abordagem *Integrated Nested Laplace Approximations*.

O quarto Capítulo apresenta os principais resultados para os três conjuntos de dados analisados. O quinto e último Capítulo faz uma discussão geral dos procedimentos utilizados, bem como, os possíveis pontos para pesquisas futuras. É apresentado também um apêndice com um exemplo de código desenvolvido para o ajuste de modelos espaço-temporais.

2 APLICAÇÕES

Esta seção descreve os três conjuntos de dados que serão analisados no decorrer da dissertação. O primeiro conjunto referente a Qualidade da Água em reservatórios operados pela COPEL no estado do Paraná, foi analisado por Ribeiro et al. (2008) usando análise de variância e modelos aditivos generalizados. O segundo conjunto trata de contagens de ovos do mosquito *Aedes aegypti* coletados em ovitrampas em Recife/PE, foi analisado em Bonat et al. (2009) com o uso de modelos aditivos generalizados. O terceiro conjunto se refere à incidência de leprose-dos-citros foi analisado por Franciscon et al. (2008) com o modelo autologístico.

Desta forma, fica claro que a contribuição desta dissertação não está somente nas análises de dados, mas também na exemplificação de uma nova estratégia de inferência para modelos Gaussianos latentes. Cabe ressaltar que as análises desenvolvidas aqui não tem interesse em reproduzir os resultados obtidos por outros autores.

2.1 Qualidade da água em reservatórios operados pela COPEL no estado do Paraná.

2.1.1 Motivação e objetivos

A companhia Paranaense de Energia (COPEL) opera no estado do Paraná, dezessete usinas hidrolétricas, com geração total de mais de $4,500$ MW de energia. A maior delas, em potência instalada, é a Usina Hidrolétrica (UHE) Governador Bento Munhoz da Rocha Netto, no Rio Iguaçu, localizada no município de Pinhão, com $1,676$ MW.

Os reservatórios constituídos para a geração de energia elétrica têm sido utilizados para inúmeras outras finalidades, destacando-se lazer, navegação e captação de água para abastecimento público. A qualidade da água, por si só e como determinante do crescimento de algas, plantas e outros organismos, é fundamental para que tais usinas apresentem máxima funcionalidade.

Pensando em abastecimento de água e no impacto que tais empreendimentos tem sobre o meio-ambiente, é fundamental saber como tais reservatórios atuam sobre a qualidade da água. Neste sentido, a COPEL realiza o monitoramento dos reservatórios, bem como dos rios represados a montante e jusante dos mesmos, em atendimento às condicionantes das licenças de operação destes empreendimentos.

O monitoramento realizado pela concessionária, envolve a avaliação de parâmetros da qualidade da água e o cálculo do Índice de Qualidade da Água - CETESB (IQA), que será a variável chave para a análise. A partir de um estudo realizado em 1970 pela *National Sanitation Foundation* dos Estados Unidos da América, a CETESB adaptou e desenvolveu o IQA - Índice de Qualidade das Águas, que incorpora nove parâmetros (Oxigênio Dissolvido, Temperatura, Coliformes fecais, pH, DBO, Nitrogênio Total, Fósforo Total, Turbidez, Sólidos totais) considerados relevantes para a avaliação da qualidade das águas, tendo como determinante principal a utilização das mesmas para abastecimento público (PHILIPPI; ROMERO; BRUNA, 2004).

Para cada uma das nove variáveis da qualidade da água que compreendem o IQA (DERISIO, 1992) corresponde uma curva de variação da qualidade da água, que a correlaciona a um sub-índice, q e um peso de importância w.

A formulação do índice é o produtório ponderado dos índices parciais (q_i) de cada parâmetro (i):

$$IQA = \prod_{i=1}^{9} q_i^{w_i}$$

De acordo com o valor do IQA, que varia de 0 a 100, a qualidade da água pode ser classificada como ótima ($79 < IQA < 100$), boa ($51 < IQA < 79$), aceitável ($36 < IQA < 51$), ruim ($19 < IQA < 36$) e péssima ($IQA < 19$).

A escolha pelo IQA deve-se ao fato de que o índice é um facilitador na interpretação geral das condições da qualidade dos corpos d'água, além de ser amplamente utilizado por instituições brasileiras como a CETESB, SANEPAR, IAP, entre outros, fazendo com que os resultados sejam comparáveis com outras localidades.

O objetivo da análise estatística de dados históricos do monitoramento da qualidade das águas dos reservatórios de usinas hidrolétricas operadas pela COPEL, no estado do Paraná, é identificar possíveis impactos e alterações na qualidade da água decorrente da existência dos reservatórios. A qualidade da água do rio pode ser definida pelas características da estação de montante, onde ainda não há influência do empreendimento no curso de água. Desta forma, utilizou-se dados da estação de montante como situação de referência a qual será comparada com os dados das estações de reservatório e jusante, verificando assim a melhora ou piora da qualidade da água após a passagem pelo reservatório.

2.1.2 O experimento

As usinas hidrolétricas estão distribuídas nas bacias do Tibagi, Ribeira, Piquiri, Ivaí, Iguaçu, e Litorânea. A localização de cada empreendimento encontra-se ilustrada na figura 2.1.

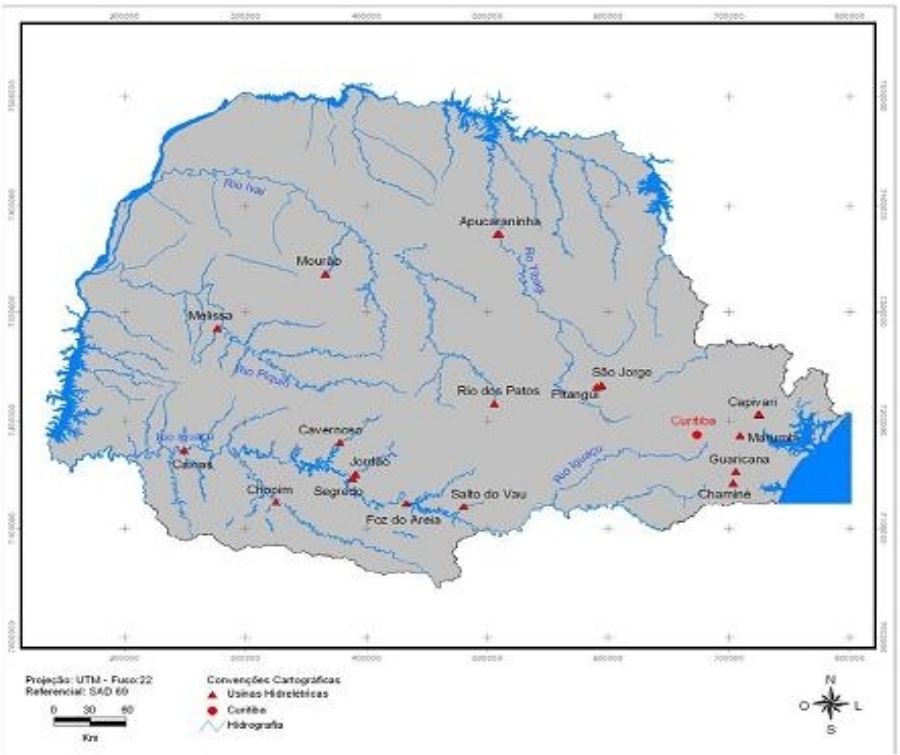

Figura 2.1: Mapa temático das usinas hidrolétricas da COPEL.

O monitoramento realizado pela Concessionária envolve a avaliação de parâmetros da qualidade da água, bem como o cálculo do Índice de Qualidade da Água (IQA), a partir

de campanhas realizadas trimestralmente desde o ano de 2003. A primeira coleta ocorreu no dia 06/03/2003 e a última no dia 30/10/2008.

Na maior parte das usinas são avaliadas três estações de monitoramento codificadas de acordo com a concessionária, sendo uma localizada a montante do reservatório, uma no próprio corpo do reservatório, e uma localizada a jusante da casa de máquinas da usina.

Nas demais, a malha amostral é maior, variando de acordo com a morfologia do reservatório e também do uso do solo no entorno. No entanto, esta malha amostral sempre atende a estrutura citada acima, com no mínimo uma estação na montante, estações de reservatório e uma estação à jusante de cada casa de força. No caso das usinas que possuem mais de três estações de monitoramento, foram selecionadas as estações de montante e jusante, e a estação de reservatório selecionada foi aquela mais próxima da barragem, onde o ambiente é caracteristicamente lêntico.

Para a análise que será apresentada foram escolhidas oito usinas, por estas apresentarem o mesmo número de coletas no tempo para todas as estações. São 23 datas de coletas em três localizações (montante, reservatório, jusante), totalizando 69 observações para cada usina. Sendo assim, o banco de dados conta com 184 observações em cada local, com um total de 552 observações.

2.2 Investigando fatores associados a ocorrência de ovos de *Aedes aegypti* coletados em ovitrampas, em Recife/PE.

2.2.1 Motivação e objetivos

O dengue é uma doença febril aguda, cujo agente etiológico é um vírus do gênero Flavivírus. São conhecidos atualmente quatro sorotipos, antigenicamente distintos: DEN-1, DEN-2, DEN-3 e DEN-4. As manifestações variam de uma síndrome viral, inespecífica e benigna, até um quadro grave e fatal de doença hemorrágica com choque.

O dengue é uma arbovirose transmitida ao homem pela picada do mosquito *Aedes aegypti*, tal mosquito tem hábitos domésticos, pica durante o dia e tem preferência acen-

tuada por sangue humano (TAUIL, 2002).

Durante quase 60 anos, de 1923 a 1982, o Brasil não apresentou registro de casos de dengue em seu território. Porém, desde 1976, o *Aedes aegypti* havia sido re-introduzido no país, a partir de Salvador, Bahia, e estava presente em muitos países vizinhos. Países da América Central, México, Venezuela, Colômbia, Suriname e alguns outros do Caribe já vinham apresentando a doença desde os anos 70.

Até os dias atuais não se dispõe de uma vacina eficaz para uso preventivo contra o dengue, apesar de todos os esforços de pesquisa para a sua produção e desenvolvimento. Enquanto não se puder contar com esta medida de controle, o único elo vulnerável da cadeia epidemiológica é o vetor.

Pensando neste elo vulnerável um dos esforços nacionais foi a elaboração do Projeto SAUDAVEL,[1] o qual pretende contribuir para aumentar a capacidade do setor de saúde no controle de doenças transmissíveis, demonstrando ser necessário desenvolver novos instrumentos para a prática da vigilância entomológica, incorporando aspectos ambientais, identificadores de risco e proteção, além de métodos automáticos e semi-automáticos, que permitam a detecção de surtos e seu companhamento no espaço e no tempo (MONTEIRO et al., 2006).

Com a intensa circulação do vírus do dengue no Brasil a partir da década de 1980, epidemias explosivas têm atingido todas as regiões brasileiras (BRAGA; VALLE, 2007). Em vista disso, também a partir desta década, diversas metodologias para a vigilância do vetor vêm sendo desenvolvidas e utilizadas no país.

Nos programas de controle do dengue, a vigilância entomológica é feita principalmente a partir da coleta de larvas, de acordo com a proposta de Connor e Monroe (1923) para medir a densidade de *Aedes aegypti* em áreas urbanas. Esta metodologia consiste em vistoriar os depósitos de água e outros recipientes localizados nas residências e demais imóveis, como borracharias, ferros-velhos, cemitérios, entre outros tipos de imóveis considerados estratégicos, por produzirem grande quantidade de mosquitos adultos, para o cálculo dos índices de infestação predial (IIP) e de Breteal (IB).

A coleta de larvas (ou pesquisa larvária, como é comumente chamada no Brasil) é importante para verificar o impacto das estratégias básicas de controle da doença, dirigidas a eliminação das larvas do vetor. Entretando, não é um bom indicador para se medir a abundância do adulto, e ineficaz para estimar o risco de trasmissão (BRAGA; VALLE,

[1]Sistema de Apoio Unificado para a Detecção e Acompanhamento em Vigilância Epidemiológica (http://saudavel.dpi.inpe.br)

2007), embora venha sendo usada com essa finalidade (FOCKS, 2000).

Apesar disso, para avaliar a densidade do vetor instalam-se armadilhas de oviposição e armadilhas para coleta de larvas, que visam estimar a atividade de postura. A armadilha de oviposição, também conhecida no Brasil como "ovitrampa", é destinada a coleta de ovos. Em um recipiente de cor escura, adere-se um material àspero que permite a fixação dos ovos depositados. Em 1965, iniciou-se o uso de ovitrampas para a vigilância de populações adultas de *Aedes aegypti* (FAY; ELIASON, 1965). Posteriormente, ficou demonstrada a superioridade dessas armadilhas em relação a pesquisa larvária, para a verificação da ocorrência do vetor (FAY; ELIASON, 1966).

As ovitrampas fornecem dados úteis para a investigação da distribuição espacial e temporal (sazonal) de ovos do mosquito. Dados obtidos com essa metodologia também são usados para verificar o impacto de vários tipos de medidas de controle, que envolvem a redução do vetor com inseticidas.

O objetivo da análise deste conjunto de dados é ajustar modelos que permitam determinar fatores de risco e proteção associados à ocorrência de ovos do mosquito *Aedes aegypti*, com base em dados de um experimento conduzido pelo "Projeto SAUDAVEL"na cidade de Recife/PE. Entende-se aqui, como fatores de risco/proteção tanto covariáveis associadas à armadilha, como presença de recipientes grandes ou pequenos que possam conter água em suas proximidades, como também aspectos abióticos (climáticos) como temperatura, precipitação e umidade. Possíveis relações espaciais entre as armadilhas, a possibilidade de uma relação temporal e ainda de uma interação espaço-temporal entre as coletas serão investigadas.

2.2.2 O experimento

O experimento está sendo desenvolvido pelo projeto SAUDAVEL na cidade de Recife/PE. Neste experimento foram instaladas 464 armadilhas (ovitrampas) para capturar ovos do mosquito *Aedes aegypti*. Estas armadilhas começaram a ser monitoradas em março de 2004. O experimento está sendo realizado em 7 dos 94 bairros da cidade. Os dados que serão apresentados aqui referem-se apenas ao bairro Brasília Teimosa, por este apresentar uma quantidade expressiva de observações.

A coleta de dados neste bairro teve início em 04/01/2005 e até 15/05/2007 período para o qual os dados estão disponíveis, foram realizadas 2480 observações em 80 armadi-

lhas no período correspondente a 124 semanas. A rede de armadilhas foi instalada de modo a cobrir toda a extensão do bairro, a cada sete dias é feita a contagem de ovos encontrados em cerca de um quarto das armadilhas, assim um ciclo de 28 dias é necessário para que todas as armadilhas sejam monitoradas. Ressalta-se que as coletas são semanais porém apenas 1/4 das armadilhas são observadas em cada semana, tem-se assim quatro grupos de armadilhas, de acordo com a semana em que ela é observada. A figura 2.2 mostra a localização das armadilhas, destacando cada grupo.

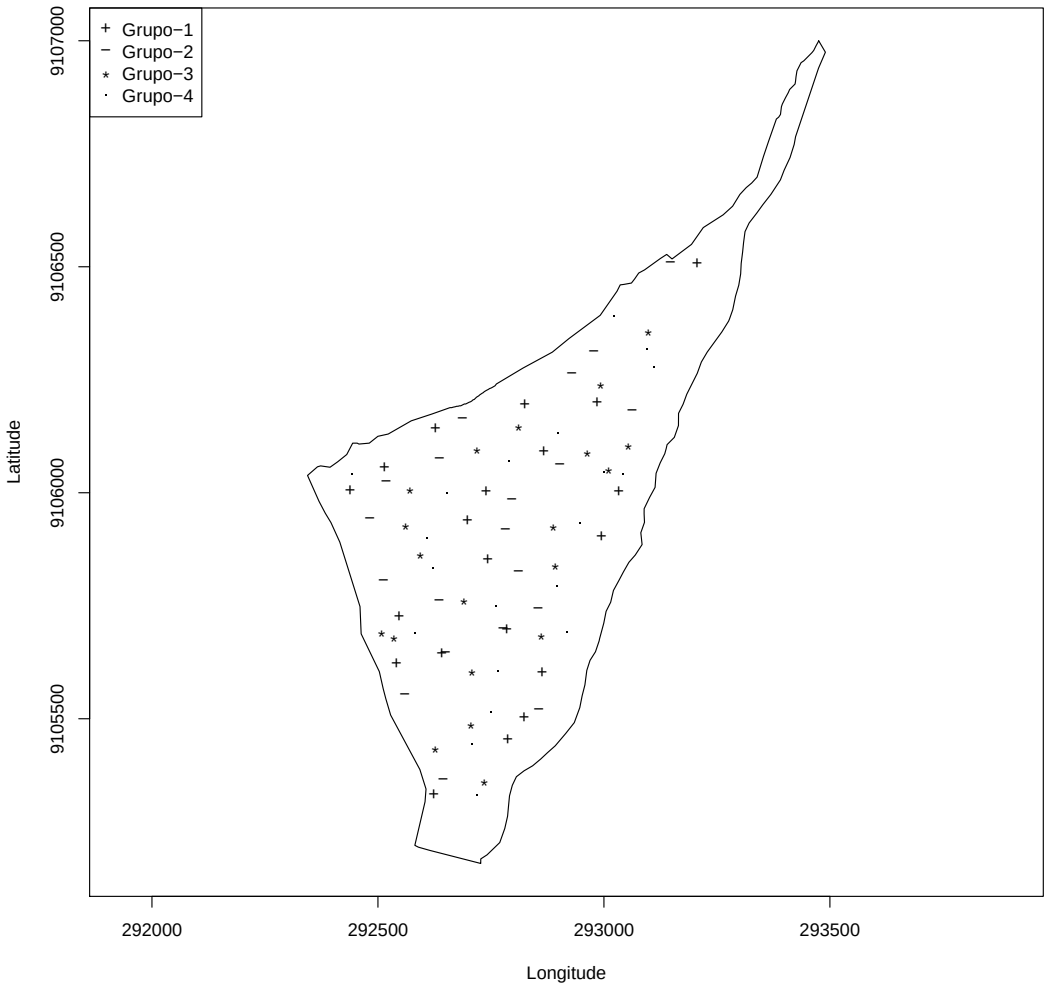

Figura 2.2: Mapa do bairro Brasilia Teimosa, Recife/PE. Malha de armadilhas classificadas por grupos de coletas.

Cada armadilha contém três lâminas de material áspero na qual a fêmea do mosquito coloca os ovos, quando recolhidas são levadas para um laboratório especializado onde a contagem de ovos é realizada. O departamento de entomologia da FIOCRUZ/PE e os serviços de saúde locais são os coordenadores operacionais e logísticos e responsáveis

pela realização do experimento (MONTEIRO et al., 2006).

Regis et al. (2008) descrevem de forma ampla o experimento SAUDAVEL/Recife, bem como todo o escopo do projeto que visa desenvolver metodologias e tecnologias para o monitoramento de populações de *Aedes aegypti* através de contagens de ovos coletados em ovitrampas.

O banco de dados do SAUDAVEL/Recife, possibilita a construção de diversas covariáveis a serem utilizadas nos modelos, seguindo Bonat et al. (2009), o conjunto de covariáveis foi dividido em dois grupos, o primeiro das covariáveis 'locais' referentes às armadilhas e o segundo de covariáveis 'ambientais' referentes a fatores abióticos, representados por variáveis climáticas.

A tabela 2.1 resume o conjunto de covariáveis locais e a codificação adotada para as análises.

Covariáveis	Níveis	Descrição
Tipo de imóvel	0	Residencial
	1	Não residencial
Quintal	0	Apresenta quintal
	1	Não apresenta quintal
Água ligada a rede geral	0	Ligada
	1	Não ligada
Abastecimento de Água	0	Diário
	1	Não diário
Água canalizada no cômodo	0	Canalizada
	1	Não canalizada
Fatores de risco	0	Apresenta
	1	Não apresenta
Recipientes grandes sem tampa	0	Apresenta
	1	Não apresenta
Recipientes grandes com tampa	0	Apresenta
	1	Não apresenta
Recipientes pequenos sem tampa	0	Apresenta
	1	Não apresenta
Recipientes pequenos com tampa	0	Apresenta
	1	Não apresenta

Tabela 2.1: Relação de covariáveis locais.

São considerados fatores de risco, plantas em vasos, charco/poça, garrafas, fossa externa, piscina, poço elevador, laje sem telhado ou calhas. Na categoria de recipientes grandes são considerados, tanques, caixa d'água ou tóneis. Na categoria de recipientes

pequenos são considerados, jarros de barro ou baldes.

Uma outra covariável não elencada na tabela, mas que também foi levada em consideração na análise é denominada de **Grupos**. Esta covariável corresponde a semana em que a armadilha é monitorada, como as coletas são realizadas semanalmente é feita em grupos de armadilhas.

As covariáveis climáticas disponíveis no banco de dados e utilizadas nas análises foram precipitação, umidade relativa do ar, temperatura máxima e mínima. A mensuração destas covariáveis é feita diariamente por uma estação de monitoramento. Seguindo novamente Bonat et al. (2009), foram tomadas médias mensais destas covariáveis para relacionar com as contagens de ovos que são realizadas semanalmente. Foram contabilizadas estas covariáveis com uma defasagem de até três meses da observação.

Desta forma, o banco de dados utilizado nesta análise conta com 2480 observações divididas em 80 armadilhas (localizações) observadas em 124 datas de coletas. Para a construção do modelo tem-se 11 covariáveis locais além de quatro medidas climáticas tomadas em diferentes pontos no tempo.

2.3 Análise do padrão espaço-temporal da leprose-dos-citrus.

2.3.1 Motivação e objetivos

O Brasil é um dos maiores produtores de citrus do mundo, aproximadamente 28% do suco de laranja produzido no mundo é proveniente do Brasil. Além disso, cerca de 80% do suco concentrado ofertado no mercado internacional é brasileiro. Citricultores, industriais e cientistas brasileiros criaram um setor de ponta na agroindústria nacional e trabalham para o aumento da produtividade e também controle e manutenção da capacidade produtiva agrícola.

Apesar de muitos estudos, a produção é limitada devido à presença de doenças nos pomares produtores (MARQUES et al., 2007). Patógenos transmitidos por vetores (insetos e ácaros) são especialmente difíceis de serem controlados e, quando são, a custos financeiros e ambientais extremamente elevados (GUIRADO, 2000).

Algumas doenças como a leprose dos citros, podem causar a erradicação de talhões inteiros de plantas, na fazenda produtora, comprometendo a quantidade e a qualidade das frutas cítricas. A leprose dos citrus foi considerada nos últimos tempos, uma das mais importantes viroses na citricultura nacional, pois reduz a produção e o período de vida das plantas de citros (RODRIGUES, 2000).

Os dados de plantas cítricas são, em geral, coletados ao longo do tempo nas inspeções periódicas, as plantas estão dispostas dentro de um talhão, em linhas e colunas com certa estrutura de vizinhança. A disposição das plantas, referenciadas por coordenadas locais, permite estudar o padrão espacial de doenças cítricas. Conhecer o movimento e a distribuição espacial de vetores que transmitem doenças e que afetam a disseminação nas plantas e o progresso da doença auxiliam na determinação de práticas de controle. Existem relativamente poucos trabalhos que buscam descrever o padrão espaço-temporal relacionados a dinâmica de doenças cítricas na população de plantas hospedeiras (FRANCISCON et al., 2008).

Em geral, nos 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 *quadrats*, 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). São exemplos de modelos estatísticos com componentes espaciais, o modelo autologístico (KRAINSKI et al., 2008), o modelo CAR (*Conditional Autoregressive*) (BESAG, 1975) e os modelos geoestatísticos (DIGGLE; RIBEIRO, 2006).

Mesmo com estes modelos a característica temporal da doença não é levada em consideração, sendo assim, é ainda necessário incorporar nos modelos espaciais o componente temporal.

O interesse aqui é propor um modelo realístico dado a estrutura do experimento e que possa ser usado para entender a dinâmica de infestação das plantas pela leprose-dos-citrus, ou seja, propor um modelo que leve em consideração o espaço e o tempo simultâneamente.

2.3.2 O experimento

Os dados que serão utilizados para a construção do modelo espaço-temporal, são de incidência de leprose-dos-citros em um talhão de laranjeira 'Valência', enxertada sobre o limoeiro 'Cravo', plantado em 1996 e localizado no município de Santa Cruz do Rio Pardo, SP ($22^o53'56''S, 49^o37'58''W$). O talhão apresentava 20 linhas de plantas, com 58 plantas em cada linha. O espaçamento entre linhas era de $7,5$ e $3,8$ metros, na linha. Os dados foram coletados em 45 avaliações, entre 01/2002 e 11/2004. Neste período, não foram realizadas pulverizações com acaricidas no talhão, de modo a não afetar a população do ácaro da leprose.

Nas avaliações, foram coletadas informações sobre os sintomas da doença e a presença do ácaro transmissor, em cada planta do talhão. A avaliação dos frutos, ramos e folhas, em 25 frutos, 25 folhas e 25 ramos por quadrante dossel, amostrados a esmo (CZERMAINSKI, 2006), num total de 100 unidades de observação, por tipo de estrutura por planta. O número de ácaros da leprose, em todas as plantas, foi obtido pela observação em cinco frutos, localizados no interior das copas, provenientes da florada principal, e em cinco ramos externos da copa (CZERMAINSKI, 2006). A partir dos registros de infestação de ácaros e de infecção pelo CiLV (sintomas da doença) nas plantas, foram obtidas as incidências codificadas de maneira binária: 0 representa a ausência do evento "presença de ácaro ou sintoma"; e 1 representa a presença do evento.

A figura 2.3 mostra o talhão que será analisado.

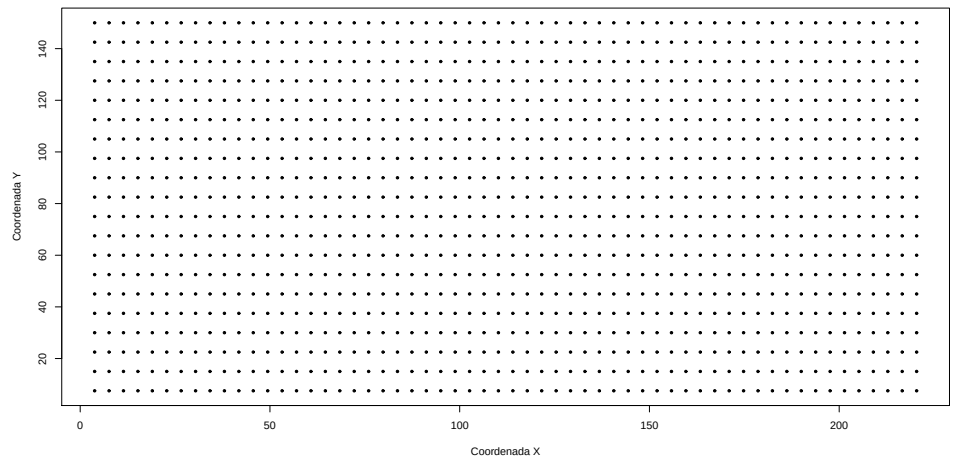

Figura 2.3: Representação do talhão

3 METODOLOGIA

Em todas as aplicações descritas no capítulo 2, o interesse está em analisar a influência de um conjunto de covariáveis em uma variável resposta levando em consideração a estrutura espacial e/ou temporal do experimento. Ressalta-se que o tipo da variável resposta não é o mesmo em cada aplicação, por exemplo, no caso de ovos do mosquito *Aedes aegypti* a resposta é uma contagem, enquanto que para o caso de leprose-dos-citrus a variável resposta é dicotômica.

Ao pensar em postular um modelo essas diferenças devem ser levadas em consideração, porém o ideal é ter um único conjunto de procedimentos que seja adequado para a maioria das situações. 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 na próxima seção.

A notação utilizada neste capítulo segue a do artigo Rue, Martino e Chopin (2009) onde os autores não fazem distinção entre y_i para a variável aleatória e dados observados.

3.1 Modelos aditivamente estruturados

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

$$P(y|\mathbf{z}) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\right)c(y,\phi) \tag{3.1}$$

onde $b(.)$,$c(.)$ são funções que dependem da distribuição atribuída a variável resposta, ϕ é um parâmetro de escala ou dispersão comum para todas as observações, e θ é o

parâmetro natural da família exponencial. Uma lista das distribuições mais comuns e seus parâmetros pode ser encontrada em Fahrmeir e Tutz (2001). A média de cada variável aleatória $E(y_i|\mathbf{z},\beta)$ é ligada ao preditor linear da forma

$$g(\mu_i) = \eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{3.2}$$

onde $g(.)$ é uma função conhecida chamada de função de ligação,α e β são parâmetros de regressão desconhecidos, e ε_i é um termo não estruturado. É importante notar que no preditor assume-se que todas as covariáveis $\mathbf{z}$ têm um relacionamento estritamente linear com a resposta. Esta restrição impossibilita uma análise adequada do conjunto de dados, em situações mais gerais como as apresentadas no capítulo 2, onde 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 em 3.2 por um preditor aditivamente estruturado

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{3.3}$$

A diferença de 3.2 para 3.3 são as funções $f()$ 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. Dado as diferentes suposições sobre as funções $f()$ tem-se diversos modelos encontrados na literatura. Por exemplo, ao assumir que tais funções são *splines* tem-se os modelos aditivos generalizados (GAM) (WOOD, 2006), que considera que o efeito das covariáveis u são fixos. Outro exemplo é quando se atribui para as funções $f()$ um campo aleatório gaussiano, por exemplo no contexto dos modelos geoestatísticos (DIGGLE; RIBEIRO, 2006).

Modelos Gaussianos latentes são um subconjunto de todos os modelos Bayesianos aditivos onde assume-se uma priori gaussiana para α, $f^{(j)}(.)$,β_k e ε_i. Denote por $\underline{x}$ o vetor de todas as variáveis gaussianas latentes, e $\underline{\theta}$ o vetor de hiperparâmetros, que não são necessariamente Gaussianos. Na literatura de aprendizagem de máquina o termo Modelos para processos Gaussianos é frequentemente usado.

Modelos Gaussianos latentes tem uma numerosa lista de aplicações. Sendo muitos dos modelos Bayesianos estruturados desta forma. Sem pretender ser exaustivo nos casos

especiais de modelos Gaussianos latentes pode-se citar: modelos lineares generalizados Bayesianos (DEY; GHOSH; MALLICK, 2000). Modelos P-splines (LANG; BREZGER, 2004), modelos *random walk* (FAHRMEIR; TUTZ, 2001) ; (RUE; HELD, 2005). Modelos dinâmicos (KITAGAWA; GERSCH, 1996) ; (WEST; HARRISON, 1997). Modelos espaciais e espaço-temporal (BESAG; YORK; MOLLIé, 1991), (KNORR-HELD, 2000), modelos Gaussianos indexados continuamente (BANERJEE; CARLIN; GELFAND, 2004) ; (DIGGLE; RIBEIRO, 2006), modelos de textura (MARROQUIN et al., 2001), entre outros.

Em muitas aplicações, o modelo final vai consistir na soma de vários componentes, tal como um efeito espacial, efeitos aleatórios, e efeitos lineares e não-lineares de covariáveis. Além disso, restrições lineares de soma zero, são às vezes impostas para separar efeitos de vários componentes em 3.3.

Dito isto, a questão que surge é qual é a forma das prioris Gaussianas para que todos os efeitos sejam levados em consideração, pode-se pensar nisso no sentido de como montar o preditor estruturado aditivamente. Nesta dissertação o foco será em modelos que levem em consideração efeitos espaciais e/ou temporais através de Campos Aleatórios Gaussianos Markovianos (CAMG), conforme descrito na próxima seção.

3.2 Modelos Gaussianos latentes para processos espaço-temporais

Esta seção está fundamentada no artigo de Knorr-Held (2000), onde o autor propõe diversas estratégias para construir modelos Bayesianos com estrutura espaço-temporal não-separável para o mapeamento de risco de doenças. O objetivo aqui é extender os modelos em Knorr-Held (2000) para a situação geral onde a distribuição da resposta pertence a família exponencial, e mostrar como o preditor pode ser estruturado para levar em consideração efeitos espaciais, temporais e interações espaço-temporais.

3.2.1 Modelos com efeitos principais

Denote por y_{it} a variável resposta na área i $(i = 1,\ldots,n)$ e no tempo t $(t = 1,\ldots,T)$. Assuma que a distribuição de y_{it} tem distribuição de probabilidade pertencente a família exponencial, com termo de média μ_{it} e possivelmente um parâmetro de escala ou dispersão ϕ que pode ou não depender do parâmetro de média. Assumindo propriedades de independência condicional a verossimilhança é simplesmente o produto da distribuição de probabilidade atribuída para a resposta. Assim tem-se, $g(\mu_{it}) = \eta_{it}$, onde η_{it} tem a seguinte decomposição

$$\eta_{it} = \alpha + \rho_t + \gamma_t + \varphi_i + \phi_i \tag{3.4}$$

onde α é o nível médio do processo e ρ_t e γ_t são efeitos temporais, representando desvios da média geral do processo para o tempo t que respectivamente tem e não tem estrutura temporal a priori respectivamente. Similarmente, φ_i e ϕ_i representam desvios da média geral para a área i que respectivamente possui e não possui estrutura espacial.

A formulação 3.4 é completada definindo as distribuições a priori para os quatro blocos $\rho = (\rho_1,\ldots,\rho_T)^T$, $\gamma = (\gamma_1,\ldots,\gamma_T)^T$, $\varphi = (\varphi_1,\ldots,\varphi_n)^T$ e $\phi = (\phi_1,\ldots,\phi_n)^T$ em η_{it}. Cada priori é assumida como sendo gaussiana multivariada com média zero e matriz de precisão kK, onde k é um escalar desconhecido e K é uma matriz de estrutura conhecida. A matriz de estrutura K vai ser diferente para cada bloco a fim de descrever diferentes suposições a priori sobre o relacionamento entre os parâmetros dentro de cada bloco. Para ρ, assume-se a priori que os efeitos de pontos vizinhos no tempo tendem a ser semelhantes. O mais simples dos modelos dinâmicos é o passeio aleatório (*random walk*) com incrementos gaussianos independentes

$$\pi(\rho|k_\rho) \propto \exp\left(-\frac{k_\rho}{2}\sum_{t=2}^{T}(\rho_t - \rho_{t-1})^2\right) \tag{3.5}$$

que tem matriz de estrutura

$$K_\rho = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \vdots & \vdots & \vdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}$$

ver por exemplo Rue e Held (2005). Esta referência descreve outras possíveis alternativas, como o passeio aleatório de segunda ordem, que segundo Clayton e Bernardinelli (1987), deve ser preferido, se um dos interesses é predizer o processo em tempos futuros. Para γ, assume-se permutabilidade (*exchangeability*) dos componentes tomando $K_\gamma = I$, a matriz identidade.

Para o bloco estruturado espacialmente φ, escolhe-se uma simples autoregressão Gaussiana; ver, por exemplo, Besag, York e Mollié (1991). Assim, a matriz de estrutura K_φ tem elementos fora da diagonal $k_{ij} = -1$ para áreas geográficamente conectadas (vizinhas) $i \sim j$ e elementos na diagonal k_{ii} igual ao número de áreas, m_i, que são geograficamente contíguas a área i. Todos os outros elementos de K_φ são zero. A priori para φ pode ser escrita como

$$\pi(\varphi|k_\varphi) \propto \exp\left(-\frac{k_\varphi}{2}\sum_{i\sim j}(\varphi_i - \varphi_j)^2\right) \tag{3.6}$$

Esta priori de campo aleatório Markoviano é o analogo espacial do passeio aleatório sendo assim, é não estacionário. Esta abordagem pode ser extendida introduzindo pesos na formulação a priori (BESAG; YORK; MOLLIé, 1991). Finalmente, heterogeneidade espacial não estruturada é acomodada por tomar $K_\phi = I$. Uma representação simbólica do modelo de efeitos principais é dada na figura 3.1. Em que círculos vazados representam indepência a priori, retângulos representam dependência a priori. Observações no espaço tempo são indicadas por círculos sólidos.

3.2.2 Especificando prioris para efeitos de interação

A formulação anterior, considera efeitos temporais e espaciais separados, e requer uma extensão apropriada para a presença de interações espaço-tempo. Formalmente adiciona-se um parâmetro de interação δ_{it}, $i = 1,\ldots,n$ e $t = 1,\ldots,T$, o preditor agora é dado por:

$$\eta_{it} = \alpha + \rho_t + \gamma_t + \varphi_i + \phi_i + \delta_{ij}. \tag{3.7}$$

O vetor de parâmetros $\delta = (\delta_{11},\ldots,\delta_{nT})^T$ é assumido como Gaussiano com matriz de precisão $k_\delta K_\delta$. Como para os efeitos principais, k_δ é um escalar desconhecido e K_δ é uma matriz de estrutura pré-especificada. Note que 3.7 se reduz a 3.4 se todos os $\delta_{ij} = 0$, assim δ captura somente a variação que não pode ser explicada pelos efeitos principais.

Clayton (2004) sugere que K_δ seja especificada como o produto Kronecker das

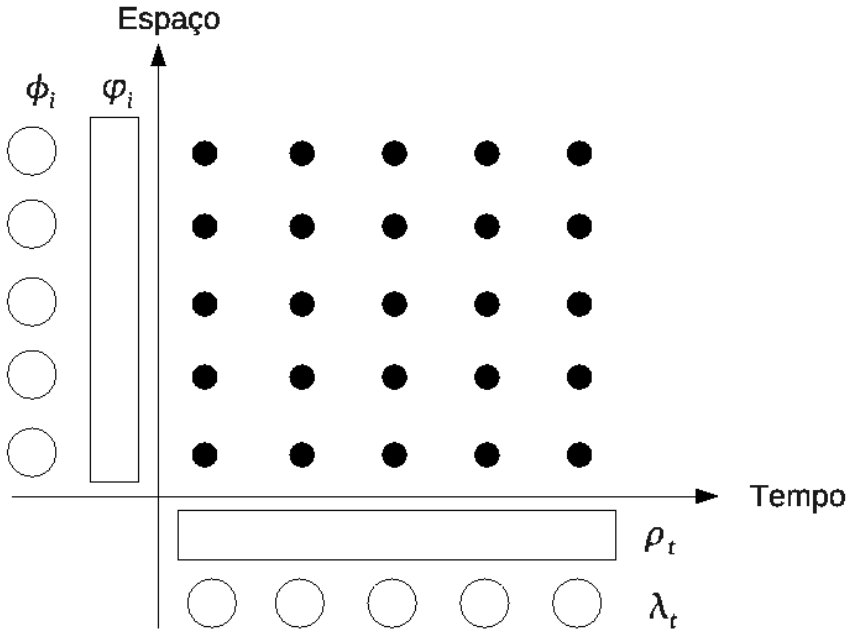

Figura 3.1: Representação simbólica do modelo de efeito principais. Círculos representam independência a priori, retângulos representam dependência a priori. Observações no tempo espaço são indicadas por círculos sólidos.

matrizes de estruturas dos efeitos principais que se assume interagir. Esta formulação pode ser vista como o análogo Bayesiano, da modelagem de interações por *tensor products* no contexto de regressão por splines (STONE et al., 1997). Na formulação proposta acima, $2*2=4$ combinações são possíveis dependendo de qual dos dois efeitos temporais assume-se interagir com qual dos dois efeitos espaciais. Estes quatro tipos de interação implicam em diferentes prioris para o relacionamento entre os δ_{it}, como ilustrado na figura 3.2. Neste momento, se discutirá cada tipo separadamente, ordenados pelo grau de dependência a priori.

- **Interação tipo I** - Se os dois efeitos principais não estruturados γ e ϕ são esperados interagirem, pela proposta de Clayton (2004), tem-se $K_\delta = K_\gamma \otimes K_\phi = I \otimes I$, assim todos os parâmetros de interação δ_{it} são a priori independentes:

$$\pi(\delta|k_\delta) \propto \exp\left(-\frac{k_\delta}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}(\delta_{it})^2\right) \tag{3.8}$$

  Este efeito pode ser visto como de covariáveis não observadas para cada pixel (i,t), que não tem nenhuma estrutura no espaço-tempo.

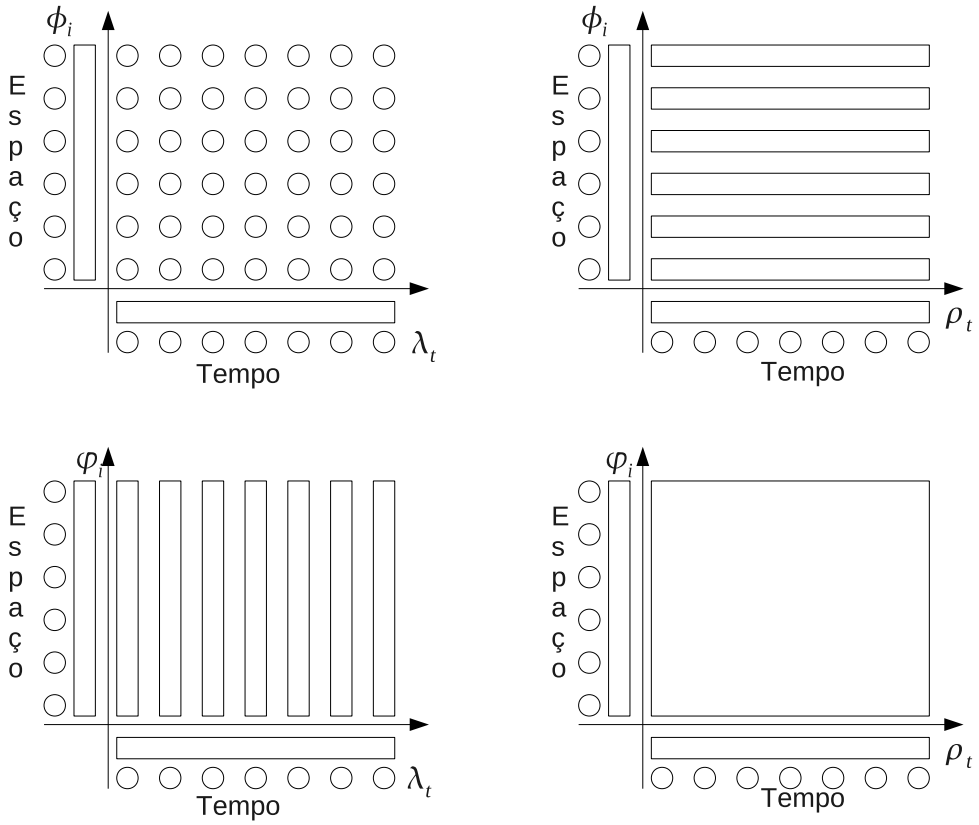

Figura 3.2: Representação simbólica para os quatro tipos de interação. Círculos representam independência a priori, retângulos representam dependência a priori.

- **Interação tipo II** - Se combinar o efeito principal *random walk* $\boldsymbol{\rho}$ com o bloco não estruturado $\boldsymbol{\phi}$, então cada $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_i T)^T$, $i = 1, \ldots, n$ segue um *random walk*, independente para cada área, ou seja, cada área tem a sua própria série temporal, porém não existe dependência espacial entre as séries. A matriz de estrutura K_δ tem rank $n(T-1)$ e a priori para $\boldsymbol{\delta}$ pode ser escrita como

$$\pi(\boldsymbol{\delta}|k_\delta) \propto \exp\left(-\frac{k_\delta}{2}\sum_{i=1}^{I}\sum_{t=2}^{T}(\delta_{it} - \delta_{i,t-1})^2\right) \tag{3.9}$$

  O modelo 3.7 com $\boldsymbol{\delta}$ do tipo II vai ser adequado, se as tendências temporais são diferentes para cada área, mas não tem nenhuma estrutura no espaço.

- **Interação tipo III** - Ao assumir, que os efeitos principais $\boldsymbol{\gamma}$ e $\boldsymbol{\varphi}$ interagem, então cada $\boldsymbol{\delta}_t = (\delta_{1t}, \ldots, \delta_{nt})^T$, $t = 1, \ldots, T$, segue uma autoregressão intrinseca (indepen-

dente). O rank de K_δ é agora $(n-1)$ e a priori para δ pode ser escrita como

$$\pi(\delta|k_\delta) \propto \exp\left(-\frac{k_\delta}{2}\sum_{t=1}^{T}\sum_{i\sim j}(\delta_{it}-\delta_{jt})^2\right) \tag{3.10}$$

Tal especificação vai ser adequada, se tendências espaciais são difentes em cada ponto no tempo, porém sem nenhuma estrutura temporal.

- **Interação tipo IV** - Do ponto de vista teórico, a mais interessante forma de interação é com o produto de dois efeitos principais dependentes, o *random walk* ρ e a autoregressão intrinseca φ. Agora δ é completamente dependente sobre o espaço e tempo e já não pode ser fatorada em blocos independentes. Pode ser mostrado que a priori para δ nesta estrutura pode ser escrita como

$$\pi(\delta|k_\delta) \propto \exp\left(-\frac{k_\delta}{2}\sum_{t=2}^{T}\sum_{i\sim j}(\delta_{it}-\delta_{jt}-\delta_{i,t-1}+\delta_{j,t-1})^2\right) \tag{3.11}$$

com constraste independentes $\delta_{it}-\delta_{jt}-\delta_{i,t-1}+\delta_{j,t-1}$. Esta distribuição é invariante a transformações do tipo

$$\widetilde{\delta_{it}} = \delta_{it}+c_i, \qquad t=1,\ldots,T,$$

$$\widetilde{\delta_{it}} = \delta_{it}+c_t, \qquad i=1,\ldots,n,$$

para algumas constantes $c_1,\ldots,c_n$, $c_1,\ldots,c_T$ e consequentemente sua matriz de estrutura tem rank altamente deficiente K_δ tem rank $(n-1)(T-1)$.

A distribuição condicional para o pixel δ_{it}, dado todos os outros, pode ser derivada (KNORR-HELD, 2000) vindo de $K_\delta = K_\rho \otimes K_\varphi$ e tem média

$$\mu_{it} = \begin{array}{ll} \delta_{i,t+1}+\frac{1}{m_i}\sum_{j\sim i}\delta_{j,t+1} & t=1 \\ \delta_{i,t-1}+\frac{1}{m_i}\sum_{j\sim i}\delta_{jt}-\frac{1}{m_i}\sum_{j\sim i}\delta_{j,t-1} & t=T \\ \frac{1}{2}(\delta_{i,t-1}+\delta_{i,t+1})+\frac{1}{m_i}\sum_{j\sim i}\delta_{jt}-\frac{1}{2m_i}\sum_{j\sim i}(\delta_{j,t-1}+\delta_{j,t+1}) & t=2,\ldots,T-1 \end{array} \tag{3.12}$$

e precisão

$$\tau_{it} = \begin{array}{ll} mi_i k_\delta & t=1 \quad \text{ou} \quad t=T \\ 2m_i k_\delta & t=2,\ldots,T-1. \end{array} \tag{3.13}$$

Assim, a interação do tipo IV é um campo Markoviano, onde não somente vizinhos (primeira ordem) temporais ($\delta_{i,t-1}$ e/ou $\delta_{i,t+1}$) e espaciais ($\delta_{jt}, j\sim i$) entram na condicional completa para δ_{it}, mas também vizinhos de segunda ordem ($\delta_{j,t-1}$

e/ou $\delta_{i,t+1}$), ou seja, vizinhos espaciais de vizinhos temporais ou, equivalentemente, vizinhos temporais de vizinhos espaciais. Esta priori força com que a tendência temporal (em termos de diferenças de primeira ordem) para áreas vizinhas sejam parecidas. Equivalentemente, esta priori força com que a dependência espacial para pontos vizinhos no tempo (t-1,t+1) sejam parecidas. Isto pode ser melhor visto olhando a média condicional μ_{it}, que satisfaz ambos

$$\mu_{it} - \overline{\delta}_{i\sim} = \overline{\delta}_{\sim t} - \overline{\delta}_{\sim\sim}$$

e

$$\mu_{it} - \overline{\delta}_{\sim t} = \overline{\delta}_{i\sim} - \overline{\delta}_{\sim\sim}.$$

Aqui $\overline{\delta}_{i\sim}$ é a média dos vizinhos temporais, $\overline{\delta}_{\sim t}$ é a média dos vizinhos espaciais, e $\overline{\delta}_{\sim\sim}$ é a média dos vizinhos de segunda ordem. Tal modelo a priori vai ser adequado, se tendências temporais são diferentes de área para área, mas são mais parecidas para áreas vizinhas.

3.2.3 Hiperprioris

Para completar a especificação Bayesiana do modelo, falta designar as distribuições a priori para os hiperparâmetros dos campos Gaussianos. Os hiperparâmetros para o modelo 3.4 são k_ρ, k_γ, k_φ e k_ϕ, eles determinam a variação de cada bloco, e precisam ser estimados através dos dados. Adicionalmente, k_δ tem que ser estimado no modelo 3.7. Seguindo a proposta de Knorr-Held (2000) assume-se para todos estes parâmetros prioris Gamma próprias, $k \sim G(a,b)$, para evitar problemas com hiperprioris impróprias. Prioris Gamma são computacionalmente convenientes, porque a condicional completa de k vai também ser Gamma. Para todas as aplicações nesta dissertação, hiperprioris Gamma de alta dispersão foram escolhidas para todos os blocos com valores $a = 1$ e $b = 0.01$. Em uma segunda rodada, para estudar a sensibilidade dos modelos às prioris troca-se os valores para $a = b = 0.01$.

3.3 Inferência Bayesiana Aproximada via INLA para modelos Gaussianos latentes

Esta seção está baseada no artigo Rue, Martino e Chopin (2009), o texto segue de muito perto o original, enfatizando alguns pontos considerados importantes para o entendimento da nova metodologia.

Para simplificar a discussão seguinte, denote genericamente $\pi(.|.)$ como a densidade condicional de seus argumentos, e faça $\underline{x}$ ser todas as n variáveis Gaussianas η_i, α, $f^{(j)}$, β_k. A densidade $\pi(\underline{x}|\theta_1)$ é Gaussiana com média zero (assumida), matriz de precisão $Q(\theta_1)$ com hiperparâmetro θ_1. Denote por $N(\underline{x};\mu,\Sigma)$ a $N(\mu,\Sigma)$ densidade Gaussiana com média μ e covariância (inverso da precisão) Σ na configuração $\underline{x}$. Note que foi incluído η_i ao invés de ε_i em $\underline{x}$, isto simplificará a notação no seguir do texto.

A distribuição para as n_d variáveis observadas $\underline{y} = y_i : i \in I$ é denotada por $\pi(\underline{y}|\underline{x},\theta_2)$ e assume-se que $y_i : i \in I$ são condicionalmente independentes dado $\underline{x}$ e θ_2. Este ponto é muito importante, pois ao assumir independência condicional a $\pi(\underline{y}|\underline{x},\theta_2)$ se resume a um produto da distribuição assumida para a variável resposta, ou seja, a verossimillhança. Para simplificar, denote por $\underline{\theta} = (\theta_1^T, \theta_2^T)^T$ com $dim(\underline{\theta}) = m$. A posteriori para uma matriz $Q(\underline{\theta})$ não singular, fica dada por

$$\begin{aligned}
\pi(\underline{x},\underline{\theta}|\underline{y}) &\propto \pi(\underline{\theta})\pi(\underline{x}|\underline{\theta})\prod_{i\in I}\pi(y_i|x_i,\underline{\theta}) \\
&\propto \pi(\underline{\theta})\|Q(\underline{\theta})\|^{n/2}\exp\left(-\frac{1}{2}\underline{x}^T Q(\underline{\theta})\underline{x} + \sum_{i\in I}\log\pi(y_i|x_i,\underline{\theta})\right). \qquad (3.14)
\end{aligned}$$

A imposição de restrições lineares (se necessário) são denotadas por $A\underline{x} = e$ para uma matriz A $k \times n$ de rank k. O objetivo principal é aproximar as marginais a posteriori $\pi(x_i|\underline{y})$, $\pi(\underline{\theta}|\underline{y})$ e $\pi(\theta_j|\underline{y})$. Muitos, mas não todos os modelos Gaussianos Latentes na literatura satisfazem duas propriedades básicas que serão assumidas através desta dissertação. A primeira é que o campo latente $\underline{x}$, que usualmente é de grande dimensão, admite propriedades de independência condicional. Assim, o campo latente é um Campo Aleatório Markoviano Gaussiano (CAMG) com uma matriz de precisão $Q(\underline{\theta})$ (RUE; HELD, 2005). Isto significa que pode-se usar métodos numéricos para matrizes esparsas, que em geral são muito mais rápidos que os métodos gerais para matrizes densas (RUE; HELD, 2005). A segunda propriedade é que o número de hiperparâmetros m, é pequeno, $m \leq 6$. Ambas propriedades são usualmente necessárias para produzir inferência rápida, embora existam excessões (EIDSVIK; MARTINO; RUE, 2009).

A abordagem INLA trabalha usando o fato que a marginal a posteriori de interesse pode ser escrita como

$$\pi(x_i|\underline{y}) = \int \pi(x_i|\underline{\theta},\underline{y})\pi(\underline{\theta}|\underline{y})\partial\underline{\theta} \quad \text{e} \quad \pi(\theta_j|\underline{y}) = \int \pi(\underline{\theta}|\underline{y})\partial\underline{\theta}_{-j}. \tag{3.15}$$

O fato chave da abordagem é usar esta forma para construir aproximações aninhadas,

$$\widetilde{\pi}(x_i|\underline{y}) = \int \widetilde{\pi}(x_i|\underline{\theta},\underline{y})\widetilde{\pi}(\underline{\theta}|\underline{y})\partial\underline{\theta} \quad \text{e} \quad \widetilde{\pi}(\theta_j|\underline{y}) = \int \widetilde{\pi}(\underline{\theta}|\underline{y})\partial\underline{\theta}_{-j}. \tag{3.16}$$

Aqui, $\widetilde{\pi}(.|.)$ é a densidade (condicional) aproximada de seus argumentos. Aproximações para $\pi(x_i|\underline{y})$ são calculadas aproximando $\pi(\underline{\theta}|\underline{y})$ e $\pi(x_i|\underline{\theta},\underline{y})$, e usando integração numérica (soma finita) para integrar fora $\underline{\theta}$. A integração é possível quando a dimensão de $\underline{\theta}$ é pequena, em geral menor ou igual a 6. Como vai ficar claro no decorrer do texto. A abordagem aninhada torna a aproximação de Laplace muito acurada quando aplicada para modelos Gaussianos latentes. A aproximação de $\pi(\theta_j|\underline{y})$ é calculada integrando fora θ_{-j} vindo de $\widetilde{\pi}(\underline{\theta}|\underline{y})$; este ponto será retomado na sequência para detalhes práticos.

A abordagem INLA é baseada na seguinte aproximação $\widetilde{\pi}(\underline{\theta}|y)$ para a marginal posteriori de $\underline{\theta}$.

$$\widetilde{\pi}(\underline{\theta}|y) \propto \left.\frac{\pi(\underline{x},\underline{\theta},\underline{y})}{\widetilde{\pi}_G(\underline{x}|\underline{\theta},\underline{y})}\right|_{\underline{x}=\underline{x}^*(\underline{\theta})} \tag{3.17}$$

onde, $\widetilde{\pi}_G(\underline{x}|\underline{\theta},\underline{y})$ é a aproximação Gaussiana para a condicional completa de $\underline{x}$ o que caracteriza a aproximação como de Laplace, e $\underline{x}^*(\underline{\theta})$ é a moda para a condicional completa de $\underline{x}$, para um dado $\underline{\theta}$. Note que a expressão 3.17 só é válida em um ponto, e portanto, para obter a aproximação da distribuição completa esta expressão precisa ser avaliada para um dado conjunto de $\underline{\theta}$. A proporcionalidade em 3.17 segue do fato que a constante normalizadora para $\pi(\underline{x},\underline{\theta}|\underline{y})$ é desconhecida. Note que $\widetilde{\pi}(\underline{\theta}|\underline{y})$ tende a ser bastante diferente da Gaussiana. Isto sugere que a aproximação Gaussiana direta para $\pi(\underline{\theta}|\underline{y})$ não é acurada o bastante. Um aspecto crítico da abordagem INLA é explorar e manipular $\widetilde{\pi}(\underline{\theta}|\underline{y})$ e $\widetilde{\pi}(x_i|\underline{y})$ de uma forma não paramétrica. Rue e Martino (2007) usaram 3.17 para aproximar marginais posteriori para vários modelos Gaussianos latentes. Suas conclusões foram que $\widetilde{\pi}(\underline{\theta}|\underline{y})$ é particularmente acurada, mesmo rodando um longo MCMC não puderam detectar nenhum erro nesta aproximação. Para as marginais posteriori do campo latente, eles propõem começar pela $\widetilde{\pi}_G(\underline{x}|\underline{\theta},\underline{y})$, ou seja,

$$\widetilde{\pi}(x_i|\underline{\theta},\underline{y}) = N(x_i;\mu_i(\underline{\theta}),\sigma_i^2(\underline{\theta})). \tag{3.18}$$

Aqui, $\mu(\underline{\theta})$ é a média (vetor) para a aproximação Gaussiana, considerando que $\sigma^2(\underline{\theta})$

é o vetor correspondente de variâncias marginais. Esta aproximação pode ser integrada numericamente com respeito a $\underline{\theta}$, ver 3.16,para obter aproximações para as marginais de interesse do campo latente,

$$\widetilde{\pi}(x_i|\underline{y}) = \sum_k \widetilde{\pi}(x_i|\underline{\theta}_k, \underline{y}) \times \widetilde{\pi}(\underline{\theta}_k|\underline{y}) \times \Delta_k. \tag{3.19}$$

A soma é sobre os valores de $\underline{\theta}$ com pesos Δ_k. Rue e Martino (2007) mostraram que a marginal posteriori para $\underline{\theta}$ foi acurada, enquanto o erro na aproximação Gaussiana 3.18 foi grande. Em particular, 3.18 pode apresentar erro na locação e/ou falta de assimetria (*skewness*). As dificuldades de Rue e Martino (2007) foram em detectar os $x_i's$ nas quais a aproximação era menos acurada e a falta de abilidade para melhorar a aproximação nestas localizações. Além disso, eles não conseguiam controlar o erro da aproximação e escolher os pontos de integração θ_k de uma forma adaptativa e automatica. Rue, Martino e Chopin (2009) resolveram os problemas em Rue e Martino (2007), e apresentam uma abordagem completa para inferência aproximada em modelos Gaussianos latentes que eles nomearam de *Integrated Nested Laplace Approximations*(INLA). A principal mudança é aplicar mais uma vez a aproximação de Laplace, desta vez para $\pi(x_i|\underline{y}, \underline{\theta})$. Eles também apresentam uma alternativa rápida que corrige a aproximação Gaussiana 3.18 para o erro de locação e falta de assimetria a um custo extra moderado. Estas correções são obtidas por uma expansão em séries de Laplace. Esta alternativa rápida é uma primeira escolha natural, porque é de baixo custo computacional e de alta acurácia. Os autores demonstram como várias aproximações podem ser usadas para derivar ferramentas para testar a aproximação, aproximar marginais a posteriori para um subconjunto de $\underline{x}$, e calcular várias medidas de interesse, tais como, verossimilhança marginal, Critério de Informação da *Deviance* (DIC) e várias medidas preditivas Bayesianas. Antes de apresentar a proposta de Rue, Martino e Chopin (2009) é interessante apresentar de forma rápida a aproximação Gaussiana.

3.4 Aproximações Gaussianas

A abordagem INLA é baseada em aproximações Gaussianas para densidades da forma:

$$\pi(\underline{x}) \propto \exp\left(-\frac{1}{2}\underline{x}^T Q \underline{x} + \sum_{i \in I} g_i(x_i)\right) \tag{3.20}$$

onde $g_i(x_i)$ é $\log \pi(y_i|x_i, \underline{\theta})$. A aproximação Gaussiana $\widetilde{\pi}_g(\underline{x})$ é obtida encontrando a moda e a curvatura na moda. A moda é calculada iterativamente usando o método de Newton Raphson, também conhecido como o algoritmo escore e sua variante o escore de Fisher (FAHRMEIR; TUTZ, 2001). Seja $\underline{\mu}^{(0)}$ um valor inicial, expande-se $g_i(x_i)$ em torno de $\mu_i^{(0)}$ até o termo de segunda ordem

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2} c_i x_i^2 \tag{3.21}$$

onde b_i e c_i depende de $\mu^{(0)}$. A aproximação Gaussiana é obtida, com matriz de precisão $Q + diag(c)$ e moda dada pela solução de $(Q + diag(c))\underline{\mu}^{(1)} = \underline{b}$. Este processo é repetido até a convergência para a distribuição Gaussiana com, média $\underline{x}^*$ e matriz de precisão $Q^* = Q + diag(c^*)$. Como o termo não quadrático em 3.21 é somente função de x_i e não uma função de x_i e x_j, então a matriz de precisão para a aproximação Gaussiana é da forma $Q + diag(c)$. Isto é computacionalmente conveniente, porque as propriedades Markovianas do CAMG são preservadas. Existem outras sugestões na literatura de como construir aproximações Gaussianas, ver (RUE, 2001), (RUE; HELD, 2005) e (KUSS; RASMUSSEN, 2005).

3.5 Integração aproximada aninhada de Laplace

Nesta seção será apresentada a abordagem INLA para aproximar marginais a posteriori para campos latentes Gaussianos, $\pi(x_i|\underline{y})$, $i = 1, \ldots, n$. A aproximação é feita em três passos. O primeiro passo aproxima a marginal a posteriori de $\underline{\theta}$ usando a aproximação de Laplace 3.17. O segundo passo calcula a aproximação de Laplace, ou a aproximação de Laplace simplificada, para $\pi(x_i|\underline{y}, \underline{\theta})$, para valores selecionados de $\underline{\theta}$, a fim de melhorar a aproximação Gaussiana 3.18. O terceiro passo combina os dois anteriores usando integração numérica 3.19.

Cabe ressaltar aqui que toda a seção foi escrita usando o artigo Rue, Martino e Chopin (2009), sendo uma revisão deste onde alguns pontos considerados importantes, são explicitados para facilitar a leitura e entendimento. Ainda, antes de descrever propriamente a abordagem INLA, será apresentado a metodologia como um todo através de um simples algoritmo, passo-a-passo. Considere,

1. Selecione um conjunto de $\Theta = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$

2. Para $k = 1$ até K faça

3. Calcule $\widetilde{\pi}(\underline{\theta}_k|\underline{y})$

4. Calcule $\widetilde{\pi}(x_i|\underline{\theta}_k,\underline{y})$ como uma função de x_i

5. Fim para

6. Calcule $\widetilde{\pi}(x_i|\underline{y}) = \sum_k \widetilde{\pi}(x_i|\underline{\theta}_k,\underline{y})\widetilde{\pi}(\underline{\theta}_k|\underline{y})\Delta_k$

O algoritmo começa selecionando um conjunto possívelmente pequeno de $\underline{\theta}'s$. O procedimento para selecionar este conjunto é descrito na seção 3.5.1, e é feito explorando a distribuição $\widetilde{\pi}(\underline{\theta}|\underline{y})$.

Após ter os vetores duas aproximações são calculadas, a primeira é exatamente a distribuição $\widetilde{\pi}(\underline{\theta}|\underline{y})$ que é calculada como na equação 3.17. Para a segunda aproximação existem três possibilidades.

A primeira e mais barata computacionalmente é a Gaussiana que é explicada na seção 3.4 e retomada na seção 3.5.3, a segunda e mais acurada é a aproximação de Laplace que é explicada na seção 3.5.3. Como ficará claro no decorrer do texto o último passo que consiste na integração numérica é bastante facilitado pelo procedimento explicado na seção 3.5.1 induzir que todos os pesos Δ_k sejam iguais, o que torna o último passo simplesmente uma soma com todos os pesos iguais. Esta é uma descrição bastante geral da abordagem INLA, e os detalhes seguem nas próximas seções.

Para o cálculo da aproximação $\widetilde{\pi}(x_i|\underline{\theta}_k,\underline{y})$ Rue, Martino e Chopin (2009) propõem também uma terceira possibilidade denominada de Aproximação de Laplace Simplicada, esta abordagem não será considerada neste texto.

3.5.1 Explorando $\widetilde{\pi}(\theta|y)$

O primeiro passo da abordagem INLA é calcular uma aproximação para a posteriori marginal de $\underline{\theta}$, ver 3.17. O denominador em 3.17 é a aproximação Gaussiana para a condicional completa de $\underline{x}$, e é calculado como descrito na seção 3.4. O principal uso de $\widetilde{\pi}(\underline{\theta}|\underline{y})$ é integrar fora a incerteza com respeito a $\underline{\theta}$ quando aproximando a marginal a posteriori de x_i, ver 3.19. Para fazer isto, não é necessário representar $\widetilde{\pi}(\underline{\theta}|\underline{y})$ paramétricamente, mas basta explorá-la suficientemente bem para encontrar bons pontos para o cálculo da integração numérica. Até o fim desta seção, será discutido como as marginais a posteriori $\pi(\theta_j|\underline{y})$ podem ser aproximadas. Assuma por simplicidade que $\underline{\theta} = (\theta_1,\ldots,\theta_m) \in \Re^m$, que pode sempre ser obtido usando uma reparametrização.

- **Passo 1** Localize a moda de $\widetilde{\pi}(\underline{\theta}|\underline{y})$, otimizando $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ com respeito a $\underline{\theta}$. Isto pode ser feito usando algum método quasi-Newton que vai usar alguma aproximação da derivada de segunda ordem de $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ usando sucessivas diferenças entre vetores gradientes. O gradiente é aproximado usando diferenças finitas. Denote por $\underline{\theta}^*$ como sendo a configuração modal.

- **Passo 2** Na configuração modal $\underline{\theta}^*$ calcule a matriz Hessiana $H > 0$, usando diferença finita. Seja $\Sigma = H^{-1}$, que vai ser a matriz de covariância para $\underline{\theta}$ se a densidade fosse Gaussiana. Para facilitar a exploração, use variáveis padronizadas z ao invés de $\underline{\theta}$: Seja $\Sigma = V\Lambda V^T$ a decomposição em autovalores e autovetores de Σ, e defina θ via z, como segue

$$\theta(z) = \underline{\theta}^* + V\Lambda^{1/2}z. \tag{3.22}$$

  Se $\widetilde{\pi}(\underline{\theta}|\underline{y})$ é a densidade Gaussiana, então z é $N(0,I)$. Esta reparametrização corrige a escala e rotação, e simplifica a integração numérica, ver (SMITH et al., 1987).

- **Passo 3** Explore $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ usando a z-parametrização. A figura 3.5.1 ilustra o procedimento quando $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ é unimodal. O painel (a) mostra um gráfico de contorno para $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ para $m = 2$, a localização da moda e o novo eixo de coordenada para z. Deve-se explorar $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ com o objetivo de encontrar pontos com a maior massa de probabilidade. O resultado deste processo é mostrado no painel (b). Cada sinal é um ponto onde $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$ é considerado como significante, e que será usado na integração numérica 3.19. Detalhes são como segue. Começa-se vindo da moda ($z = 0$), e caminha-se na direção positiva de z_1 com passos de tamanho δ_z por exemplo $\delta_z = 1$, contanto que

$$\log\widetilde{\pi}(\underline{\theta}(0)|\underline{y}) - \log\widetilde{\pi}(\underline{\theta}(z)|\underline{y}) < \delta_\pi \tag{3.23}$$

  onde, por exemplo $\delta_\pi = 2.5$. Então troca-se a direção e procede-se similarmente. As outras coordenadas são tratadas da mesma forma. Isto produz os sinais pretos. Pode-se agora preencher todos os valores intermediários tomando-se todas as diferentes combinações do sinais pretos. Estes novos sinais (mostrados em cinza) são incluídos se 3.23 for satisfeita. Como o *layout* dos pontos $\underline{\theta}_k$ é uma grade regular, pode-se tomar todos os pesos Δ_k em 3.19 como sendo iguais.

Em (a) a moda é localizada, a Hessiana e o sistema de coordenadas z são calculados. Em (b) cada coordenada é explorada (pontos pretos) até um certo limite da log-densidade. Finalmente os pontos cinza são explorados.

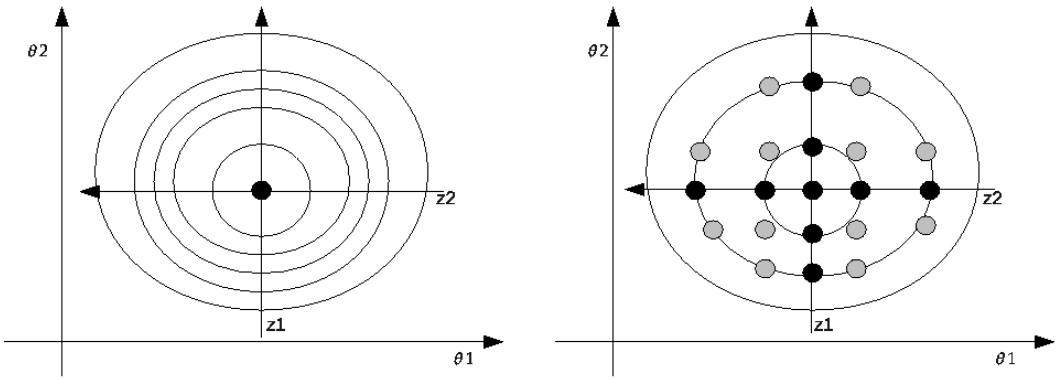

Figura 3.3: Ilustração da exploração da marginal posteriori para $\boldsymbol{\theta}$.

3.5.2 Aproximando $\pi(\theta_j|y)$.

Marginais a posteriori para θ_j podem ser obtidas diretamente vindo de $\widetilde{\pi}(\underline{\theta}|\underline{y})$ usando integração numérica. Entretanto, isto é computacionalmente caro, porque é necessário calcular $\widetilde{\pi}(\underline{\theta}|\underline{y})$ para um grande número de configurações. Uma abordagem mais factível é usar os pontos obtidos durantes os passos 1-3 para construir um interpolante para $\log\widetilde{\pi}(\underline{\theta}|\underline{y})$, e calcular marginais usando integração numérica vinda deste interpolante. Se uma alta acurácia é necessária, na prática basta obter uma configuração mais densa, por exemplo $\delta_z = 1/2$ ou $1/4$.

3.5.3 Aproximando $\pi(x_i|\theta, y)$

Até aqui tem-se um conjunto de pontos ponderados $\underline{\theta}_k$ para ser usado na integração 3.19. O próximo passo é providenciar aproximações acuradas para a posteriori marginal dos x_i's condicionado nos valores selecionados de $\underline{\theta}$. Nesta seção são discutidas, duas aproximações para $\widetilde{\pi}(x_i|\underline{y}, \theta_k)$, a Gaussiana e a de Laplace.

Usando aproximação Gaussiana

Uma aproximação simples e barata computacionalmente para $\pi(x_i|\underline{\theta}, \underline{y})$ é a aproximação Gaussiana $\widetilde{\pi}_G(x_i|\underline{\theta}, \underline{y})$, onde a média $\mu_i(\underline{\theta})$ e a variância marginal $\sigma_i^2(\underline{\theta})$ são derivadas usando propriedades de Campos Aleatórios Markovianos Gaussianos, e algoritmos para matrizes esparsas (RUE; MARTINO, 2007). Durante a exploração de $\widetilde{\pi}(\underline{\theta}|\underline{y})$, ver seção 3.4, já foi calculada $\widetilde{\pi}(x|\underline{\theta}, \underline{y})$, assim só variâncias marginais precisam ser calculadas. A aproximação Gaussiana apresenta resultados razoáveis, mas pode apresentar

erros na locação e/ou erros devido a falta de assimetria (RUE; MARTINO, 2007).

Usando aproximação de Laplace

A forma natural de melhorar a aproximação Gaussiana é calcular a aproximação de Laplace

$$\widetilde{\pi}_{LA}(x_i|\underline{\theta},\underline{y}) \propto \left.\frac{\pi(x,\underline{\theta},\underline{y})}{\widetilde{\pi}_{GG}(\underline{x}_{-i}|x_i,\underline{\theta},\underline{y})}\right|_{\underline{x}_{-i}=\underline{x}^*_{-i}(x_i,\underline{\theta})} \tag{3.24}$$

Aqui, $\widetilde{\pi}_{GG}$ é a aproximação Gaussiana para $x_{-i}|x_i,\theta,\underline{y}$, e x^*_{-i} é a configuração modal. Note que $\widetilde{\pi}_{GG}$ é diferente da densidade condicional correspondente a $\widetilde{\pi}_G(x|\underline{\theta},\underline{y})$. Infelizmente, 3.24 implica que $\widetilde{\pi}_{GG}$ precisa ser recalculada para cada valor de x_i e $\underline{\theta}$, uma vez que sua matriz de precisão depende de x_i e $\underline{\theta}$. Isto é muito caro, porque requer n fatorizações da matriz de precisão completa. Rue, Martino e Chopin (2009) propõem duas modificações em 3.24 que tornam isto factível computacionalmente. A primeira modificação consiste em evitar o passo de otimização ao calcular $\widetilde{\pi}_{GG}(x_{-i}|x_i,\underline{\theta},\underline{y})$ aproximando a configuração modal,

$$\underline{x}^*_{-i}(x_i,\underline{\theta}) \approx E_{\widetilde{\pi}_G}(\underline{x}_{-i}|x_i). \tag{3.25}$$

O lado direito é calculado sob a densidade condicional derivada da aproximação Gaussiana $\widetilde{\pi}_G(x|\underline{\theta},\underline{y})$. O benefício computacional é imediato. Primeiro, a média condicional pode ser calculada por uma atualização *rank-one* vindo da média incondicional. Outro fato positivo é que a média condicional é contínua com respeito a x_i, que não é o caso quando otimização numérica é usada para calcular $x^*_i(x_i,\underline{\theta})$. A segundo modificação materializa a seguinte intuição: somente aqueles x_j que estão 'próximos' de x_i vão ter impacto na marginal de x_i. Se a dependência entre x_j e x_i decai com o aumento da distância entre i e j, somente aqueles x_j's em uma 'região de interesse' acerca de i, $R_i(\theta)$, determinam a marginal de x_i. A esperança condicional em 3.25 implica que

$$\frac{E_{\widetilde{\pi}_G}(x_j|x_i)-\mu_j(\underline{\theta})}{\sigma_j(\underline{\theta})} = a_{ij}(\underline{\theta})\frac{x_i-\mu_i(\underline{\theta})}{\sigma_i(\underline{\theta})} \tag{3.26}$$

para alguma $a_{ij}(\underline{\theta})$ quando $i \neq j$. Assim, uma simples regra para construir o conjunto $R_i(\underline{\theta})$ é

$$R_i(\underline{\theta}) = j : |a_{ij}(\underline{\theta})| > 0.001. \tag{3.27}$$

A importância computacional de usar $R_i(\underline{\theta})$ segue do cálculo do denominador de 3.24, onde agora só é necessário fatorar uma $|R_i(\underline{\theta})| \times |R_i(\underline{\theta})|$ matriz esparsa. A expressão 3.24, simplificada como explicado acima, precisa ser calculada para diferentes valores de x_i para encontrar a densidade. Para selecionar estes pontos, usa-se a média e variância da aprox-

imação Gaussiana 3.18, e combina-se, diferentes valores para as variáveis padronizadas

$$x_i^{(s)} = \frac{x_i - \mu_i(\underline{\theta})}{\sigma_i(\underline{\theta})} \tag{3.28}$$

de acordo com as correspondentes escolhas de abscissas dadas pela regra de quadratura de Gauss-Hermite. Para representar a densidade $\widetilde{\pi}_{LA}(x_i|\underline{\theta},\underline{y})$, usa-se

$$\widetilde{\pi}_{LA}(x_i|\underline{\theta},\underline{y}) \propto N(x_i;\mu_i(\underline{\theta}),\sigma_i^2(\underline{\theta})) \times \exp\{\text{cubic spline}(x_i)\} \tag{3.29}$$

O spline cúbico é ajustado para a diferença entre a log-densidade $\widetilde{\pi}_{LA}(x_i|\underline{\theta},\underline{y})$ e $\widetilde{\pi}_G(x_i|\underline{\theta},\underline{y})$ nas abscisas selecionadas, e então a densidade é normalizada usando integração por quadratura.

3.6 Aproximando a Verossimilhança marginal

A verossimilhança marginal $\pi(\underline{y})$ é uma quantidade útil para comparar modelos, como o fator de Bayes é definido pela razão de verossimilhanças marginais entre dois modelos competidores.

$$\widetilde{\pi}(\underline{y}) = \int \frac{\pi(\underline{\theta},\underline{x},\underline{y})}{\widetilde{\pi}_G(\underline{x}|\underline{\theta},\underline{y})}\Bigg|_{\underline{x}=\underline{x}^*(\underline{\theta})} d\underline{\theta}. \tag{3.30}$$

onde $\pi(\underline{\theta},\underline{x},\underline{y}) = \pi(\underline{\theta})\pi(\underline{x}|\underline{\theta})\pi(\underline{y}|\underline{x},\underline{\theta})$. Uma alternativa, mais simples para estimar a verossimilhança marginal é obtida por assumir que $\underline{\theta}|\underline{y}$ é Gaussiana; então 3.30 torna-se uma quantidade conhecida vezes $|H|^{-1/2}$, onde H é a matriz Hessiana da seção 3.4. Este método pode falhar no caso em que a marginal posteriori $\pi(\underline{\theta}|\underline{y})$ é multi-modal, mas isto não é específico do cálculo de verossimilhança marginal mas se aplica a abordagem geral. Felizmente, modelos Gaussianos latentes geram distribuições posterioris unimodais em muitos casos.

3.7 Critério de informação da *Deviance*

O critério de informação da *Deviance* (SPIEGELHALTER et al., 2001) é um critério de informação popular para modelos hierárquicos, e (em muitos casos) é bem definido para prioris impróprias. Sua principal aplicação é a seleção de modelos bayesianos,

mas ele também providência uma noção do número efetivo de parâmetros, que está sendo usado no modelo. Neste contexto, a *deviance* é

$$D(\underline{x}, \underline{\theta}) = -2 \sum_{i \in I} \log \pi(y_i | x_i, \underline{\theta}) + const \tag{3.31}$$

DIC é definido como duas vezes a média da *deviance* menos a *deviance* para a média. O número efetivo de parâmetros é a média da *deviance* menos a *deviance* da média.

4 RESULTADOS

Neste Capítulo apresentam-se os resultados da aplicação da metodologia do Capítulo 3 aos problemas descritos no Capítulo 2. Os três problemas têm em comum o fato dos experimentos apresentarem estrutura espaço-temporal e portanto a adoção de modelos latentes que contemplem tais estruturas é uma escolha natural.

Por outro lado diferem na dimensão do conjunto de dados, tipo de distribuição designada à variável resposta, estrutura auxiliar de covariáveis e interesses principais na análise.

Não é o interesse principal desta dissertação, mas em alguns dos exemplos analisados os principais resultados obtidos com o ajuste via INLA serão contrastados com resultados provenientes de um Modelo Aditivo Generalizado (GAM) (WOOD, 2006), tais modelos encontram-se disponíveis no pacote *mgcv* (WOOD, 2008) do software R (R Development Core Team, 2009). Uma rápida revisão sobre tais modelos pode ser encontrada em Bonat et al. (2009).

No decorrer das análises foram utilizados os pacotes *tripack* (EGLEN; ZUYEV; WHITE, 2009) e *gpclib* (MURTA, 2009) na construção da tecelagem de voronoi no segundo exemplo. Para manipulação de objetos espaciais foram utilizados os pacotes *sp* (PEBESMA; BIVAND, 2005) e *spdep* (BIVAND et al., 2009).

Para comparação entre os modelos em todos os exemplos foram calculados o Critério de informação da *Deviance*, o número estimado de parâmetros e a verossimilhança marginal. Tais medidas foram usadas apenas como um guia geral para a escolha, ressalta-se que a verossimilhança e consequentemente o fator de Bayes deve ser olhado com cuidado quando a distribuição a priori considerada for imprópria.

4.1 Qualidade da água em reservatórios operados pela COPEL no estado do Paraná

Como mencionado no Capítulo 2, o principal objetivo da análise do índice de Qualidade da água em reservatórios operados pela COPEL no estado do Paraná, é avaliar o eventual efeito do reservatório sobre tal índice. Desta forma, foram coletadas amostras de água em três locais (montante, reservatório e jusante), além disso outros atributos (covariáveis) no experimento que podem afetar tal índice. O experimento ocorreu em diferentes pontos no tempo com coletas trimestrais, então pode-se esperar que exista algum efeito climático (temporal). Além disso, foram realizadas em diferentes usinas, tal condição pode influenciar no índice de qualidade da água, assim o processo de modelagem deve levar em consideração todas estas possibilidades, a fim de isolar apenas o efeito da covariável de interesse que são os locais de coleta. Além destes efeitos principais podem ocorrer interações entre por exemplo, trimestres com locais de coleta, tempo com locais de coleta e ainda tempo com usinas (UHEs) e locais de coleta.

Como forma inicial de investigar tais relações, a Figura 4.1 apresenta uma análise descritiva para o IQA após a transformação logística. Tal análise consiste da construção de diagramas *boxplot*, histograma da variável e gráfico de interação entre os trimestres do ano e as estações de monitoramento das diferentes usinas (UHEs), trazendo uma visão geral da série histórica monitorada. É importante notar que a escala gráfica apresenta os valores transformados da variável IQA após a transformação logística. Para uma avaliação estatística com vistas à qualidade da água, as categorias do IQA estão representadas nos gráficos por meio de linhas cheias. Apesar de todos os valores de IQA estarem incluídos na escala, os valores observados incluem apenas as classes BOM (IQA entre 51 e 79) e ÓTIMO(IQA entre 79 e 100).

É possível observar pela análise do *boxplot* (4.1B) que o IQA das estações reservatório, são sempre maiores do que o de jusante e montante. Em termos de qualidade da água, isto indica que a qualidade da água do reservatório, com base no IQA, é melhor do que a observada nas estações de montante e jusante. As medianas do IQA em cada uma das três estações encontram-se acima do limite da classe BOM, indicando águas de boa qualidade. As estações de reservatório e jusante se enquadram na classe ÓTIMO. Em

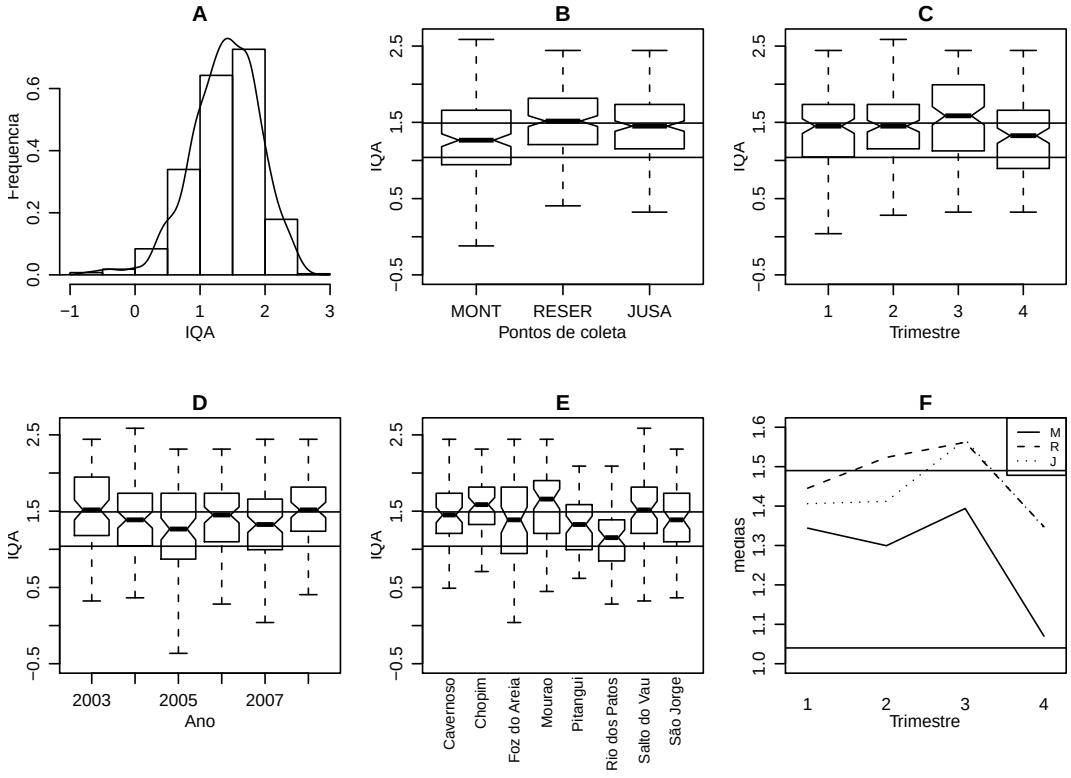

Figura 4.1: Análise descritiva para o Índice de Qualidade da Água. As linhas cheias representam os limites da classe Bom e Ótimo.

relação à variação temporal, não há uma diferença clara entre os anos de monitoramento, uma vez que o nível mediano do IQA é praticamente o mesmo, podendo-se ressaltar que nos anos de 2005 e 2006 foram observados os menores valores do índice (4.1D). Já em relação aos meses, é no terceiro trimestre que foram verificados os maiores valores de IQA (4.1C). A Figura 4.1E apresenta *boxplot's* para cada uma das usinas hidrolétricas consideradas na análise. É possível observar que os dados apresentam grande variabilidade tanto dentro das usinas (série temporal individual) como entre as usinas (diferenças entre as localidades). Dentre as UHEs, destaca-se a usina de Rio dos Patos que apresentou a menor mediana de IQA. No entanto, apesar da grande variabilidade dos dados, as medianas dos índices de todas as usinas encontram-se acima do valor da categoria BOM.

A fim de testar todos os efeitos principais e possíveis interações, foram ajustados uma série de modelos com crescente grau de complexidade, conforme Tabela 4.1. Na notação usada para apresentar os modelos, Y sempre representará a variável resposta, efeitos espaciais e temporais seguem a notação introduzida no Capítulo 3, efeitos fixos serão representados pela letra $\boldsymbol{\beta}$, seguida de um índice e, todos os modelos incluem o intercepto. Para possibilitar uma comparação e consequente escolha do modelo que melhor descreve a resposta, foram calculados o Critério de Informação da *Deviance* (DIC), o

número estimado de parâmetros e a verossimilhança marginal. Para esta aplicação sempre que possível foi ajustado também um modelo aditivo generalizado, adequado em cada situação. Para comparação dos resultados foi calculado também o Critério de Informação de *Akaike*. As medidas de ajuste de modelos são apresentadas na Tabela 4.1.

Tabela 4.1: Modelos ajustados, critério de informação da *Deviance*, número estimado de parâmetros, verossimilhança marginal e critério de informação de *Akaike*.

Modelos	Preditor Linear	DIC	NP	MV	AIC
1	$Y \sim 1$	$837,18$	$1,658$	$-427,32$	$837,18$
2	$Y \sim \beta_{uhe} + \beta_{loc}$	$797,86$	$10,66$	$-415,08$	$798,50$
3	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_t$	$755,72$	$24,01$	$-402,16$	$768,68$
4	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_{loc:t}$	$773,72$	$41,14$	$-563,16$	$784,85$
5	$Y \sim \beta_{uhe} + \beta_{loc} + \rho_{uhe:loc:t}$	$741,15$	$58,11$	$-440,12$	$---$

O modelo 1 apenas com o intercepto, foi ajustado apenas para servir de base para comparações. O modelo 2 leva em consideração efeitos das usinas (UHE) e dos locais de coleta, o modelo 3 complementa o 2 incluindo uma única série temporal, através de um modelo *random walk* de primeira ordem. O modelo 4 atende a intuição de que a série temporal de cada local de coleta pode ter um comportamento diferenciado, então uma série para cada local é ajustada. E finalmente o modelo 5 extende o 4 dizendo que as séries temporais podem ser diferentes não somente entre locais de coleta, mas também, entre usinas.

De acordo com os resultados apresentados na Tabela 4.1 o modelo que apresenta o menor DIC é o 5. Entretando, observando o número estimado de parâmetros e a verossimilhança marginal verifica-se que tal modelo pode estar superparametrizado, desta forma, opta-se pelo modelo 3 por este apresentar um baixo ***DIC***, um número estimado de parâmetros reduzido e a maior verossimilhança marginal. A mesma forma de preditor linear é escolhida pelo critério de informação de *Akaike* quando ajustando pelo modelo aditivo generalizado. A distribuição a posteriori para o intercepto e efeitos dos locais de coleta é apresentada na Figura 4.2.

A Tabela 4.2 apresenta o ajuste do modelo 3 pelas abordagens INLA através da média a posteriori, para o GAM pela estimativa pontual e o desvio padrão correspondente para ambos.

Como pode ser visto na Tabela 4.2 o ajuste pelos dois métodos é muito parecido, tanto em termos de médias como desvios padrão. Apenas para o intercepto é possível identificar diferença entre a estimativa do GAM e a média da posteriori.

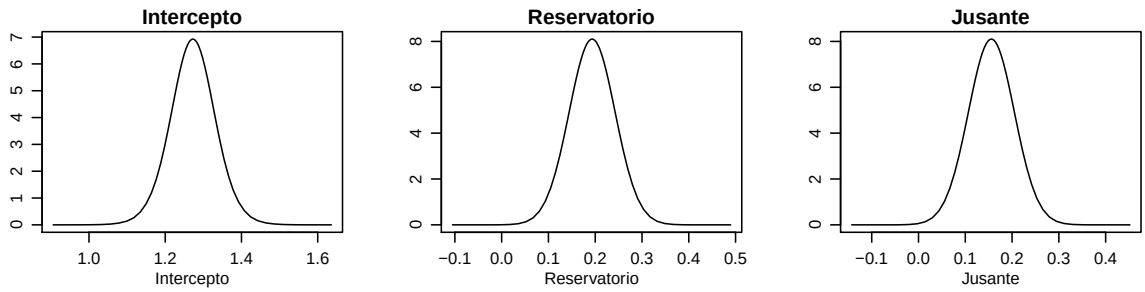

Figura 4.2: Distribuição a posteriori para o intercepto e efeitos dos locais de coleta.

Tabela 4.2: Resultados do modelo 3 via INLA e GAM.

Parâmetros	Média Posteriori	Desvio Padrão	Estimativa	Desvio Padrão
Intercepto	1,2725	0,06053	1,3398	0,06454
Reservatorio	0,1933	0,04930	0,1933	0,05016
Jusante	0,1557	0,04930	0,1557	0,05016

Pela parametrização usada no ajuste, a estação de montante é tomada como referência. Sendo assim, os valores estimados para os efeitos de reservatório e jusante, indicam a mudança que ocorre no índice de qualidade da água (transformado). Pelos dois métodos existe uma mudança significativa no índice de qualidade da água quando esta passa da estação de montante para o reservatório, sendo que o reservatório apresenta um efeito benéfico na qualidade da água. O mesmo tipo de efeito é observado quando se passa da estação de montante para a de jusante, porém com menor intensidade.

O modelo aditivo leva em consideração o efeito temporal como uma função suave das datas de coleta. Já pela abordagem INLA tal efeito é considerado através de um modelo *random walk* de primeira ordem. A fim de comparar como as duas abordagens captam tal efeito a Figura 4.3 sobrepõe os efeitos temporais estimados pelas duas abordagens.

Pela Figura 4.3 é possível verificar uma grande semelhança entre os efeitos temporais estimados pelos dois métodos. O modelo *random walk* apresenta variações mais abruptas que a função suave estimada pelo GAM. Também nota-se que as bandas de credibilidade obtidas pelo *random walk* são mais largas que as bandas de confiança obtidas pelo GAM. As bandas de credibilidade do *random walk* são calculadas via quantis e o intervalo de confiança do GAM é baseado na teoria assintótica usando os quantis da distribuição normal. Os dois intervalos possuem nível nominal de 95%.

Como forma de testar à sensibilidade do modelo a escolha das prioris, trocou-se os hiperparâmetros da distribuição gama dos parâmetros de precisão do modelo conforme discutido na Seção 3.2.3. As distribuições a posteriori para os parâmetros de interesse

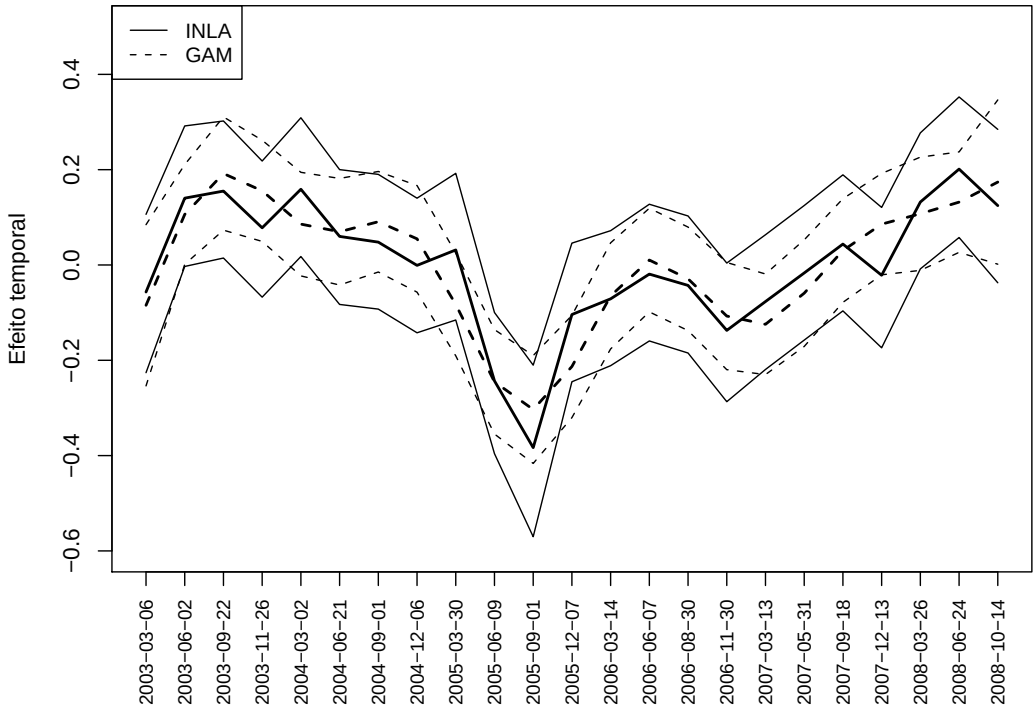

Figura 4.3: Sobreposição do efeito temporal ajustado pelo INLA e GAM.

do modelo nas duas configurações de prioris são sobrepostas na Figura 4.4. Optou-se por considerar as UHEs como um efeito aleatório não estruturado, sendo assim, este efeito também conta com um parâmetro de precisão.

Na Figura 4.4 pode-se observar que a escolha de uma priori mais vaga provocou mudanças nítidas na forma da distribuição a posteriori dos parâmetros de precisão do efeito temporal, e das UHEs. Porém em termos de moda a posteriori as mudanças foram pequenas. Para o parâmetro de precisão das observações Gaussianas o modelo é pouco sensível a troca de priori. Para o efeito dos locais de coleta é possível verificar que as médias a posteriori não são sensíveis a mudança da priori. A posteriori do intercepto apresentou uma cauda levemente mais pesada com a troca de priori, refletindo a priori mais *flat* que está sendo usada.

Considerando que tem-se dois efeitos aleatórios é interessante verificar como a troca da priori afeta estes efeitos. A Figura 4.5 apresenta a sobreposição do efeito temporal estimado conforme o modelo 3 com as duas configurações de prioris. A mudança da priori não teve um grande impacto nos resultados do efeito temporal, com apenas um alargamento das bandas de credibilidade como era esperado devido a priori Gama(0.01,0.01) ser mais vaga que a Gama(1,0.01). Resultado semelhante é mostrado na Figura 4.6 que

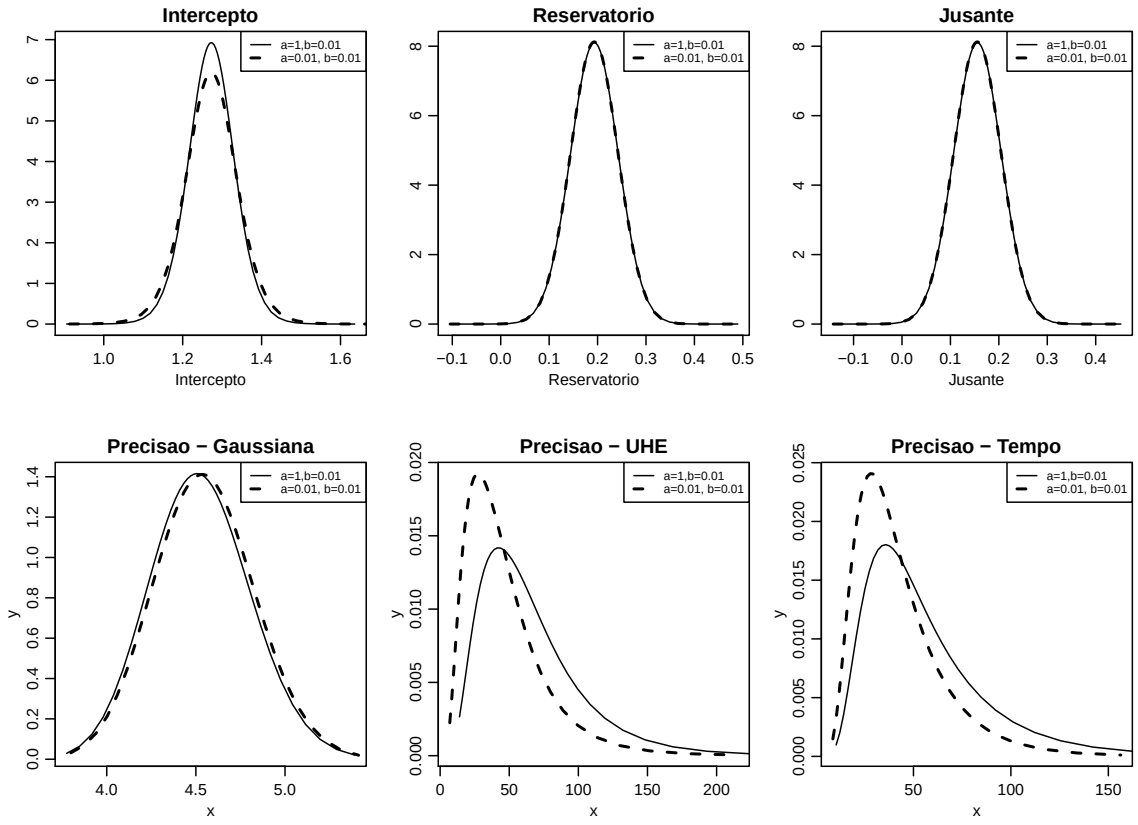

Figura 4.4: Posterioris para os parâmetros de interesse do modelo 3.

apresenta a sobreposição do efeito das UHEs usando as diferentes priori's.

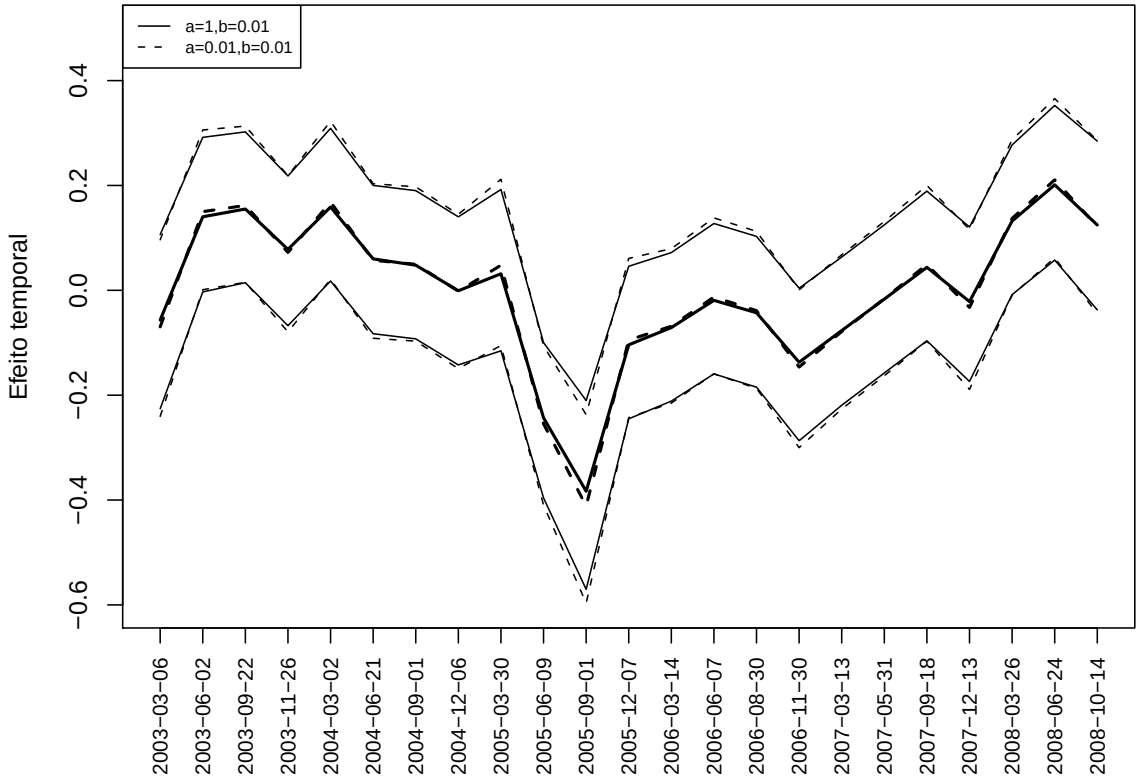

Figura 4.5: Sobreposição do efeito temporal usando diferentes prioris.

Com estes resultados pode-se concluir que os modelos são consistentes e pouco sensíveis à escolha de diferentes prioris. A diferença entre os locais de coleta é consistente em todos os modelos ajustados, com a estação dentro do reservatório apresentando um índice de qualidade da água melhor do que o índice da estação de montante. Resultado semelhante é encontrado quando se compara a montante com a jusante, apenas com diferente grau de intensidade.

Para finalizar a análise pode-se comparar as duas abordagens no sentido de verificar qual produz predições que acompanham melhor o conjunto de dados. Para isto, foram retiradas as observações de todos as usinas e locais de coleta para o último tempo. Os modelos foram reajustados e a predição para as observações retiradas foi realizada. Após feita a predição foram calculadas algumas medidas de concordância. As medidas foram: erro quadrático médio, erro absoluto médio, correlação entre os observados e os preditos e a proporção de pontos observados que estão cobertos pelos intervalos obtidos pelas duas abordagens (cobertura). Os resultados desta análise são apresentados na Tabela 4.3.

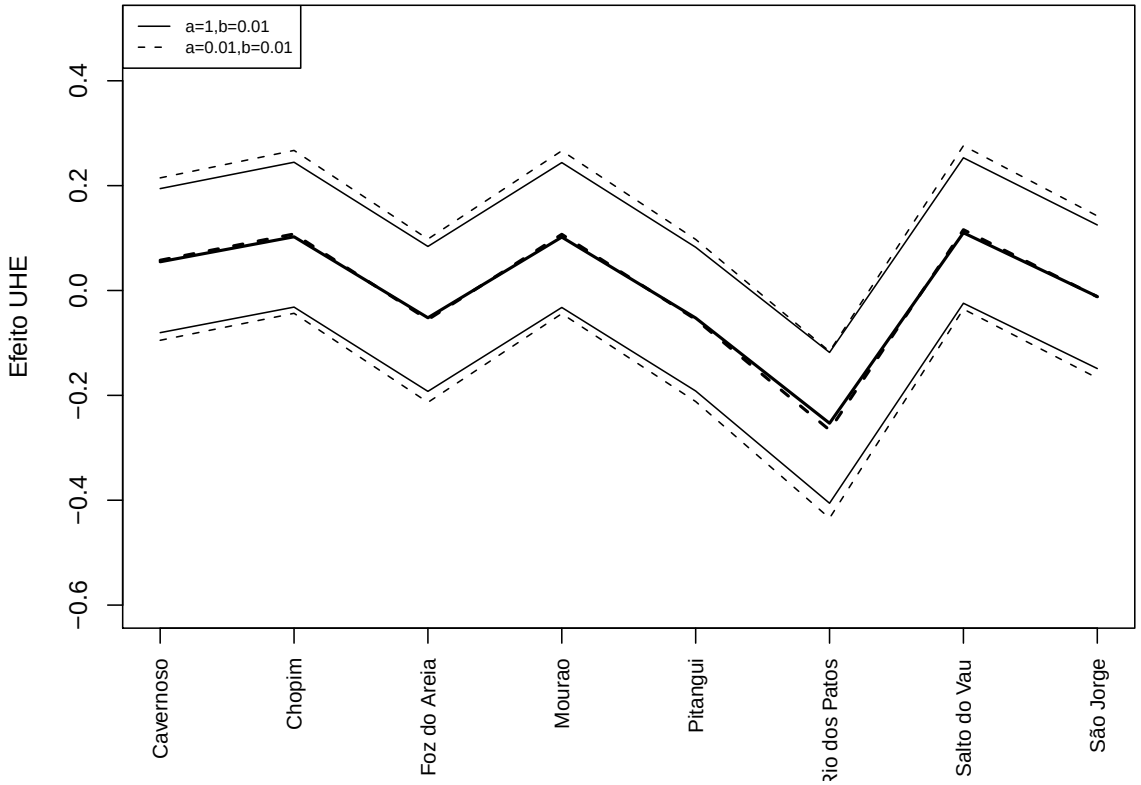

Figura 4.6: Sobreposição do efeito das usinas (UHEs) usando diferentes prioris.

Tabela 4.3: Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados.

Abordagem	Erro quadrático	Erro absoluto	Correlação	Cobertura
INLA	0,1921	0,3390	0,5663	0,6700
GAM	0,2679	0,3918	0,5628	0,5416

Os resultados das medidas de qualidade de predição utilizadas são melhores para a abordagem INLA. A cobertura dos intervalos de credibilidade e confiança ficou abaixo do nível nominal de 95% porém, neste exemplo, esta medida deve ser encarada com cuidado porque a predição foi baseada em apenas 24 observações.

4.2 Investigando fatores associados a ocorrência de ovos de *Aedes aegypti* coletados em ovitrampas em Recife/PE

O objetivo da análise deste conjunto de dados é investigar fatores de risco e proteção associados à ocorrência de ovos do mosquito *Aedes aegypti*, com base em dados do experimento conduzido pelo projeto SAUDAVEL na cidade de Recife/PE, ver Capítulo 2. Entende-se aqui, como fatores de risco/proteção tanto covariáveis associadas à armadilha, como presença de recipientes grandes ou pequenos que possam conter água em suas proximidades, como também aspectos abióticos (climáticos) como temperatura, precipitação e umidade. Possíveis relações espaciais entre as armadilhas e também a possibilidade de uma relação temporal entre as coletas são investigadas. Com o objetivo de identificar possíveis padrões de interesse, tais como, períodos e regiões de maior incidência.

Neste experimento as coletas são feitas em armadilhas, a fim de modelar o relacionamento espacial destas na área em estudo pela metodologia descrita no Capítulo 3, ou seja, usando campos Markovianos é necessário que seja montada uma matriz de precisão. Basicamente, para a construção desta matriz é necessário informar ao modelo qual é a estrutura de vizinhança dos dados. Nesta análise foi feita uma tecelagem de Voronoi, para obter pequenas áreas com base na malha de armadilhas. Após este processo foi construída a matriz pelo critério de adjacência. A Figura 4.7 ilustra este procedimento.

Dado a estrutura espaço-temporal do experimento, o primeiro passo da modelagem deste conjunto de dados, foi a escolha de modelos que levassem em consideração esta estrutura de diferentes formas. Sendo assim, foram ajustados uma série de modelos com grau crescente de complexidade, partindo de um modelo simples apenas com o intercepto até chegar em um modelo que contempla uma interação completa entre espaço e tempo. A série de modelos ajustados é apresentada na Tabela 4.4. A variável resposta é a contagem de ovos e seguindo Bonat et al. (2009) a distribuição Binomial Negativa com função de ligação logaritmo foi assumida para a variável resposta.

Na notação usada para apresentar os modelos, Y sempre representará a variável

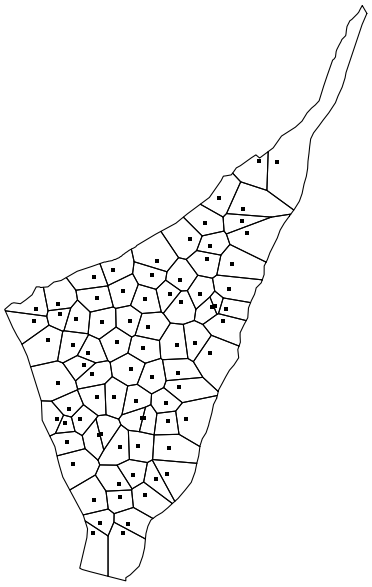

Figura 4.7: Tecelagem de Voronoi com base na malha de armadilhas para a construção da matriz de vizinhança.

resposta, efeitos espaciais e temporais seguem a notação introduzida no Capítulo 3 e efeitos fixos serão representados pela letra β, seguida de um índice. Em todos os modelos o intercepto está presente. Para possibilitar uma comparação e consequente escolha do modelo que melhor descreve a resposta, foram calculados o Critério de Informação da *Deviance* (DIC), o número estimado de parâmetros e a verossimilhança marginal.

Para ajustar os modelos apresentados na tabela 4.4 foi usada a aproximação Gaussiana ao invés da aproximação de Laplace para o Campo Aleatório Markoviano Gaussiano, por não estar claro como usar a aproximação de Laplace na situação de modelos com interação espaço-temporal. No decorrer das análises foi usada sempre usada a aproximação de Laplace.

Tabela 4.4: Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal.

Modelos	Preditor Linear	DIC	NP	MV
1	$Y \sim 1$	$35691,77$	$1,973$	$-17853,93$
2	$Y \sim \gamma_t + \phi_i$	$35004,85$	$172,34$	$-17624,54$
3	$Y \sim \rho_t + \varphi_i$	$34969,34$	$123,04$	$-17635,08$
4	$Y \sim \rho_t + \varphi_i + \gamma_t + \phi_i$	$34968,48$	$126,34$	$-17633,14$
5	$Y \sim$ Tipo I	$34972,28$	$150,79$	$-17637,76$
5	$Y \sim$ Tipo II	$34979,30$	$156,41$	$-17990,40$
5	$Y \sim$ Tipo III	$35121,00$	$264,38$	$-23964,46$
5	$Y \sim$ Tipo IV	$35105,00$	$245,19$	$-24176,06$

O modelo 1 representa a variabilidade total dos dados e foi ajustado apenas para servir de base para comparações. O modelo 2 leva em consideração efeitos espaciais e temporais, porém assume a priori que tais efeitos não tem nenhuma estrutura. O modelo 3 assume a priori que os efeitos espaciais e temporais são estruturados, conforme descrito no Capítulo 3. O modelo 4 condensa os modelos 2 e 3 dizendo que os efeitos espaciais e temporais podem ser divididos em uma parte estruturada e outra não estruturada. Os modelos de interação espaço-tempo seguem a nomenclatura dada na seção 3.2.2.

De acordo com os resultados apresentados na Tabela 4.4, o modelo que apresenta o menor DIC é o 4, com efeitos espaciais e temporais, estruturados e não estruturados. Porém, olhando o ajuste dos efeitos não estruturados, verifica-se que estes não ajudam a explicar a variável resposta, ou seja, não apresentam significância. Além disso, a diferença em DIC para o modelo com apenas efeitos estruturados é de apenas $0,86$ unidades, pode-se verificar também que as verossimilhanças marginais dos modelos são muito próximas. Calculando o fator de Bayes tem-se um valor de $0,999$, mostrando que não se tem nenhuma perda significativa ao trocar o modelo 4 pelo modelo 3. A última consideração a ser feita é com relação ao número estimado de parâmetros, sendo, o modelo 3 mais parsimonioso que o 4. Sendo assim, para descrever as condições espaço-temporais em que o experimento foi realizado, o modelo 3 é escolhido.

Dado que a estrutura espaço-temporal do experimento já está representada no modelo, pode-se a partir de agora investigar o efeito de cada uma das covariáveis ou possíveis fatores de risco/proteção sobre a variável resposta. As condições ligadas às armadilhas (locais) são de papel fundamental, pois elas podem orientar as políticas de prevenção da propagação do vetor através de campanhas educacionais promovidas a fim de evitar criadouros do mosquito. Assim, a identificação das características associadas é importante para orientar as ações de tais campanhas.

O conjunto de gráficos da Figura 4.8 faz uma comparação das contagens de ovos (em escala logarítmica) entre as categorias de cada uma das doze covariáveis locais. A análise destes gráficos permite identificar de forma exploratória e inicial, os fatores que mais afetam as contagens de ovos, orientando a seleção e a escolha de modelos, antecipando e explicando possíveis resultados da modelagem.

Pelo conjunto de gráficos da Figura 4.8 é possível observar que as diferenças são pequenas e que existem apenas em algumas das covariáveis. Pode-se observar que o tipo de imóvel residencial apresenta contagens médias maiores que os imóveis não residenciais. A presença de fatores de risco também apresenta um leve aumento na contagem média de

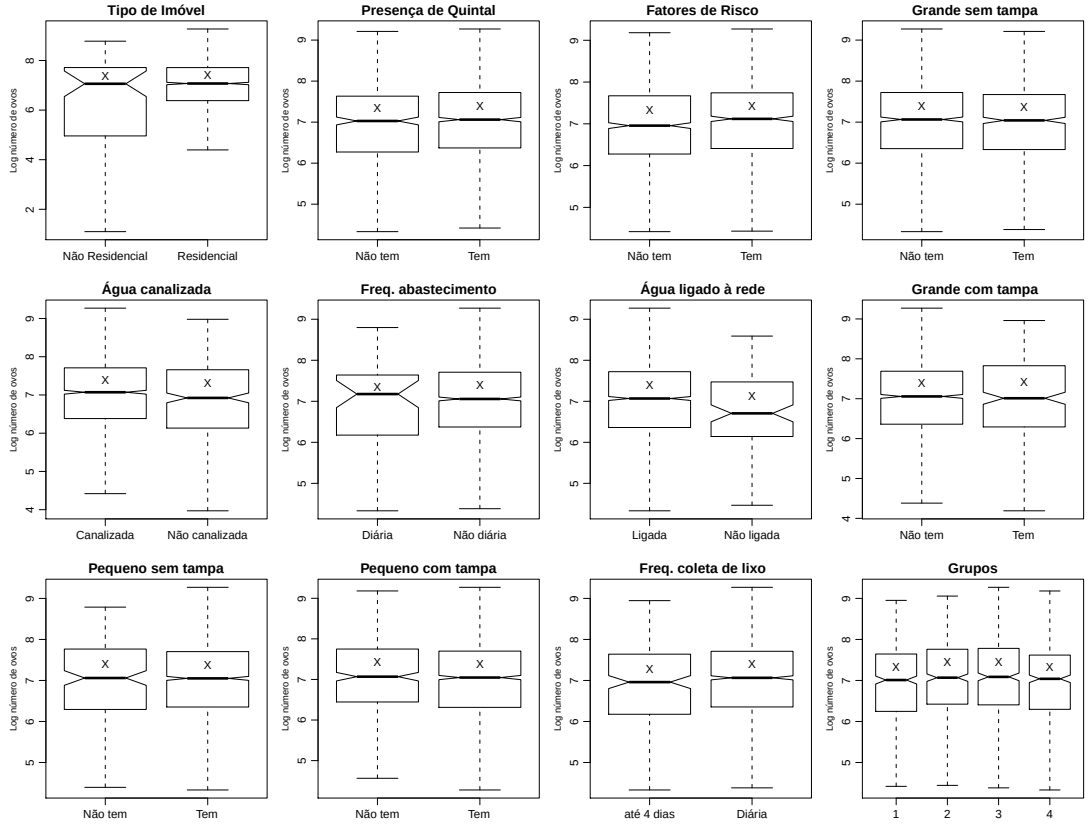

Figura 4.8: Boxplots do log das contagens de ovos por categorias para cada covariável (X representa a média amostral), bairro Brasilia Teimosa Recife/PE.

ovos. As condições ligadas a água, água canalizada, frequência de abastecimento e água ligada a rede geral, apresentam uma pequena diferença entre as categorias, sendo sempre maior nas categorias que indicam presença de água abundante. Quanto à frequência de coleta de lixo, os dados mostram uma maior contagem média de ovos nas localizações que apresentam coleta de lixo diária. Para as outras covariáveis não é possível identificar nenhuma relação aparente.

Para melhor explorar o efeito das covariáveis locais foi ajustado um modelo para cada uma dessas covariáveis controlando os efeitos espaciais e temporais conforme o modelo 3 escolhido anteriormente. Além disso, foram consideradas as abordagens INLA e GAM para comparações dos resultados. Na abordagem GAM os efeitos espaciais e temporais foram controlados por funções suaves das coordenadas geográficas e datas de coletas. O resumo desta análise é apresentada na Tabela 4.5 através de médias das posterioris e intervalos de credibilidade baseados em quantis para a abordagem INLA e estimativas pontuais com intervalos de confiança assintóticos para a abordagem GAM. O nível nominal de 95% foi fixado para ambos os casos.

Tabela 4.5: Ajustes dos modelos para cada covariável na presença dos efeitos espaciais e temporais, abordagens INLA e GAM.

Parâmetros	Média Post.	Int. Cred.	Estimativa	Int. Conf.
Tipo de imóvel	$-0,040$	$(-0,231;0,148)$	$-0,273$	$(-0,628;0,081)$
Quintal	$0,074$	$(-0,133;0,282)$	$0,065$	$(-0,053;0,184)$
Água ligada a rede geral	$-0,066$	$(-0,407;0,274)$	$-0,199$	$(-0,405;0.007)$
Abastecimento de água	$0,080$	$(-0,319;0,479)$	$0,092$	$(-0,238;0,423)$
Água canal. no cômodo	$-0,116$	$(-0,344;0,109)$	$-0,242$	$(-0,376;-0,113)$
Fatores de risco	$0,143$	$(-0,020;0,307)$	$0,083$	$(0,003;0,192)$
Rec. grandes sem tampa	$-0,054$	$(-0,232;0,123)$	$-0,125$	$(-0,227;-0,024)$
Rec. grandes com tampa	$-0,026$	$(-0,276;0,221)$	$-0,015$	$(-0,155;0,124)$
Rec. pequenos sem tampa	$-0,040$	$(-0,371;0,289)$	$0,029$	$(-0,152;0,212)$
Rec. pequenos com tampa	$-0,086$	$(-0,298;0,125)$	$-0,136$	$(-0,252;-0,019)$
Fre. coleta de lixo	$-0,024$	$(-0,226;0,176)$	$0,123$	$(-0,039;0,286)$
Grupos 1 - 2	$0,150$	$(-0,081;0,382)$	$0,079$	$(-0,043;0,202)$
Grupos 1 - 3	$0,155$	$(-0,084;0,394)$	$0,120$	$(-0,004;0,244)$
Grupos 1 - 4	$-0,001$	$(-0.232;0,229)$	$0,004$	$(-0,121;0,129)$

Como pode ser visto na Tabela 4.5 os resultados diferem entre os modelos em termos de média posteriori e estimativativas pontuais. De forma geral, a abordagem INLA parece tender a ser mais conservadora, não indicando nenhuma covariável como significativa na presença do efeito espaço-temporal.

Pela abordagem GAM a interpretação dos intervalos de confiança indica que as

covariáveis Fatores de Risco, Recipientes grandes sem tampa, água canalizada e recipientes pequenos com tampa são indicadas como significativas. Nota-se que destas quatro covariáveis, três delas apresentam uma das bordas do intervalo estimado muito próximo do zero, o que indicaria não significância do efeito para intervalos apenas ligeiramente mais conservadores. Com excessão da covariável água canalizada, o que de certa forma reforça os resultados obtidos pela abordagem INLA.

Pensando que pela abordagem INLA tem-se uma distribuição a priori Gaussiana para os efeitos fixos de média zero e variância infinita, a covariável tem que trazer muito mais informação para ser considerada significativa do que pela abordagem GAM, onde não se tem priori, e isto pode ser uma explicação para as diferenças entre os resultados. Outra diferença entre os modelos é a forma em que o efeito espaço-temporal é considerado no modelo. O INLA trabalha com propriedades de independência condicional, decodificada de forma conveniente através de uma matriz de precisão na distribuição a priori para os efeitos estruturados o que torna este um efeito aleatório. Já no GAM estes efeitos são suavizados por uma *spline* (efeito fixo), o que impede que mudanças bruscas ocorram na função estimada dos efeitos espaciais e temporais. Isto faz com que uma porção menor da variabilidade da resposta possa ser atribuída a função suave do que a um processo estocásticos subjacente, e consequentemente ao menos parcialmente captada pelo intercepto e possívelmente pelas covariáveis, o que explica as estimativas pontuais do GAM serem em geral mais distantes do zero do que as médias a posterioris obtidas com o INLA.

Considerando que o GAM faz uso da teoria assintótica atribuída aos estimadores de máxima verossimilhança, por mais que na prática um algoritmo de mínimos quadrados modificado seja usado para obter as estimativas, pode-se desenhar a distribuição assintótica das estimativas obtidas pelo GAM e compará-la com as distribuições a posteriori obtidas com o INLA. Lembrando que existe uma constante de proporcionalidade desconhecida nas posterioris, para as duas distribuições poderem ser colocadas no mesmo gráfico é preciso que sejam padronizadas. Sendo assim, basta fazer com que a moda da posteriori e a moda da distribuição assintótica da estimativa GAM corresponda ao valor 1. Desta maneira, as distribuições podem ser colocadas no mesmo gráfico e comparadas visualmente com relação a locação e variabilidade. Essa visualização é apresentada na Figura 4.9 para as doze covariáveis locais em análise.

Analisando as distribuições apresentadas na Figura 4.9 é possível ver que as posterioris apresentam maior variabilidade. Sendo que, em geral, a posteriori cobre a distribuição assintótica das estimativas obtidas pelo GAM. A forma das posterioris são bem

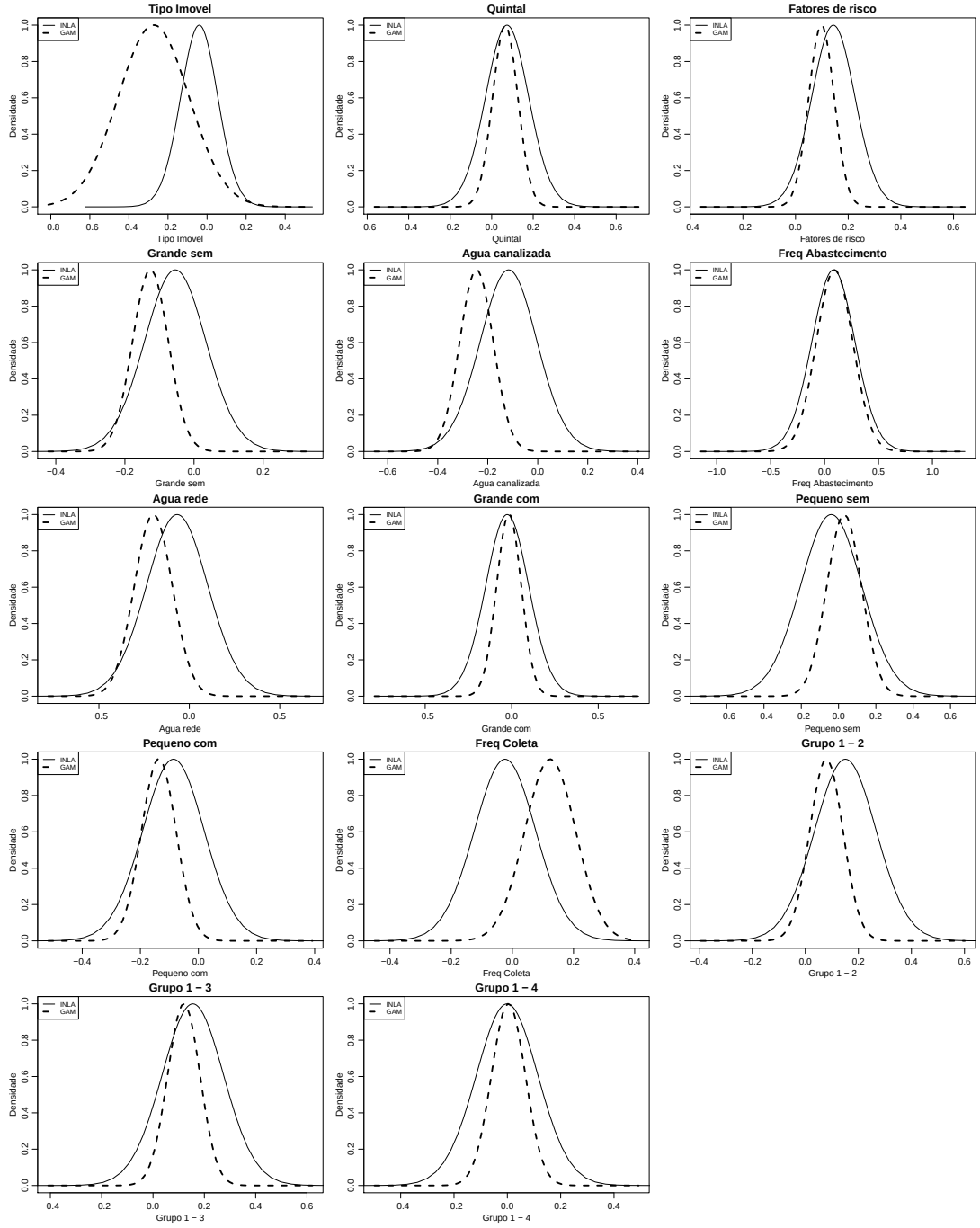

Figura 4.9: Comparação entre as distribuições a posteriori e assintótica das estimativas dos efeitos das covariáveis locais.

definidas em formato aparentemente próximo o de uma distribuição normal.

Seguindo a mesma idéia básica da análise anterior, para investigar o efeito das condições climáticas sobre o número de ovos. Foi ajustado um modelo para cada covariável separadamente, considerando que não se sabe a priori a forma do relacionamento destas covariáveis com a resposta, atribuiu-se à priori para este efeito um modelo *random walk* de primeira ordem, a fim de permitir inferir também sobre a forma do relacionamento. Novamente para o ajuste do modelo a variável resposta foi assumida Binomial Negativa com função de ligação logarítmica. Também para comparações dos resultados foram ajustados GAM's onde os efeitos espaciais e temporais foram controlados por funções suaves das coordenadas geográficas das armadilhas e das datas de coleta. Para o efeito das covariáveis climáticas uma função suave (*spline*) foi ajustada.

A Figura 4.10 apresenta o efeito de cada uma das doze covariáveis climáticas que estão sob análise. Pela Figura é possível ver que as duas abordagens trazem resultados diferentes em praticamente todas as covariáveis ambientais analisadas. Novamente a abordagem INLA tende a ser mais conservadora, além disso, as funções suaves estimadas pelo GAM tendem a ter um comportamento oscilante, dando uma idéia de que o relacionamento da covariável com a resposta é não-linear na escala da função de ligação. É conhecido (VENABLES; DICHMONT, 2004) que o GAM tende a seguir mais o comportamento dos dados, sendo influenciado por pequenas variações na covariável tornando o modelo exploratório e particular para aquele conjunto de dados. Enquanto que modelando o efeito da covariável como aleatório esse comportamento tende a ser melhor captado como um processo subjacente tornando o modelo mais geral. Este resultado confirma um dos achados de (BONAT et al., 2009) onde os autores concluem que mesmo com estes comportamentos oscilantes dos *splines* o modelo final considerava os efeitos das covariáveis climáticas como lineares na escala da função de ligação.

Pelos resultados dos modelos via INLA nenhuma das covariáveis ambientais apresenta relação com as contagens de ovos. Pela abordagem GAM, aparentemente mais flexível, aparecem diversas formas de relacionamento, porém nada de forma clara. Uma análise detalhada com base na abordagem GAM é encontrada em (BONAT et al., 2009) e não será repetida aqui.

Seguindo o artigo de Bonat et al. (2009) os autores através de um esquema de seleção de covariáveis ambientais e locais chegam a um modelo, que é composto pelas covariáveis Precipitação no mês da observação (PREC.MES1) e no mês anterior (PREC.MES2). A covariável Umidade de dois meses antes da observação (UMID.MES3),

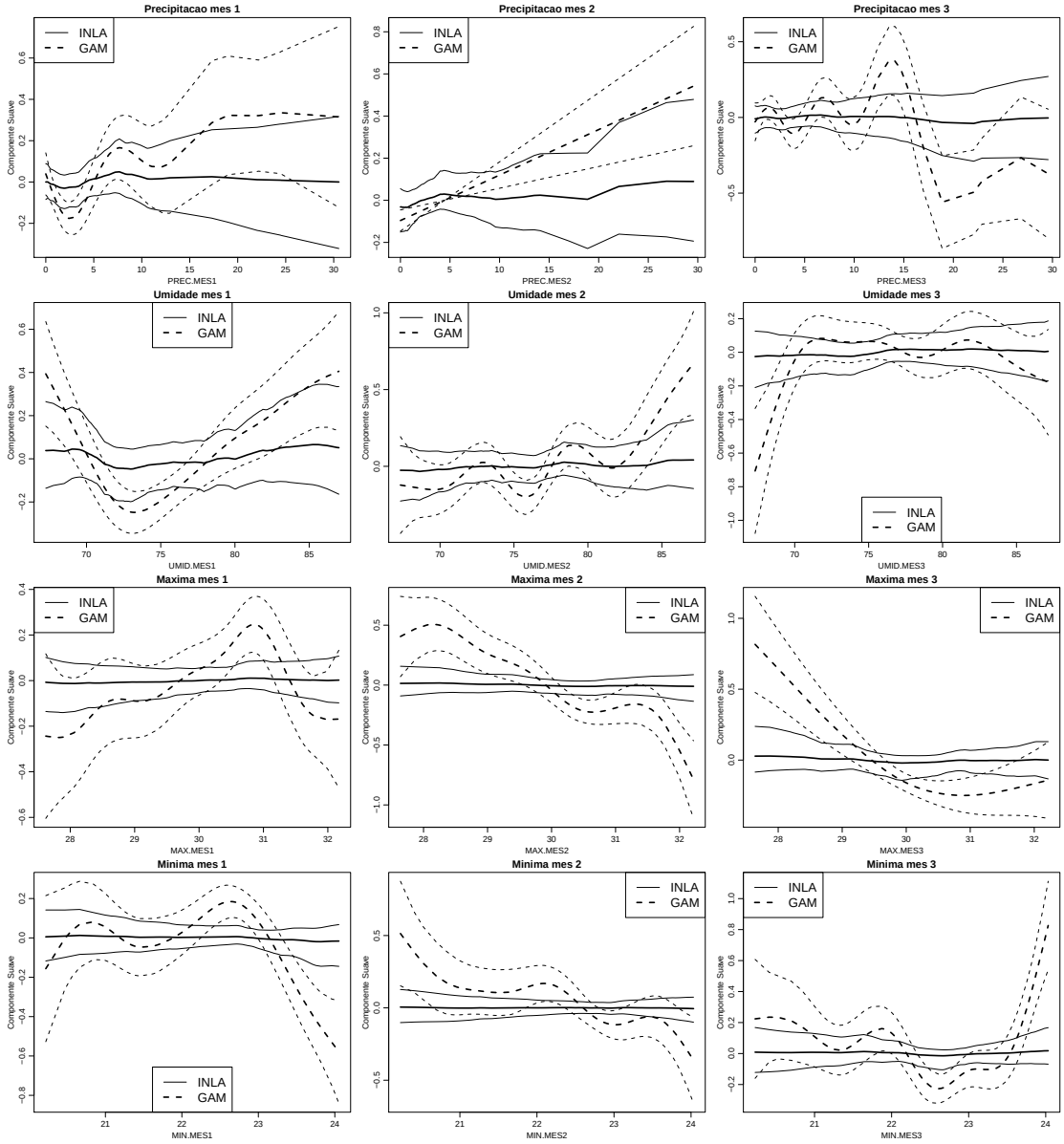

Figura 4.10: Sobreposição do efeito estimado e respectivas faixas de confiança de cada covariável ambiental pelas abordagens INLA e GAM.

as covariáveis locais foram a presença de Água Canalizada e recipientes grandes sem tampa, além dos efeitos espaciais e temporais considerados como funções suaves das coordenadas geográficas e das datas de coletas. Sendo assim, foi ajustado o mesmo modelo pela abordagem INLA para permitir comparações dos resultados. O resumo desta análise é apresentada na Tabela 4.6 através de médias das posterioris e intervalos de credibilidade baseados em quantis para a abordagem INLA e estimativas pontuais com intervalos de confiança assintóticos para a abordagem GAM, ambos com nível nominal de 95%.

Tabela 4.6: Ajuste do modelo proposto em Bonat et al. (2009) pelas abordagens INLA e GAM.

Parâmetros	Média Posteriori	Int. Cred.	Estimativa	Int. Conf.
Intercepto	6,487	(3,878;9,150)	5,125	(3,344;6,905)
PREC.MES1	0,006	(−0,015;0,027)	0,027	(0,016;0,039)
PREC.MES2	0,009	(−0,013;0,031)	0,027	(0,016;0,039)
UMID.MES3	0,009	(−0,025;0,043)	0,025	(0,002;0,0484)
Canalizada	−0,110	(−0,343;0,120)	−0,241	(−0,375;−0,107)
Grande sem tampa	−0,045	(−0,225;0,134)	−0,111	(−0,216;−0,007)

A Tabela 4.6 mostra que pela abordagem INLA nenhuma das covariáveis consideradas no modelo apresentam significância, já que, os intervalos de credibilidade baseados no quantil de 95% cobrem o zero, contradizendo os resultados do modelo GAM. Pode ser observado que as médias a posteriori tendem a ser mais próximas do zero, que as estimativas pontuais obtidas pelo GAM. Conforme já identificado na análise anterior. As mesmas possíveis explicações cabem novamente para este resultado. É interessante sobrepor a distribuição assintótica do estimador obtido pelo GAM com as posterioris obtidas pelo INLA para verificar a diferença entre as estimativas dos efeitos destas covariáveis. A Figura 4.11 apresenta esta comparação.

Pelas distribuições apresentadas na Figura 4.11 fica claro que as médias a posteriori (INLA) estão sempre mais próximas do zero, que as estimativas pontuais (GAM), além disso, a incerteza estimada pelo INLA é maior em todas as covariáveis consideradas, mostrando que a abordagem INLA tende a ser mais conservadora.

As duas abordagens estão considerando os efeitos espaciais e temporais. Sendo assim, é interessante ver como cada abordagem capta tais efeitos. O efeito temporal no INLA é modelado segundo um modelo *random walk* de primeira ordem. Enquanto que no GAM este efeito é um *spline* das datas de coletas. A Figura 4.12 sobrepõe o efeito temporal estimado pelas duas abordagens.

Pela Figura 4.12 é clara a alta suavização do efeito temporal pela abordagem

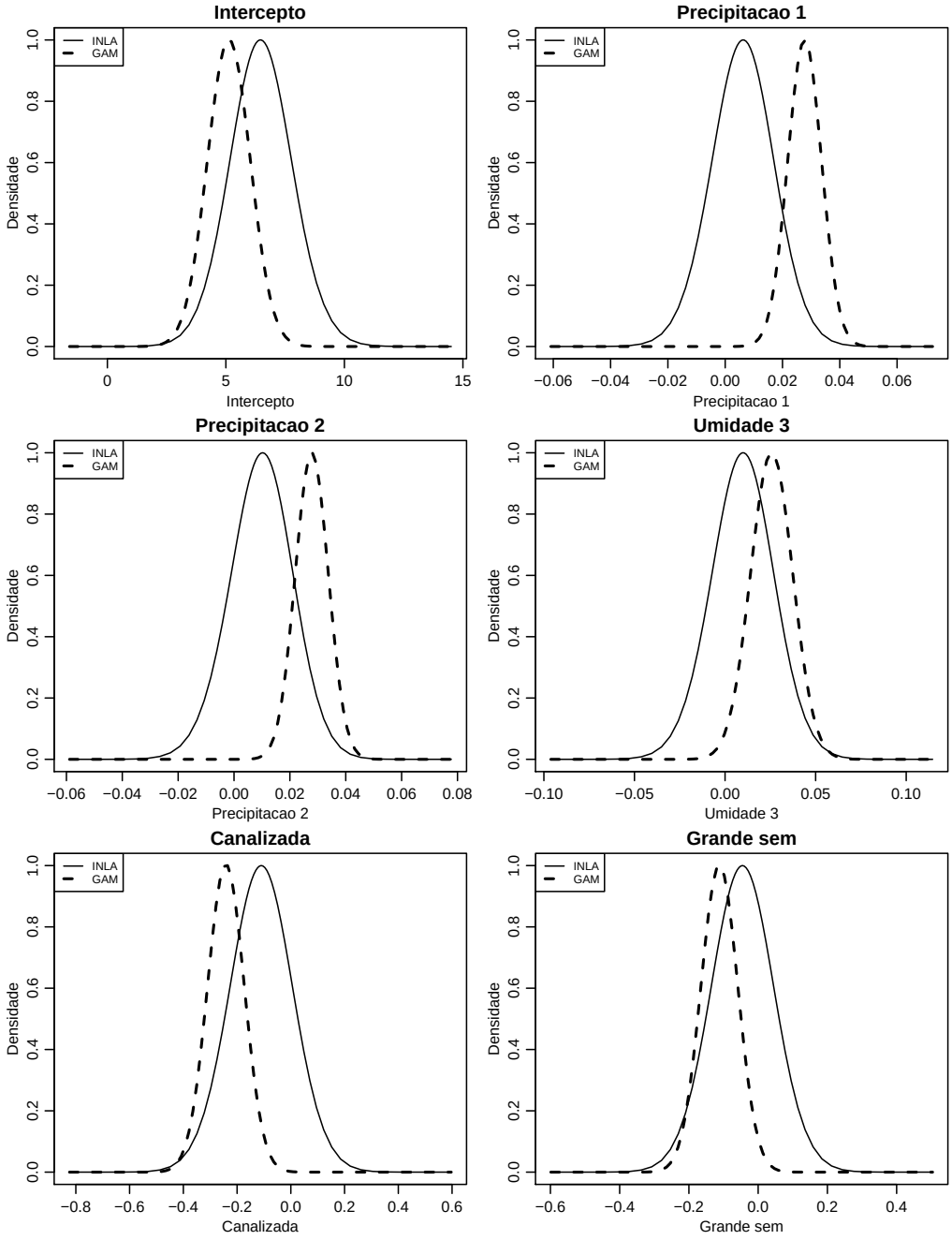

Figura 4.11: Sobreposição do efeito estimado de cada covariável conforme o modelo proposto em (BONAT et al., 2009).

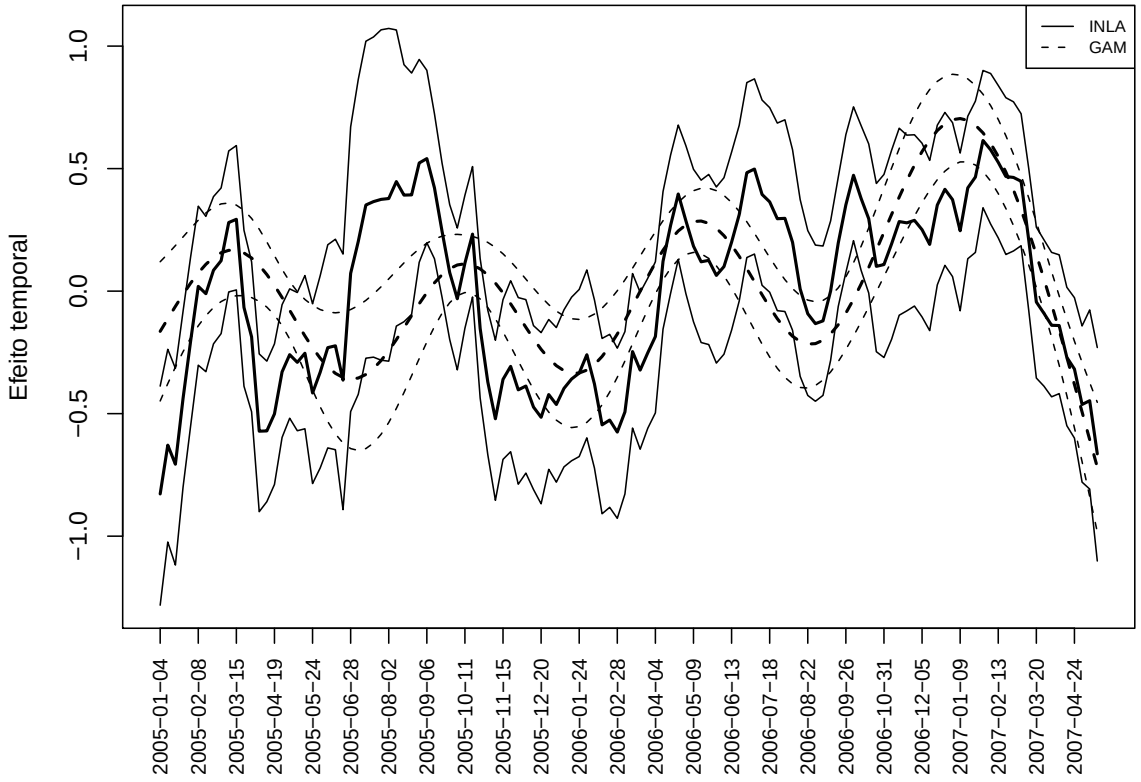

Figura 4.12: Sobreposição do efeito temporal estimado conforme o modelo proposto em (BONAT et al., 2009) pelas abordagens INLA e GAM.

GAM. Como era esperado o modelo *random walk* permite que mudanças bruscas ocorram na série temporal, o que não é possível quando ajusta-se um *spline*. Apesar da grande variabilidade do *random walk* de forma geral ele acompanha o comportamento do *spline*, aparentando ser uma versão mais ruidosa deste.

O último efeito que compõe o modelo é o efeito espacial. Pela abordagem GAM este efeito é modelado por um *spline* bidimensional das coordenadas das armadilhas. Na abordagem INLA este efeito é modelado através de uma autoregressão intrínseca, conforme explicado no Capítulo 3. Para comparar como cada abordagem capta tal efeito, cada par de coordenadas é associado a uma área (ver Figura 4.7), assim pode-se isolar o efeito espacial pelas duas abordagens para cada área.

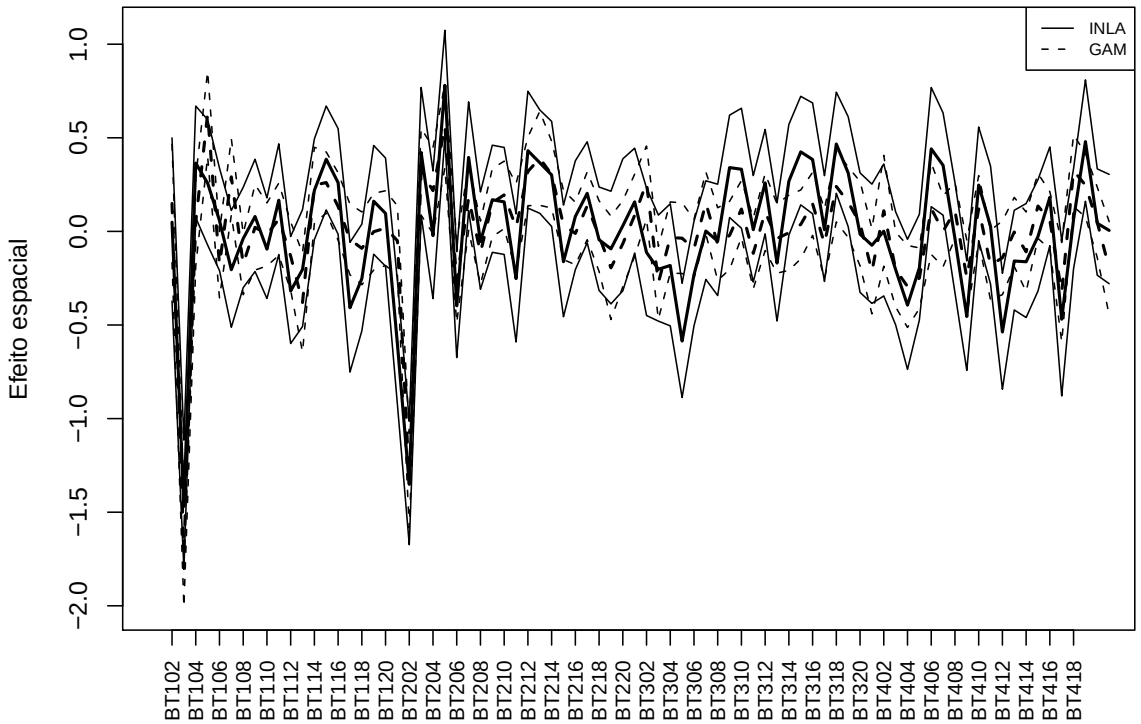

Figura 4.13: Sobreposição do efeito espacial estimado conforme o modelo proposto em (BONAT et al., 2009) pelas abordagens INLA e GAM.

Pela Figura 4.13 é clara a semelhança entre as duas abordagens. Para melhor explorar este efeito é interessante ainda que seja construído um mapa, espacializando os resultados. Como é apresentado na Figura 4.14, usando o mapa construído via tecelagem de voronoi.

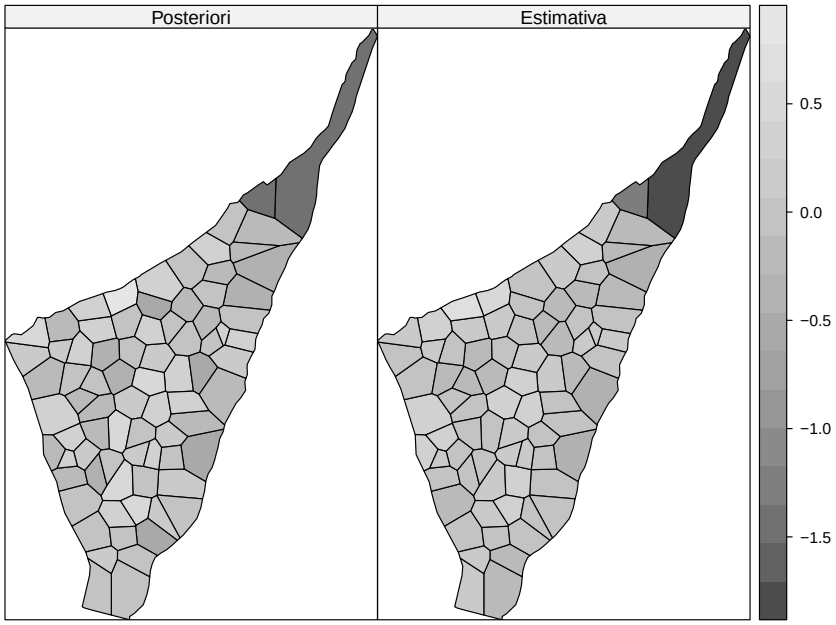

Figura 4.14: Mapas das estimativas do efeito espacial pelas abordagens INLA e GAM

Para facilitar a visualização do mapa da figura 4.14 a figura 4.15 apresenta um digrama de dispersão entre o efeito espacial calculado pelas abordagens INLA e GAM.

Apresenta-se na Figura 4.16 um mapa do desvio padrão da estimativa do efeito espacial para cada localização obtidos pelas duas abordagens.

Como já foi visto em praticamente todos os resultados até aqui, a abordagem GAM associa menores intervalos aos efeitos que a abordagem INLA. E isso se repetiu nas estimativas para o efeito espacial. A média do desvio padrão pelo INLA é de 0.1535 enquanto que pelo GAM é de 0.1028, uma diferença considerável. As duas abordagens identificam as áreas de borda principalmente a parte superior direita dos mapas como sendo a de maior incerteza.

Para verificar a robustez dos resultados obtidos pela abordagem INLA, os modelos foram reajustados mudando a priori dos parâmetros de precisão, conforme explicado na seção 3.2.3 sendo os resultados obtidos praticamente iguais. Uma diferença pequena é verificada apenas nas estimativas de desvio padrão que tem um aumento como era esperado. Desta forma, optou-se por não repetir a apresentação de toda a análise, pois seria muito cansativa e os resultados são os mesmos.

Para finalizar a análise pode-se comparar a capacidade preditiva das duas abordagens de modelagem no sentido de verificar qual acompanha melhor o conjunto de dados.

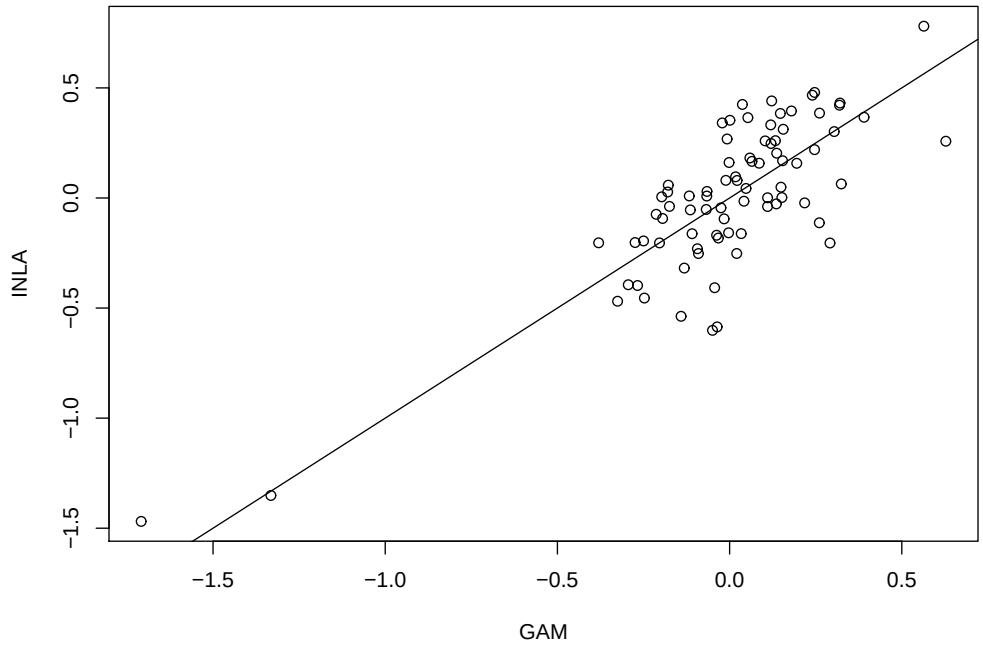

Figura 4.15: Diagrama de dispersão entre o efeito espacial calculado pelas abordagens INLA e GAM.

Para isto, foram retiradas as 80 últimas observações que correspondem aos quatro últimos tempos. Os modelos foram reajustados e as predições obtidas para cada modelo. Pela abordagem INLA foi usado o modelo 3 que não leva em consideração nenhuma covariável, já que, por esta abordagem nenhuma mostrou-se significativa. Para abordagem GAM foi utilizado o modelo proposto em Bonat et al. (2009). Após isto, algumas medidas de similaridade foram calculadas: erro quadrático médio, erro absoluto médio, correlação entre observado e preditos e nível de cobertura. Os resultados são apresentados na Tabela 4.7.

Tabela 4.7: Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados.

Abordagem	Erro quadrático	Erro absoluto	Correlação	Cobertura
INLA	958264	685,23	0,3440	0,6029
GAM	1009266	681,21	0,1962	0,3432

Os resultados da Tabela 4.7 mostra que a abordagem INLA apresenta melhores resultados em todas as medidas com excessão ao Erro absoluto médio, embora neste caso a diferença seja pequena. O nível de cobertura pelas duas abordagens ficou abaixo do nível nominal de 95%, porém a abordagem INLA se aproximou mais deste, do que, a abordagem GAM.

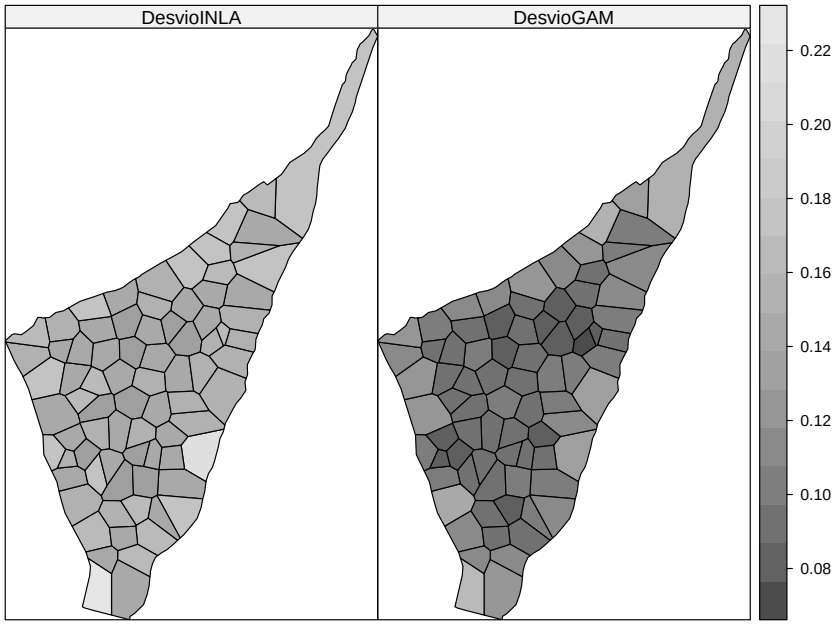

Figura 4.16: Mapas dos desvios padrão para cada localização espacial pelas abordagens INLA e GAM.

4.3 Análise do padrão espaço-temporal da leprose-dos-citrus

O objetivo da análise deste conjunto de dados é verificar a existência de padrão e compreender o caminho espaço-temporal da doença leprose-dos-citrus em um talhão com 1160 plantas, que foram avaliadas em 45 tempos aproximadamente igualmente espaçados. Os primeiros 15 tempos foram excluídos da análise por não apresentarem nenhum caso da doença. Sendo assim, a base de dados conta com 30 tempos com um total de 34800 observações. Cada planta em cada tempo é classificada como não doente(0) e doente (1).

Para todos os modelos ajustados a distribuição de probabilidade atribuída à variável resposta foi a Binomial com função de ligação *logit*. As condições espaço-temporais em que o experimento foi conduzido são levadas em consideração nos modelos de diversas formas, a fim de encontrar aquela que melhor explica a variabilidade da variável resposta. A tabela 4.8 apresenta a série de modelos que foram ajustados.

Para possibilitar uma comparação e consequente escolha do modelo que melhor descreve a resposta, foram calculados o Critério de Informação da *Deviance* (DIC), o número estimado de parâmetros e a verossimilhança marginal. Na notação usada para apresentar os modelos Y sempre representará a variável resposta, efeitos espaciais e tem-

porais seguem a notação introduzida no capítulo 3. Para todos os modelos, o intercepto está presente.

Tabela 4.8: Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados, verossimilhança marginal e critério de informação de *Akaike*.

Modelos	Preditor Linear	DIC	NP	MV
1	$Y \sim 1$	$24592,37$	$1,022$	0
2	$Y \sim \gamma_t$	$20043,72$	$29,87$	$-10090,03$
3	$Y \sim \rho_t$	$20028,79$	$20,20$	$-10038,06$
4	$Y \sim \phi_i$	$16495,41$	$1601,32$	$-8084,46$
5	$Y \sim \varphi_i$	$15164,56$	$918,60$	$-8825,41$
6	$Y \sim \gamma_t + \phi_i$	$9885,40$	$3538,92$	$-3246,39$
7	$Y \sim \rho_t + \varphi_i$	$9716,71$	$3457,86$	$-3843,93$
8	$Y \sim \rho_t + \varphi_i + \gamma_t + \phi_i$	$9601,78$	$3399,04$	$-3844,82$

O modelo 1 apenas com o intercepto foi ajustado apenas para servir como comparação. Os modelos 2 e 3 levam em consideração apenas o efeito temporal, sendo que o 2 assume a priori que o efeito não tem nenhuma estrutura, e o 3 assume uma estrutura do tipo *random walk* de primeira ordem.

Os modelo 4 e 5 levam em consideração apenas o efeito espacial, sendo que o 4 assume a priori que o efeito não tem nenhuma estrutura no espaço, e o 5 assume uma estrutura condicional autoregressiva. O modelo 6 condensa os modelos 2 e 4 e o modelo 7 condensa os modelos 3 e 5. Finalizando o modelo 8 condensa todos os anteriores, dizendo que os efeitos espaciais e temporais podem ser divididos em uma parte com estruturada no espaço ou tempo, e outra parte é não estruturada.

A limitação que surgiu nesta análise é que não foi possível estimar os modelos com interação espaço-tempo apresentados na seção 3.2.2. Em princípio o problema com estes modelos neste conjunto de dados parece ser o número de plantas (áreas) que compõe o talhão (1160). Com esta estrutura a matriz de precisão Q do modelo torna-se intratável computacionalmente, nas rotinas implementadas no pacote INLA (MARTINO; RUE, 2008). Foram feitas várias tentativas inclusive tomando-se apenas partes do conjunto de dados, mas as análises ficaram pouco confiáveis, com resultados não consistentes. Desta forma, optou-se por não apresentá-los nesta dissertação.

Na tabela 4.8 comparando-se os modelos 2 e 3 que tratam o efeito temporal, é possível verificar que o modelo com estrutura temporal apresenta um melhor ajuste, tanto em termos de DIC quanto de número estimado de parâmetros. Porém, em termos de verossimilhança marginal, calculando o fator de Bayes entre os dois modelos tem-se um valor de 0.9948 que não mostra nenhuma melhora expressiva ao trocar o modelo 3

pelo 2.

Com relação ao efeito espacial, comparando os modelos 4 e 5, verifica-se que o modelo com efeito espacial estruturado apresenta um melhor ajuste em termos de DIC e número estimado de parâmetros. Em termos de verossimilhança marginal, calcula-se o fator de Bayes comparando o modelo 5 com o 4 tem-se um valor de 1.0916 que mostra uma pequena evidência de melhora do ajuste com o efeito espacial estruturado.

Em termos gerais o modelo 8 é o que apresenta o menor DIC o que o indica como o melhor entre os modelos ajustados para descrever o comportamento da variável resposta. Em termos de número de parâmetros ele é mais parsimonioso que o modelo 7, e em termos de verossimilhança marginal os modelos 7 e 8 são equivalentes. A figura 4.17 apresenta um gráfico do ajuste de cada bloco de parâmetros do modelo 8.

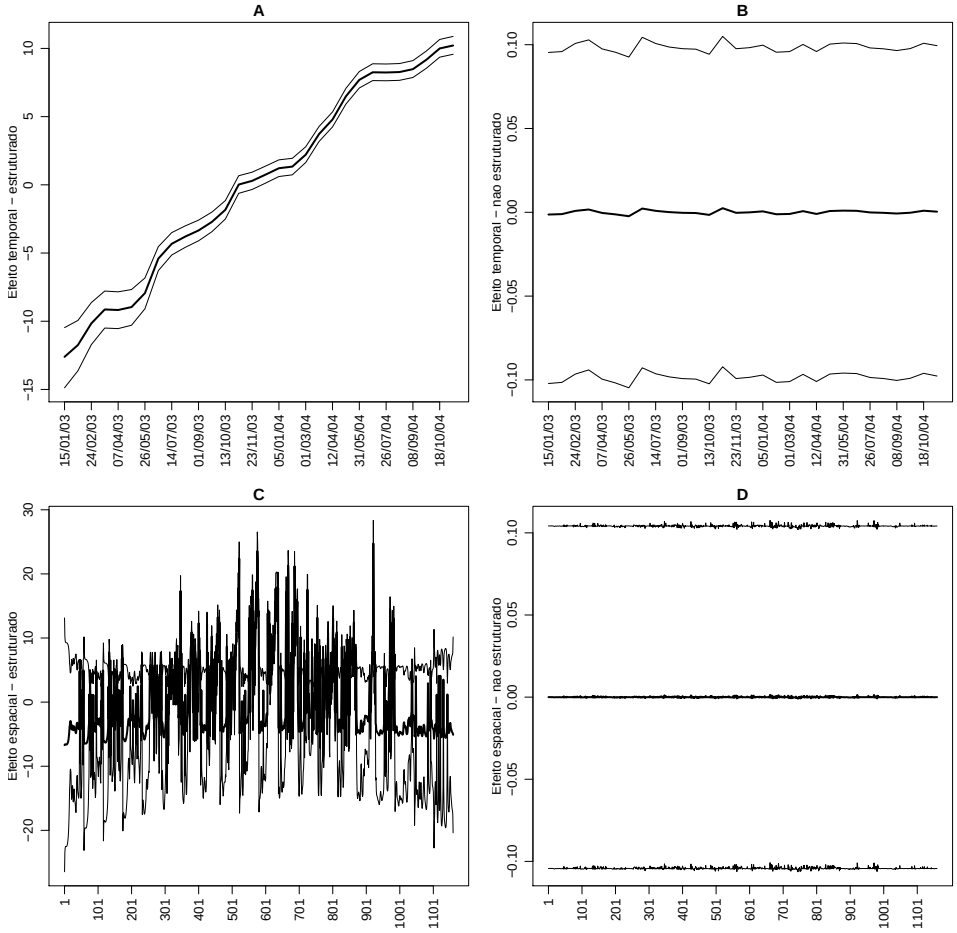

Figura 4.17: Ajuste de cada bloco de parâmetros do modelo 8.

A figura 4.17 apresenta o ajuste de cada um dos blocos de parâmetros do modelo 8. O painel A é o ajuste do efeito temporal estruturado ρ_t, o painel B é o efeito temporal não estruturado γ_t. O painel C e D é o efeito espacial estruturado e não estruturado (φ_i

e ϕ_i) respectivamente.

A figura 4.17 mostra claramente que os efeitos não estruturados não ajudam a descrever o comportamento da variável resposta, já que, não apresentam diferenciais entre os pontos no tempo ou espaço. Claramente os efeitos estruturados captam toda a variabilidade da variável resposta, mostrando que apenas os dois efeitos estruturados são responsáveis pela qualidade do ajuste, como mostrava as verossimilhanças marginais anteriormente. Desta forma, opta-se pela escolha do modelo 7 como o modelo que melhor descreve o padrão espaço-temporal da leprose-dos-citrus.

A fim de melhor explorar os resultados desta análise, ajusta-se um modelo aditivo generalizado para confrontar os resultados, e verificar como cada modelo capta cada efeito. Sendo assim, no GAM os efeitos espaciais e temporais são modelados como uma função *spline* das datas de coleta e das coordenadas x e y das plantas no talhão. A figura 4.18 mostra o resultado desta sobreposição para o efeito temporal.

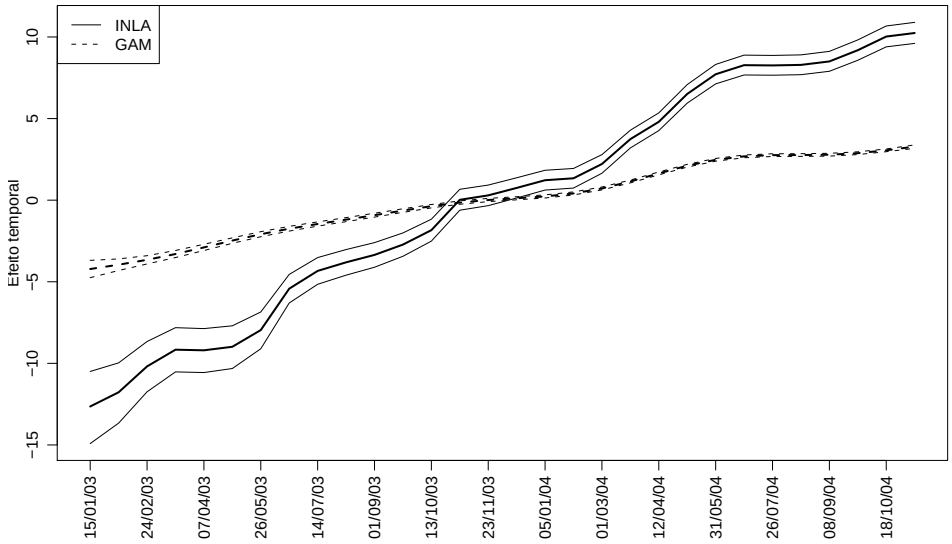

Figura 4.18: Sobreposição do efeito temporal estimado via INLA e GAM na estrutura do modelo 7.

Conforme a figura 4.18 a diferença no crescimento do efeito temporal é muito grande entre as abordagens INLA e GAM. O GAM como era esperado tende a suavizar mais o efeito, porém neste caso a suavização comparada com o efeito estimado pelo modelo *random walk* é muito grande. Pelo modelo *random walk* o efeito temporal apresenta valores que variam de $-14,91$ até $10,89$, enquanto que no *spline* ajustado pelo GAM este efeito varia de $-4,74$ até $3,38$, uma diferença bastante consideravel. Os intervalos de confiança de 95% estimados são muito menores que os intervalos de credibilidade obtidos

com o INLA. De forma geral, o único ponto em que as duas abordagens concordam é no crescimento da série com o passar do tempo.

Da mesma forma que foi comparado o efeito temporal, pode-se também comparar o efeito espacial. Para isto é interessante fazer um mapa do efeito espacial estimado pelas duas abordagens, este mapa é apresentado na figura 4.19.

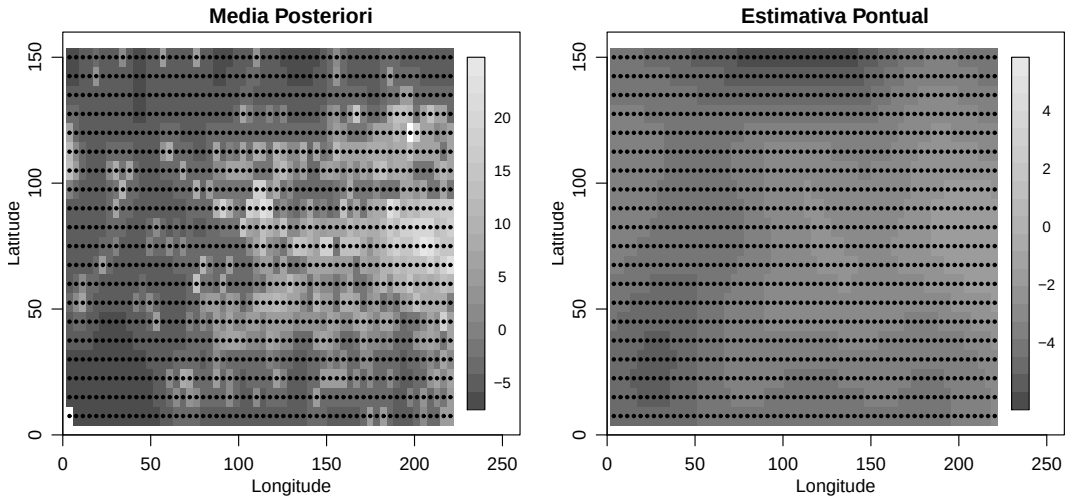

Figura 4.19: Efeito espacial estimado via INLA e GAM na estrutura do modelo 7.

Como mostra a figura 4.19 as estimativas do GAM diferem muito das médias posterioris obtidas com o INLA. Com o GAM o efeito espacial tem valores médios variando de $-5,99$ até $5,53$ ao passo que com o INLA estes valores variam de $-6,67$ até $24,84$. Com a abordagem GAM fica nítida a forte suavização do efeito, a superfície estimada não apresenta cortes abruptos como a estimada pelo INLA.

Dado esta diferença de resultados é interessante verificar qual dos modelos tem um melhor ajuste, no sentido de se aproximar mais das observações. Para isto foram feitas as predições das probabilidades de cada planta estar doente em cada ponto no tempo e espaço, a seguir foram classificadas como não doente ($p < 0,5$) e doente ($p > 0,5$). Para cada data foi calculado o percentual de plantas doentes pelos dois métodos (INLA e GAM) e o percentual observado, a figura 4.20 resume esta análise.

Pela figura 4.20 fica claro que o modelo que acompanha melhor as observações é o estimado pela abordagem INLA. É interessante notar que pela abordagem GAM o padrão na sequência de datas é muito parecido com o das observações, porém apresenta um deslocamento no sentido de subestimar as proporções em todas as datas. Olhando os intervalos de confiança obtidos pelo GAM, verifica-se claramente que são muito cur-

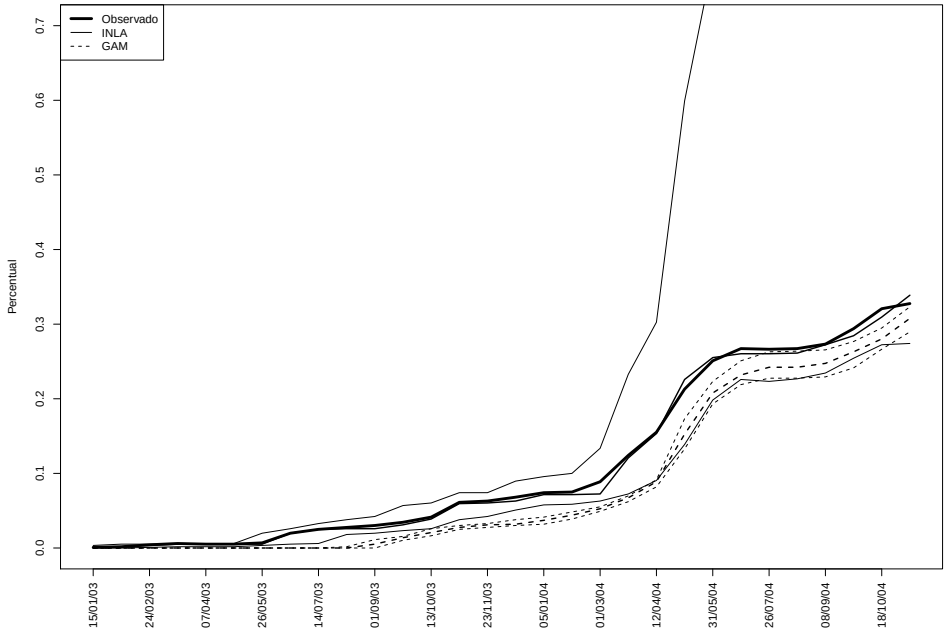

Figura 4.20: Comparação entre o percentual observado e estimado pelas abordagens INLA e GAM por data de coleta.

tos e que praticamente em nenhum momento cobrem os percentuais observados. Pela abordagem INLA o intervalo (95%) de credibilidade baseado em quantis é amplo praticamente em todas as datas cobrindo os percentuais observados. A partir de 12/04/2004 o que chama muita atenção é o aumento extremamente rápido da incerteza associada, o intervalo superior cresce muito até tocar a barreira de 1. Nesta figura o eixo y foi truncado em $0,7$ para não prejudicar a visualização dos resultados. Isso mostra que a posteriori para cada tempo, tende a ficar mais assimétrica conforme vai se aproximando das últimas datas de observações. Desta análise o que pode-se concluir é que modelar os efeitos espaciais e temporais através de *splines* e ainda olhar seu intervalo de confiança baseado na distribuição assintótica pode ser muito enganoso.

A análise apresentada na figura 4.20 compara os percentuais observados de plantas doentes para cada data de coleta com o que foi estimado por cada metodologia. Porém, ao calcular o percentual sobre todas as plantas do talhão não é possível identificar a taxa de acerto, ou seja, quando o observado para determinada planta era 'doente' e o modelo a classificou como 'doente'. A mesma situação ocorre para o caso de plantas 'não doentes'. Sendo assim, a figura 4.21 compara a taxa de acertos de cada modelo para cada tempo de coleta, ou seja, é contado quantas vezes cada modelo classificou uma planta 'doente' como 'doente' e 'não doente' como 'não doente', sobre o total de observações em cada tempo. Esta análise é apresentada na figura 4.21.

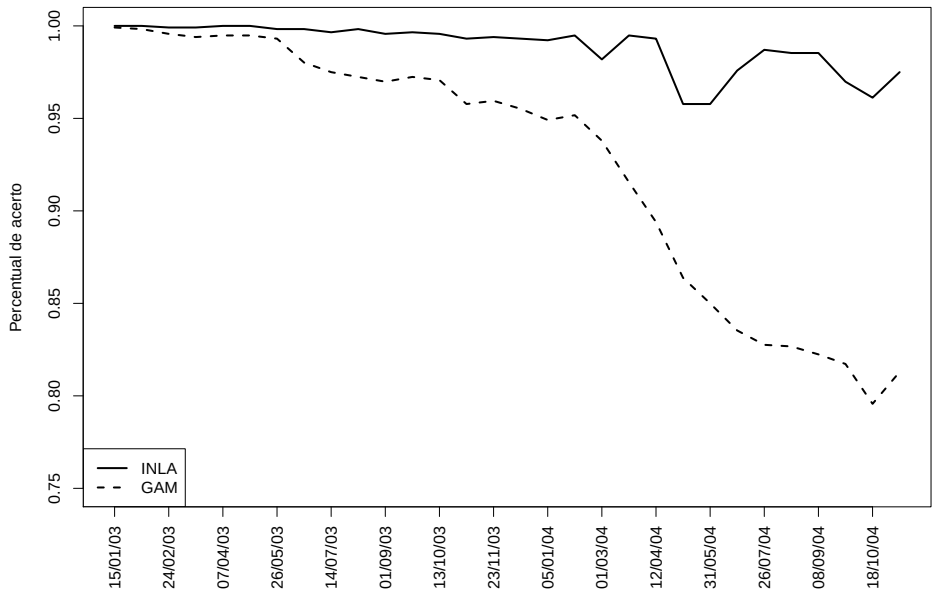

Figura 4.21: Comparação entre o percentual de acertos estimado pelas abordagens INLA e GAM por data de coleta.

A figura 4.21 mostra que a abordagem INLA tem um percentual de acertos muito bom, sempre acima do 95%. Já a abordagem GAM mostra um desempenho menos satisfatório com percentual de acertos entre 78% e 100%. Para finalizar a comparação entre as duas abordagens foram calculadas mais duas medidas de concordância, o erro quadrático médio e o erro absoluto médio. Pela abordagem INLA o erro quadrático médio foi de $0,0122$ ao passo que para a abordagem GAM esta medida foi de $0,0537$. Da mesma forma, o erro absoluto médio pelo INLA foi de $0,0454$ e pelo GAM foi de $0,1081$.Novamente os resultados mostram que a abordagem INLA apresenta melhores resultados. Embora estas medidas devem ser vistas com cautela, pois os dados são binários.

Para verificar a sensibilidade do modelo a escolha da priori, o modelo 7 foi reajustado trocando a priori Gamma(1,0.01) dos parâmetros de precisão dos efeitos espaciais e temporais, para distribuição Gamma(0.01,0.01), a distribuição a posteriori obtida com cada uma das prioris é mostrada na figura 4.22.

Pela figura 4.22 verifica-se que o modelo é pouco sensível a troca da priori. A posteriori do parâmetro de precisão do efeito espacial apresenta apenas um leve deslocamento, a média posteriori deste parâmetro é de $0.0042(0.0036; 0.0050)$ com a mudança da priori a média posteriori passou para $0.0042(0.00357; 0.0049)$, ou seja com quatro casas decimais não há diferença na média posteriori. Para o parâmetro de precisão do efeito temporal verifica-se um maior deslocamento, a média posteriori deste parâmetro é de

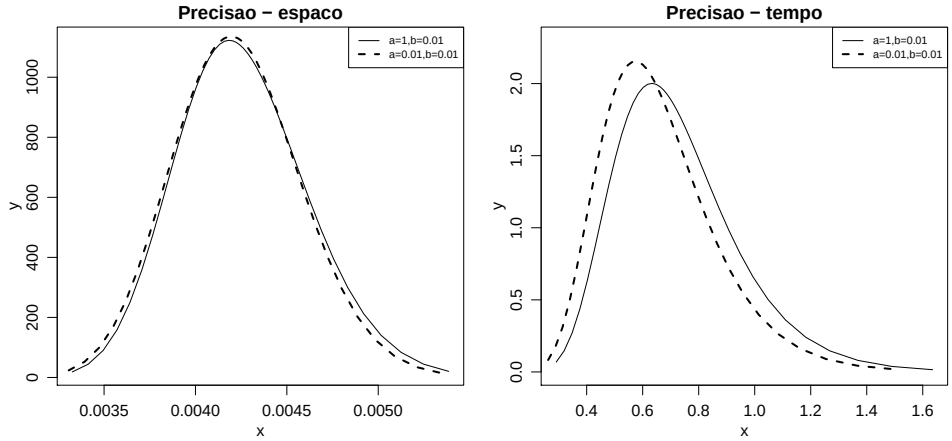

Figura 4.22: Distribuições a posteriori de acordo com a especificação de diferentes prioris para os parâmetros de precisão dos efeitos espaciais e temporais.

$0.7269(0.3878; 1.2515)$ e passa para $0.6625(0.3482; 1.1518)$ com a troca da priori. Com relação as estimativas dos efeitos espaciais e temporais, de acordo com as duas especificações de prioris não verifica-se nenhuma diferença significativa entre os efeitos, apenas um alargamento dos intervalos de credibilidade trocando a priori Gamma(1,0.01) para a Gamma(0.01,0.01). Sendo assim, optou-se por não mostrar tal análise, pois não acrescenta nenhuma nova informação.

5 CONCLUSÕES

A família dos modelos Gaussianos latentes é adaptável para uma grande quantidade de situações de análise de dados complexos, como as apresentadas no capítulo 2. Destaca-se a situação de dados espaço-temporais que é possívelmente a mais complexa que os atuais modelos estatísticos tratam. Esta dissertação revisou algumas possíveis estratégias de modelagem para dados deste tipo, inclusive tratando da situação de interação espaço-tempo.

A inferência nesta classe de modelos têm sido realizada usando métodos computacionalmente intensivos, tais como, os algoritmos MCMC *Markov Chain Monte Carlo*, porém tais métodos não estão livres de problemas. Assim, novos métodos para inferência nesta classe de modelos têm sido propostos. A dissertação revisou a abordagem INLA *Integrated Nested Laplace Approximations* proposta por Rue, Martino e Chopin (2009), que se mostrou muito eficiente para ajustar modelos altamente estruturados em diversas situações práticas.

A nova metodologia de inferência foi aplicada na análise de três conjuntos de dados previamente analisados na literatura usando outras abordagens, como modelos aditivos generalizados e modelos autoregressivos. Sempre que possível os modelos ajustados pelo INLA, foram confrontados com ajustes de modelos aditivos generalizados para verificar a concordância entre as abordagens, principalmente no que diz respeito ao modo como captam os efeitos espaciais e temporais.

Nos três conjuntos de dados analisados a concordância entre as duas abordagens foi diversificada. No primeiro conjunto referente a análise da qualidade da água em reservatórios operados pela COPEL no estado do Paraná, os resultados foram bastante parecidos, as estimativas dos efeitos dos locais de coletas foi praticamente a mesma, ou seja, com quatro casas decimais de precisão nenhuma diferença foi encontrada. Quando tratando dos efeitos temporais presente no experimento, o modelo *random walk* de primeira ordem ajustado pela abordagem INLA foi concordante com a função *spline* ajustada pelo GAM,

o que nota-se claramente é que o *spline* tende a suavizar o efeito temporal, fazendo com que o efeito não apresente mudanças abruptas como as que acontecem pelo ajuste do *random walk* via INLA, o que era esperado, devido as restrições de suavidade impostas pela abordagem GAM. Cabe ressaltar que esta foi a situação mais simples analisada na dissertação, já que, a distribuição atribuída para a variável resposta foi a Normal e não se tinha efeitos espaciais no modelo.

No segundo conjunto de dados, o interesse era investigar fatores associados a ocorrência de ovos de *Aedes aegypti* coletados em ovitrampas em Recife/PE. Neste conjunto de dados a abordagem INLA permitiu uma análise completa, considerando os quatro tipos de modelos de interação espaço-tempo, descritos na Seção 3.2. Novamente sempre que possível foi ajustado um modelo aditivo generalizado compatível para comparações dos resultados. De forma geral, os resultados foram conflitantes, as covariáveis locais, fatores de risco, recipientes grandes sem tampa, água canalizada e recipientes pequenos com tampa, que foram indicadas como significativas pelo GAM não foram significativas pelo INLA. Em geral, o que se percebe é que a abordagem INLA tende a ser mais conservadora, indicando sempre médias posterioris mais próximas do zero, do que as estimativas fornecidas pelo GAM.

Com relação as covariáveis ambientais, a forma como cada abordagem capta o efeito destas covariáveis foi muito diferente. Em termos gerais, os *splines* ajustados pelo GAM tendem a apresentar comportamento oscilante entre as covariáveis não deixando claro se existe algum tipo de relação com a resposta. Ao passo que a abordagem INLA apresentou resultados mais consistentes entre as covariáveis não indicando nenhuma covariável como relacionada com a resposta. A sequência da análise se deu, reajustando o modelo de Bonat et al. (2009), as covariáveis precipitação no mês da observação (PREC.MES1), precipitação no mês anterior a observação (PREC.MES2), umidade de dois meses antes da observação (UMID.MES3), água canalizada e recipientes grandes sem tampa, que apresentavam significância no modelo ajustado pela abordagem GAM, não mostraram significância pela abordagem INLA. Como ambos modelos apresentam termos que representam o espaço e tempo, estes termos foram comparados. Para o efeito temporal, verificou-se que a abordagem GAM suavizou bastante este efeito, o *spline* que o representa tem caminhos bem suaves. O modelo *random walk* de primeira ordem que modela o efeito temporal pela abordagem INLA, apresenta um comportamento mais volátil, porém acompanha o *spline* sendo uma versão mais ruidosa deste.

Uma análise de sensibilidade as prioris foi realizada trocando os hiperparâmetros

que indexam a distribuição Gama atribuída aos parâmetros de precisão do modelo. De forma geral, os resultados mudam pouco apenas em termos de intervalos de credibilidade que são alargados quando se troca a priori Gama(1,0.01) para Gama(0.01,0.01).

O terceiro e último conjunto de dados referente ao padrão espaço-temporal da doença leprose-dos-citrus, foi sem dúvida o mais desafiador, e o que deixou mais pontos críticos na análise. Em primeiro lugar não foi possível ajustar os modelos de interação espaço-tempo apresentados na Seção 3.2 devido a alta dimensão do talhão analisado que é de 1160 plantas em 30 tempos de observações. Dado esta restrição a análise foi restrita a modelos com apenas efeitos principais, e a comparação destes com um modelo aditivo generalizado. A distribuição atribuída a variável resposta foi a Binomial, já que, a resposta é binária. Pela abordagem INLA foi escolhido um modelo com efeito temporal e espacial estruturados. Este modelo foi confrontado com o ajuste de um modelo aditivo generalizado. Os resultados foram muito conflitantes para os dois efeitos.

Dado que os resultados foram conflitantes, optou-se por tomar alguma medida que pudesse fornecer uma forma de comparar os modelos com os dados observados, para verificar qual abordagem estava mais de acordo com as observações. Sendo assim, foi calculado o percentual de plantas doentes em cada tempo em todo o talhão, e este mesmo percentual predito por cada uma das abordagens. O resultado revelou a abordagem INLA com um resultado muito superior ao da abordagem GAM. O que chama muito atenção é os intervalos de confiança irrealísticos obtidos pela abordagem GAM, nesta situação.

Para finalizar a análise deste conjunto de dados foi feita uma análise de sensibilidade do modelo a troca de priori e verificou-se que os resultados são consistentes, e que o modelo não é sensível a troca da priori.

Da comparação entre as abordagens INLA e GAM nos três conjuntos de dados, alguns pontos foram consistentes, são eles:

- A abordagem INLA tende a ser mais conservadora que a abordagem GAM.
- Os intervalos de confiança assintóticos fornecidos pelo GAM são sempre menores que os intervalos de credibilidade baseado em quantis obtidos pelo INLA, para efeitos fixos de covariáveis.
- A modelagem de efeitos suaves de covariáveis contínuas via *splines* levou a resultados conflitantes com a modelagem de efeitos suaves de covariáveis através de modelos do tipo *random walk* no exemplo analisado.

- Efeitos espaciais e temporais modelados como funções *splines* tendem a ser super suavizados.
- Intervalos de confiança baseados na distribuição assintótica para *splines* que buscam explicar efeitos espaciais e temporais podem ser irrealísticos em algumas situações.
- As diferenças entre as abordagens se acentua em situações onde a distribuição da variável resposta não é Normal.

Deixo aqui como futuras agendas de pesquisa, simular conjuntos de dados com diferentes estruturas de interação espaço-temporal, e comparar o ajuste obtido com as abordagens INLA e MCMC. Algum tipo de comparação com a metodologia de inferência baseada em modelos mistos também deve ser investigada.

Referências Bibliográficas

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. (Ed.). **Hierarchical Modeling and Analysis for Spatial Data**. London: Chapman & Hall, 2004.

BEAL, M. J. (Ed.). **Variational Algorithms for Approximate Bayesian Inference**. Phd thesis. University College London: London, 2003.

BESAG, J. Statistical analysis of non-lattice data. **The Statistician**, v. 24, p. 179–195, 1975.

BESAG, J.; YORK, J.; MOLLIé, A. Bayesian image restoration with two applications in spatial statistics. **Annals of Institute of Statistical Mathematics**, v. 43, p. 1–59, 1991.

BISHOP, C. M. (Ed.). **Pattern Recognition and Machine Learning**. New York: Springer-Verlag, 2006. Series: Information Science and Statistics.

BIVAND, R. et al. *spdep: Spatial dependence: weighting schemes, statistics and models*. [S.l.], 2009. R package version 0.4-34.

BONAT, W. H. et al. Investigando fatores associados a contagens de ovos de Aedes Aegypti coletados em ovitrampas em Recife/PE. **Revista Brasileira de Biometria**, v. 27, p. 519–537, 2009.

BRAGA, I. A.; VALLE, D. Aedes aegypti: vigilância, monitoramento da resistência e alternativas de controle no Brasil. **Epidemiologia e Serviços de Saúde**, v. 16, p. 295–302, 2007.

CLAYTON, D. (Ed.). **Generalized linear mixed models, in W. Gilks et. al. (eds), Markov Chain MOnte Carlo in Practice**. London: Chapman & Hall, 2004.

CLAYTON, D. G.; BERNARDINELLI, L. **Bayesian methods for mapping disease risks**. Oxford University Press, 1987.

CONNOR, M. E.; MONROE, W. M. Stegomyia indices and their value in yellow fever control. **American Journal of Tropical Medicine and Hygiene**, v. 2, p. 9–19, 1923.

CZERMAINSKI, A. B. C. (Ed.). **Dinâmica espaço-temporal de populações do patossistema leprose-dos-citrus em condições naturais de epidemia**. Tese (doutorado). Universidade de São Paulo: Piracicaba, 2006.

DERISIO, J. C. **Introdução ao controle de poluição ambiental**. São Paulo, 1992.

DEY, D. K.; GHOSH, S. K.; MALLICK, B. K. (Ed.). **Generalized linear models: A Bayesian Perspective.** London: Chapman & Hall/CRC, 2000.

DIGGLE, P. J.; RIBEIRO, J. P. J. (Ed.). **Model Based geostatistics**. New York: Springer, 2006.

EGLEN, F. code by R. J. Renka. R functions by Albrecht Gebhardt. With contributions from S.; ZUYEV, S.; WHITE, D. *tripack: Triangulation of irregularly spaced data*. [S.l.], 2009. R package version 1.3-3. Disponível em: <http://CRAN.R-project.org/package=tripack>.

EIDSVIK, J.; MARTINO, S.; RUE, H. Approximate Bayesian inference in spatial generalized linear mixed models. **Scandinavian Journal of Statistics**, v. 36, p. 1–22, 2009.

FAHRMEIR, L.; TUTZ, G. (Ed.). **Multivariate Statistical Modelling Based on Generalized Linear Models**. Second. Berlin: Springer-Verlag, 2001.

FAY, R. W.; ELIASON, D. A. Laboratory studies of oviposicional preferences of aedes aegypti. **Mosquito News**, v. 25, p. 270–281, 1965.

FAY, R. W.; ELIASON, D. A. A preferred ovoposicion site as a surveillance method for aedes aegypti. **Mosquito News**, v. 26, p. 531–534, 1966.

FOCKS, D. A. A review of entomological sampling methods and indicators for dengue vectors. **Monografia na internet**, p. 1–20, 2000.

FRANCISCON, L. et al. Modelo autologístico espaço-temporal com aplicação à análise de padrões espaciais da leprose-dos-citrus. **Pesquisa agropecuária brasileira**, v. 43, p. 1677–1682, 2008.

GELMAN, A. et al. (Ed.). **Bayesian Data Analysis**. Second. Boca Raton: Chapman & HALL/CRC, 2004.

GUIRADO, N. Defensivos naturais controlam a leprose dos citrus. **O agronômico**, v. 52, p. 11–12, 2000.

HINTON, G. E.; CAMP, D. V. Keeping the neural networks simple by minizing the description length of weights. **Proceedings of the sixth annual conference on Computacinal learning theory**, p. 5–13, 1993.

KITAGAWA, G.; GERSCH, W. (Ed.). **Smoothness Priors Analysis of Time Series**. New York: Springer-Verlag, 1996.

KNORR-HELD, L. Bayesian modelling of inseparable space-time variation in disease risk. **Statistical Medical**, v. 19, p. 2555–2567, 2000.

KNORR-HELD, L.; RUE, H. On block updating in markov random field models for disease mapping. **Scandinavian Journal of Statistics**, v. 29, p. 597–614, 2002.

KRAINSKI, E. T. et al. Autologistic model with an application to the citrus sudden death disease. **Scientia Agricola**, v. 65, p. 541–547, 2008.

KUSS, M.; RASMUSSEN, C. E. Assessing approximate inference for binary Gaussian process classification. **Journal of Machine Learning Research**, v. 6, p. 1679–1704, 2005.

LANG, S.; BREZGER, A. Bayesian p-splines. **Journal of Computational and Graphical Statistics**, v. 13, p. 183–212, 2004.

LIMA, R. R. et al. Uma comparação de técnicas baseadas em quadrats para caracterização de padrões espaciais em doenças de plantas. **Revista Brasileira de Biometria**, v. 24, p. 7–26, 2006.

MARQUES, J. P. R. et al. Lesões foliares e de ramos de laranjeira-doce causadas pela leprose-dos-citros. **Pesquisa Agropecuária Brasileira**, v. 42, p. 1531–1536, 2007.

MARROQUIN, J. L. et al. Gauss-Markov measure field models for low-level vision. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 23, p. 337–348, 2001.

MARTINO, S.; RUE, H. **Implementing approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations: A manual for the inla-program**. Department of Mathematical Sciences, 2008. Technical Report no 2.

MINKA, T. P. Expectation propagation for approximate Bayesian inference. **Uncertainty in Artificial Intelligence**, v. 17, p. 451–369, 2001.

MONTEIRO, A. M. et al. **SAUDAVEL - Bridging the Gap between Research and Service in Public Health Operational Programs by Multi-Institutional Networking Development and use of Spatial Information Technology Inoovative Tools**. Divisão de processamento de imagens, 2006.

MURTA, R. D. P. with contributions from Duncan Murdoch; GPC library by A. *gpclib: General Polygon Clipping Library for R*. [S.l.], 2009. R package version 1.4-4. Disponível em: <http://CRAN.R-project.org/package=gpclib>.

NELDER, J. A.; WEDDERBURN, R. M. Generalized linear models. **Journal of the Royal Statistical Association - Series A**, v. 135, p. 370–384, 1972.

PEBESMA, E. J.; BIVAND, R. S. Classes and methods for spatial data in R. *R News*, v. 5, n. 2, p. 9–13, November 2005. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

PHILIPPI, J. A.; ROMERO, M. A.; BRUNA, G. C. **Curso de Gestão Ambiental**. Barueri, 2004.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2009. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

REGIS, L. N. et al. Developing new approaches for detecting and preventing Aedes aegypti population outbreaks: bases for surveillance, alert and control system. **Memórias do Instituto Oswaldo Cruz**, v. 103, p. 50–59, 2008.

RIBEIRO, J. P. J. et al. **Análise estatística de variáveis de qualidade da água em reservatórios da COPEL**. Divisão de Meio Ambiente - Departamento de Recursos Ambientais, 2008.

ROBERT, C. P.; CASELLA, G. (Ed.). **Monte Carlo Statistical Methods**. First. New York: Springer-Verlag, 1999.

RODRIGUES, J. C. V. (Ed.). **Relações patógeno-vetor-planta no sistema leprose-dos-citros**. Tese (doutorado). Universidade de São Paulo: Piracicaba, 2000.

RUE, H. Fast sampling of Gaussian Markov random fields. **Journal of the Royal Statistical Society, Series B**, v. 63, p. 325–338, 2001.

RUE, H.; HELD, L. (Ed.). **Gaussian Markov Random Fields: Theory and Applications**. London: Chapman & Hall, 2005.

RUE, H.; MARTINO, S. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. **Journal of statistical Planning and Inference**, v. 137, p. 3177–3192, 2007.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. **Journal Royal Statistical Society B**, v. 71, p. 319–392, 2009.

SMITH, A. F. M. et al. Progress with numerical and graphical methods for practical Bayesian statistics. **Journal of statistical Planning and Inference**, v. 36, p. 75–82, 1987.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistical Society, Series B**, v. 64, p. 583–639, 2001.

STONE, C. J. et al. Polynomial splines and their tensor products in extended linear modeling. **Annals of Statistics**, v. 25, p. 1371–1470, 1997.

TAUIL, P. L. Aspectos críticos do controle do dengue no brasil. **Cadernos de Saúde Pública**, v. 18, p. 867–871, 2002.

VENABLES, W. N.; DICHMONT, C. M. Glms, gams and glmms: an overview of theory for applications in fisheries research. *Fisheries Research*, v. 70, p. 319–337, 2004.

WEST, M.; HARRISON, J. (Ed.). **Bayesian Forecasting and Dynamic Models**. 2. ed. New York: Springer-Verlag, 1997.

WOOD, S. N. (Ed.). **Generalized additive models: Introduction with R.** Boca Raton: Chapman and Hall, 2006.

WOOD, S. N. Gams with gcv smoothness estimation and gamms by reml/pql. *R package version*, 2008. Acesso em 26/01/2010. Disponível em: <http://cran.r-project.org/web/packages/mgcv>.

ANEXO A – Apêndice

O objetivo deste apêndice é mostrar de forma rápida como especificar modelos no pacote INLA (MARTINO; RUE, 2008). Conforme visto em toda a dissertação esta abordagem permite o ajuste de modelos altamente estruturados, sendo a situação mais complexa tratada no texto, o ajuste de modelos com interação espaço-temporal no segundo exemplo. Sendo assim, optou-se por exemplificar ajuste dos modelos via INLA através dos códigos do exemplo de contagens de ovos do mosquito *Aedes aegypti* coletados em ovitrampas em Recife/PE, por este exemplo cobrir as mesmas especificações que os outros dois exemplos, e além disso, cobrir os modelos de interação espaço-temporal.

Antes de começar propriamente a análise é necessário carregar alguns pacotes adicionais.

```
> require(gpclib)
> require(tripack)
> require(sp)
> require(spdep)
> require(INLA)
```

O começo da análise é ler a base de dados. Após esta estar posta de forma adequada em um arquivo do tipo *.txt* ou *.csv*, pode ser lida no R através do comando *read.table()*. O seguinte código ilustra a leitura da base de dados.

```
> dados <- read.table("newdados",header=TRUE)
> dados <- dados[order(dados$DATA_COLETA),]
```

O INLA trabalha sempre com dados indexados para descrever os padrões espaciais e temporais, por isso é importante a base de dados estar ordenada, de uma forma que o analista consiga indexar estes efeitos sem grandes manipulações na base de dados.

```
> dados$espaco <- rep(0:79,124/4)
> dados$tempo <- rep(1:124,each=80/4)
> dados$espaco.iid <- dados$espaco
> dados$tempo.iid <- dados$tempo
```

O código acima mostra como indexar as localizações espaciais (espaço) e os períodos de tempo. Cada efeito é indexado por uma simples sequência de números. Para o efeito espacial é obrigatório que a sequência começe em ZERO, pois posteriormente este indexador vai ser linkado a um arquivo *.graph* que descreve o arranjo espacial dos dados. Para o efeito temporal não há restrições, como no exemplo tem-se 124 data de coleta foi simplesmente feita uma sequência de 1 até 124 e repetida 20 vezes. Se o modelo for contar com dois efeitos espaciais por exemplo, sendo um estruturado e outro não estruturado, a base de dados deve conter uma coluna para cada efeito e com nomes diferentes, como ilustrado no código acima.

Como neste exemplo os dados são por pontos (armadilhas) e deseja-se tratar estes como área é necessário que estas áreas sejam formadas com base na malha de armadilhas, um procedimento para formar as áreas é a tecelagem de Voronoi.

```
> nomes = sort(unique(dados$COD_ARMADILHA))
> coord.x <- c()
> coord.y <- c()
> for(i in 1:80){
+ coord.x[i] = unique(
               dados[which(dados$COD_ARMADILHA == nomes[i]),]$coords.x1)
+ coord.y[i] = unique(
               dados[which(dados$COD_ARMADILHA == nomes[i]),]$coords.x2)}
```

Os contornos do poligono podem devem ser lidos

```
> poligono <- read.table("/home/wagner/Mestrado/ModelosINLA/polBT",header=T)
```

Para construção da tecelagem foi programada uma função chamada de *voronoi()* baseada no pacote *tripack* que é disponibilizada na página *http://www.leg.ufpr.br/papercompanions*. Devido a estrutura de classe usada pelo pacote *tripack* baseado pacote *gpclib* foi necessário programar uma outra função para converter um objeto da classe *gpclib* para *sp* um formato mais geral de representação

de dados espaciais em R, que também já fornece uma forma de construir a estrutura de vizinhança entre as armadilhas através da função *poly2nb()*.

```
> source("voronoi.R")
> mapa <- voronoi(coords.x1=coord.x,coords.x2=coord.y,poligono=poligono)
> source("grp2sp.r")
> mapa1 = grp2sp(mapa,ID=nomes)
> mapa.nb1 <- poly2nb(mapa1)
```

Feito isso já tem-se condições de escrever um arquivo *.graph* para ser baseado ao INLA que descreve o arranjo espacial do conjunto de dados. Este arquivo pode facilmente ser escrito de dentro do R usando o comando *cat()*. Importante notar sempre que os índices espaciais sempre começam em ZERO.

```
> cat(80,file="recife.graph",append=TRUE,fill=TRUE,sep=" ")
> for(i in 1:80){
+ cat(c(i-1, length(mapa.nb1[[i]]),mapa.nb1[[i]]-1), sep=" ",
+     append=TRUE,fill=TRUE,file="recife.graph")}
```

Neste ponto pode-se escrever as equações dos modelos com efeitos principais, como segue:

```
> formu1 <- NRO_OVOS ~ 1
> formu2 <- NRO_OVOS ~ f(espaco.iid,model="iid") + f(tempo.iid,model="iid")
> formu3 <- NRO_OVOS ~ f(espaco,model="besag",graph.file="recife.graph") +
+                      f(tempo,model="rw1")
> formu4 <- NRO_OVOS ~ f(espaco,model="besag",graph.file="recife.graph") +
+                      f(espaco.iid,model="iid") +
+                      f(tempo,model="rw1") + f(tempo.iid,model="iid")
```

A formu1 é o modelo apenas com o intercepto. A formu2 é o modelo com efeitos espaciais e temporais não estruturados, a função *f()* diz que este é um efeito aleatório e a forma do modelo é especificado pelo argumento *model*. Da mesma forma, a formu3 apresenta o modelo com efeitos espaciais e temporais estruturados, onde aparece o arquivo *recife.graph* escrito anteriormente, o efeito temporal com o argumento *model='rw1'* mostra que agora o termo temporal tem uma estrutura seguindo um modelo *random walk* de

primeira ordem. E a última equação condensa a duas anterior para ajustar um modelo com efeitos espaciais e temporais estruturados e não estruturados.

Uma vez escrita as fórmulas dos modelos pode-se proceder o ajuste pela função *inla()*.

```
> mod1 <- inla(formu1,family="nbinomial",
+              control.inla=list(strategy="laplace"),
+              control.compute=list(dic=TRUE,cpo=TRUE,mlik=TRUE),
+              data=dados)
```

Neste exemplo um objeto chamado *mod1* recebe o ajuste do modelo. Basta indicar a fórmula para o ajuste, a distribuição da variável resposta, neste caso Binomial Negativa, a estratégia de integração neste caso Laplace, se for de interesse pode pedir para calcular medidas de diagnósticos e comparação como o DIC, CPO e verossimilhança marginal e por último indicar o conjunto de dados.

Para ajustar modelos com interação espaço-tempo é necessário que as matrizes de precisão deste modelos sejam passadas ao INLA, manualmente. O primeiro passo para o ajuste é então montar tais matrizes, para isto foram programadas algumas funções auxiliares, conforme mostra o código abaixo. Novamente as funções extras podem ser obtidas em *http://www.leg.ufpr.br/papercompanions*.

```
> source("mat.espacial.R")
> source("mat.temporal.R")
> espacial = mat.espacial(mapa.nb1,nrow=80,ncol=80)
> temporal <- mat.temporal(ncol=124,nrow=124)
> espacial.nstru <- diag(80)
> temporal.nstru <- diag(124)
```

Neste código foram montadas quatro matrizes sendo duas estruturadas e duas não estruturadas. Além disso, restrições de soma zero, devem ser impostas para garantir a identificabilidade dos termos do modelo. Estas restrições são passadas através de uma matriz.

```
> source("restricoes.R")
> A <- restri(n.dados=9920,n.tempo=124,n.espaco=80)
> e <- rep(0,204)
```

Note que como os dados são coletados em grupos, a cada tempo 60 armadilhas deixam de ser avaliadas, sendo assim, para um modelo com interação espaço-tempo estas são consideradas como NA, porque para a construção da matiz é necessário que o conjunto de dados esteja completo. O último passo é escrever a matriz do modelo da forma como o programa INLA sabe ler, para isto foi programada uma função.

```
> source("mat.INLA.R")
> inte1 <- kronecker(espacial.nstru,temporal.nstru)
> mat.INLA(inte1,name="inte1.txt")
> matriz.tipo1 <- read.table("inte1.txt")
> names(matriz.tipo1) <- c("linha","coluna","corpo")
> inte2 = kronecker(espacial.nstru,temporal)
> mat.INLA(inte2,name="inte2.txt")
> matriz.tipo2 <- read.table("inte2.txt")
> names(matriz.tipo2) <- c("linha","coluna","corpo")
> inte3 = kronecker(espacial,temporal.nstru)
> mat.INLA(inte3,name="inte3.txt")
> matriz.tipo3 <- read.table("inte3.txt")
> names(matriz.tipo3) <- c("linha","coluna","corpo")
> inte4 = kronecker(espacial,temporal)
> mat.INLA(inte4,name="inte4.txt")
> matriz.tipo4 <- read.table("inte4.txt")
> names(matriz.tipo4) <- c("linha","coluna","corpo")
```

Montadas as matrizes pode-se escrever a equação do modelo, por exemplo, para o modelo 8 é dada abaixo.

```
> forma8 = NRO_OVOS ~ f(espaco,model="besag",graph.file="recife.graph") +
+                     f(espaco.iid,model="iid")
>                     f(tempo,model="rw1") + f(tempo.iid,model="iid") +
+                     f(int,model="generic0",
+                     Cmatrix=list(i=matriz.tipo4$linha,
                      j=matriz.tipo4$coluna,
+                     Cij=matriz.tipo4$corpo),extraconstr=list(A=A,e=e),
+                     param=c(1,0.01),constr=TRUE)
```

E o ajuste é feito como anteriormente. Note que na estratégia de integração foi alterado de 'laplace' para 'GAUSSIAN', isto foi feito porque segundo (MARTINO; RUE, 2008) não está claro com a aproximação de 'laplace' trata as restrições nesta classe de modelos.

```
> mod8 <- inla(forma8,family="nbinomial",
               control.inla=list(strategy="GAUSSIAN"),
+              control.compute=list(dic=TRUE,cpo=TRUE,mlik=TRUE),
               data=dados)
```