MODELAGEM DA DISTRIBUIÇÃO ESPAÇO-TEMPORAL DA BROCA DO CAFÉ (*Hypothenemus hampei* Ferrari) EM UMA CULTURA DA REGIÃO CENTRAL COLOMBIANA

RAMIRO RUIZ CÁRDENAS

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Mestre em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
Estado de São Paulo - Brasil
Abril - 2002

MODELAGEM DA DISTRIBUIÇÃO ESPAÇO-TEMPORAL DA BROCA DO CAFÉ (*Hypothenemus hampei* Ferrari) EM UMA CULTURA DA REGIÃO CENTRAL COLOMBIANA

RAMIRO RUIZ CÁRDENAS

Engenheiro Agrônomo

Orientadora: Prof.ª Dr.ª **Clarice Garcia Borges Demétrio**

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Mestre em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
Estado de São Paulo - Brasil
Abril - 2002

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Ruiz Cárdenas, Ramiro
Modelagem da distribuição espaço-temporal da broca do café (*Hypothenemus hampei* Ferrari) em uma cultura de região central colombiana / Ramiro Ruiz Cárdenas. - - Piracicaba, 2002.
120 p.

Dissertação (mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2002.
Bibliografia.

1. Brocas (Insetos nocivos) 2. Café 3. Estatísticas espaciais I. Título

CDD 633.73

AGRADECIMENTOS

O autor expressa seus mais sinceros agradecimentos às seguintes pessoas por sua valiosa contribuição e apoio na realização deste trabalho:

Clarice Garcia Borges Demétrio

Renato Martins Assunção

Paulo Justiniano Ribeiro Jr.

Roseli Aparecida Leandro

Edna Afonso Reis

Peter Diggle

Alunos do CPG em Estatística e Experimentação Agronômica, ESALQ – USP.

SUMÁRIO

MODELAGEM DA DISTRIBUIÇÃO ESPAÇO-TEMPORAL DA BROCA DO CAFÉ (*Hypothenemus hampei* Ferrari) EM UMA CULTURA DA REGIÃO CENTRAL COLOMBIANA

Autor: RAMIRO RUIZ CÁRDENAS
Orientadora: Prof.ª Dr.ª CLARICE GARCIA BORGES DEMÉTRIO

RESUMO

O estudo da distribuição de pragas em espaço e tempo em sistemas agrícolas fornece informação importante sobre os mecanismos de dispersão das espécies e sua interação com fatores ambientais. Esse tipo de estudos também é de muita ajuda no desenvolvimento de planos de amostragem, na otimização de programas de manejo integrado de pragas e no planejamento de experimentos. O objetivo deste trabalho foi comparar vários modelos hierárquicos na modelagem da variação espaço-temporal da infestação da broca do café visando produzir mapas de risco da infestação que descrevam adequadamente o processo de infestação. Foram usadas diferentes combinações de efeitos aleatórios representando variabilidade não estruturada, com diferentes escolhas de distribuições a priori para os parâmetros e os hiperparâmetros dos modelos. Foram também usados diferentes esquemas de vizinhança para representar a correlação espacial dos dados. O ajuste dos modelos foi feito usando métodos MCMC. A estatística deviance e funções de perda quadrática foram usadas para a comparação entre modelos. Os resultados são apresentados como uma seqüência de mapas de risco de infestação.

SPATIO-TEMPORAL HIERARCHICAL MODELLING OF THE COFFEE BERRY BORER (*Hypothenemus hampei* Ferrari) DISPERSION IN COLOMBIA

Author: RAMIRO RUIZ CÁRDENAS
Adviser: Prof.ª Dr.ª CLARICE GARCIA BORGES DEMÉTRIO

SUMMARY

Study of agricultural pests distribution in space and time provides important information about the species dispersion mechanisms and its interaction with environmental factors. It also helps the development of sampling plans, the integrated pest management and planning of experiments. The aim of this work was to compare several hierarchical models in modelling the spatio-temporal variation of the coffee berry borer infestation in order to produce risk maps. Different combinations of random effects representing spatially structured and unstructured variability were used, with different prior distributions for the parameters and hyperparameters. Also different neighbourhood schemes were used to represent the spatial correlation of the data. The model fitting was done using MCMC methods and deviance and squared loss function were used for the comparison between models. The results are presented as a sequence of risk maps.

1 INTRODUÇÃO

A distribuição de artrópodos no campo tem uma considerável heterogeneidade espacial, à qual têm sido associados vários fatores ambientais. Estes fatores podem interagir entre si, e com relações inter e intra-espécies, determinando a distribuição das espécies e tendendo a estabilizar sua presença num ecossistema.

Descrições detalhadas da distribuição espacial de populações de insetos raramente têm sido abordadas na literatura, embora elas sejam de grande importância no estudo das necessidades ambientais e do comportamento do movimento das espécies, podendo ser usadas para gerar hipóteses sobre processos ecológicos subjacentes ou sugerir mecanismos que os originam (Dale, 1999). Além disso, este tipo de estudos é útil no desenvolvimento de estratégias de manipulação de *habitats* que tendam a favorecer, por exemplo, o uso de espécies benéficas em programas de manejo integrado, assim como no desenvolvimento de planos de amostragem, e no entendimento das relações presa-predador e dos processos de competição intra-específica. A distribuição espacial de insetos tem também uma implicação muito grande no planejamento de experimentos no campo, embora raramente seja levada em consideração.

O padrão espacial de populações de insetos tem sido pouco pesquisado, em parte pelo esforço intensivo de amostragem requerido para obter tal informação, mas também devido às limitações prévias em metodologia estatística. Tradicionalmente, os padrões de dispersão de insetos têm sido descritos usando-se índices baseados na variância entre amostras e sua relação com a média, tais como o índice de David & Moore (1954), a lei de poder de Taylor (1961), o índice de Morisita (1962), os índices de agregação de Lloyd (1967) e Iwao (1968), dentre outros. Estes índices porém, ignoram a localização espacial das amostras e, portanto, não determinam uma distribuição espacial como tal,

sendo sua capacidade para descrever padrões espaciais limitada a inferir se existe, ou não, aleatoriedade para alguma escala espacial desconhecida sob a qual os dados foram coletados. Além disso, eles geralmente requerem uma amplitude grande de densidades populacionais para serem efetivos, além de serem altamente dependentes do tamanho das unidades de amostragem. Métodos baseados na contagem de indivíduos em *quadrats* contíguos (Bliss, 1941; Greig-Smith, 1952) têm sido usados também mas eles ainda não incorporam, explicitamente, as coordenadas das unidades de amostragem e requerem que a amostragem seja feita a intervalos de espaço regulares. Além disso, há uma perda de resolução ao passar de dados pontuais para contagens por *quadrats.*

Entomologistas envolvidos no estudo de dinâmica populacional de insetos têm usado métodos de geoestatística convencional para caracterizar padrões espaciais (Borth & Hubert, 1987; Schotzko & O'Keeffe, 1989; Liebhold et al., 1991; Williams et al., 1992; Nestel & Klein, 1995; Darnell et al., 1999; Schotzko & Quisenberry, 1999), motivados pela freqüente necessidade de se considerarem muitas variáveis, algumas das quais variam continuamente, e pela grande quantidade disponível de *softwares* para processar grandes volumes de dados. Segundo Perry (1998), porém, aplicações geoestatísticas foram desenvolvidas originalmente para variáveis físicas estudadas comumente na área de solos, tais como fertilidade e teor de nutrientes, que são medidas em escalas contínuas e mostram uma estrutura de covariância estacionária, estável sobre uma área ampla. Entretanto, contagens de indivíduos de uma espécie animal ou vegetal em particular são discretas, e distribuídas freqüentemente em aglomerados e com uma grande quantidade de valores iguais a zero. Em contraste com variáveis físicas, tais contagens são altamente dinâmicas e poderiam não ter a estrutura de covariância espacial estável assumida pelos métodos geoestatísticos. Perry et al. (1993) afirmam que elas são freqüentemente caracterizadas por *clusters* isolados que podem atuar como metapopulações com graus variáveis de dispersão *intercluster*.

Perry & Hewitt (1991) criticaram as medidas tradicionais de agregação animal devido à carência de uma relação direta entre seus componentes e o movimento dos indivíduos envolvidos e por não usarem a informação espacial disponível nas amostras. Perry (1995) introduziu uma nova classe de índices para quantificar padrões espaciais

em ecologia denominada SADIE (*Spatial Analysis by Distance Indices*), numa tentativa de substituir a aproximação matemática abstrata de padrão espacial, por uma medida biologicamente mais adequada. Vários trabalhos (Holland et al., 1999; Turechek & Maden, 1999; Thomas et al., 2001; Winder et al., 2001), têm usado recentemente esta metodologia para descrever padrões espaciais nas áreas de entomologia e fitopatologia. A base do SADIE é quantificar o padrão espacial na população amostrada medindo uma distância mínima ou mínimo esforço total que os indivíduos, nas amostras observadas, deveriam gastar para se moverem para um arranjo completamente regular no qual o número de indivíduos seria igual em todas as unidades de amostragem. O grau de não aleatoriedade num conjunto de dados é quantificado comparando-se o padrão espacial observado com rearranjos em que as contagens amostradas são aleatoriamente redistribuídas entre as unidades (Perry, 1998). Esta metodologia foi recentemente estendida para permitir a detecçao de *clusters* (Perry et al., 1999) e para descrever o grau de associação espacial entre duas espécies amostradas na mesma unidade de amostragem ou da mesma espécie em dois tempos diferentes (Perry, 2002) mas continua sendo uma ferramenta para análise exploratória espacial, sem considerar propriamente aspectos de modelagem.

O presente trabalho pretende estudar aspectos da variação espaço-temporal da infestação da broca do café em condições de campo na Colômbia, usando modelos estatísticos que descrevam adequadamente a dispersão da praga num lote de café a partir de focos iniciais de infestação, e construir mapas de risco de infestação da praga que permitam identificar áreas de crescimento ou decréscimo da infestação no tempo, *clusters* e descontinuidades de infestação.

2 REVISÃO DE LITERATURA

2.1 Importância sócio-econômica do café

O café continua sendo um produto de grande importância para os países produtores do grão. O Brasil, seu primeiro produtor mundial, teve uma produção de 3,651 milhões de toneladas no ano de 2001, das quais exportou 1,3 milhões (correspondente a 24,5% das exportações mundiais), gerando divisas em torno de U$1393 milhões (Ministério da Agricultura, 2002). Segundo Costa et al. (1995), a cultura do café no Brasil está distribuída em 1572 municípios. Como gerador de mão-de-obra, o setor absorve quatro milhões de pessoas na produção e dez milhões nos demais segmentos tais como comércio, indústria e serviços.

O terceiro produtor mundial, a Colômbia, produziu, no ano de 2001, 720 mil toneladas de grãos, das quais exportou 564 mil, correspondente a 10,6% das exportações mundiais, por um valor de U$918 milhões (Federación Nacional de Cafeteros de Colômbia, 2002). A atividade cafeeira gera nesse país 800 mil empregos (37% do emprego agrícola e 8% do emprego total).

2.2 A broca do café

2.2.1 Generalidades

A broca do café *Hypothenemus hampei* Ferrari (Coleoptera: Scolytidae) tem sido descrita como a praga mais importante da cafeicultura no mundo (Baker et al., 1993; Murphy & Moore, 1990). Este inseto causa sérias perdas na produção e na qualidade do café ao infestar os frutos em desenvolvimento, os quais fornecem à broca um lugar para criar a sua progênie, juntar-se e se resguardar de predadores e condições climáticas

adversas (Le Pelley, 1968). Esta praga, originária da África equatorial, foi introduzida acidentalmente no Brasil, provavelmente, em 1913 (Bergamin, 1943). Atualmente, está presente em todas as regiões cafeeiras do país. Sua intensidade de infestação é variável de região para região; entretanto, seu ataque é mais intenso nas culturas onde são adotados espaçamentos mais densos ou naquelas pertencentes à espécie *C. canephora*, a exemplo do Espírito Santo e Rondônia. Na Colômbia, foi registrada pela primeira vez em 1988, e atualmente está presente em mais de 700 mil ha (85% da área plantada).

Para reduzir as perdas de café, conservando ao mesmo tempo o ambiente e evitando altos custos, é essencial o desenvolvimento de programas eficazes de manejo integrado de pragas. O sucesso de tais programas, porém, depende em grande parte do desenvolvimento e da validação de procedimentos confiáveis de amostragem. Estes, por sua vez, requerem o conhecimento dos padrões de disposição espacial das populações da praga em relação a características da planta de café, tais como padrões de floração e estado de desenvolvimento dos frutos, e de variáveis climáticas. É assim, clara, a importância deste tipo de estudo.

2.2.2 Etiologia da broca

Alguns aspectos do padrão de dispersão e estratégias de amostragem foram estudados para a broca do café na América Central por Sánchez e Ramírez (1984), Baker (1989), Baker et al. (1989), Decazy et al. (1989), Barrera (1992) e Lacayo e Guharay (1992). Entretanto, esses trabalhos não levaram em conta a localização espacial das amostras nem o efeito da escala espacial sobre a estimação desses padrões de dispersão. Remond et al. (1993) usaram semivariogramas para estudar a distribuição espacial da broca em El Salvador, mas não conseguiram ajustar uma curva de semivariograma teórica, mostrando que a distribuição da praga não é homogênea, e que o ataque parece expandir-se segundo um gradiente. Todos esses trabalhos, porém, afirmam que a praga tem um padrão de distribuição agregado no campo.

Baker (1984) e Mendoza et al. (1993) afirmam que, no caso da broca do café, odores podem ter um papel dominante na agregação já que, provavelmente, o dano feito nos frutos pelo inseto incrementa a liberação de compostos voláteis produzidos pela planta

hospedeira e/ou pela broca os quais incrementariam a atração específica de outras brocas da mesma espécie. O nível de agregação da praga pode ser influenciado também pelo microclima da parcela de café. A preferência da broca por ambientes mais sombreados foi observada por Hargreaves (1926) em Uganda, e posteriormente por Barrera (1992) na Guatemala. Outro fator que contribui para a agregação é a distribuição dos frutos nas plantas, já que a distribuição das florações dos ramos mais baixos para os mais altos, através dos sucessivos períodos de crescimento da cultura, vai concentrando a maior quantidade de frutos em certos ramos da planta cada ano. É importante considerar, também, a disponibilidade de frutos com idade adequada para serem infestados, a qual é influenciada pelo padrão de colheitas e pelo número de florações registradas ao longo do ano.

O movimento da broca do café, provavelmente, não está restrito à dispersão local, podendo, também, ocorrer em escalas maiores. Baker (1984) demonstrou, em laboratório, que a broca tem capacidade de voar durante períodos de até três horas. Baker (1999) registrou um incremento anual rápido na infestação de um lote de café na Colômbia durante a época em que culturas mais velhas e de baixa produção, que ficavam por perto, foram submetidas a uma poda drástica ou "decepa" (esta atividade, geralmente, é feita nos dois primeiros meses de cada ano), obrigando as populações de broca destas culturas a migrar para procurar refúgio em frutos de culturas mais novas. Isso evidencia o fato de que a broca pode migrar em massa fora da parcela na qual se estabeleceu originalmente, causando rapidamente infestações fortes nas culturas adjacentes. Segundo Sreedharan et al. (1994), a dispersão da broca é também grandemente ajudada pelo vento. Portanto, é razoável esperar algumas diferenças na estimação de parâmetros para um modelo em particular avaliado em diferentes escalas espaciais.

A amplitude de hospedeiros da broca do café está limitada à espécie *Coffea spp*. e seu ciclo biológico é estritamente dependente da planta de café. Estudos feitos por Mathieu et al. (1997) evidenciaram que a dinâmica das populações da broca ao longo do ano pode ser dividida em três fases: uma fase de estabelecimento, caracterizada pelo movimento do inseto de frutos da safra anterior para os primeiros frutos que começam a

amadurecer na safra seguinte; uma segunda fase, chamada de multiplicação, que toma lugar sobre várias gerações da broca ao mesmo tempo, à medida que os frutos vão amadurecendo; e uma terceira fase inter-estações, caracterizada como um período pós-colheita, em que a maioria dos frutos que ficam na cultura estão secos, embora apareçam alguns frutos verdes na planta. Este período é crucial para a dinâmica subseqüente das populações de *H. hampei*, porque a população em frutos secos constitui um refúgio através do qual ocorre a reinfestação, e os frutos verdes, embora não forneçam um substrato adequado para oviposição, representam a única fonte de nutrição disponível para as fêmeas. Durante esta terceira fase, a sobrevivência da broca depende da sua capacidade para localizar um lugar receptivo tão rapidamente quanto seja possível.

Para as condições da zona cafeeira central colombiana, Ruiz et al (1996) verificaram que durante todo o período de formação do fruto de café podem ocorrer até duas gerações da broca, mas que só depois dos 120 dias de idade do fruto, a broca começa a encontrar condições ótimas para seu estabelecimento e multiplicação, sendo que para infestações em frutos de uma idade menor as fêmeas ficam esperando dentro dos frutos o momento apropriado para ovipositar. Porém, uma vez estabelecida a colônia de insetos, sua dinâmica pode ser afetada pela oferta de frutos presentes no campo (Cure et al., 1998), ou por condições ambientais (Gutierrez et al., 1998), especialmente a temperatura (Costa e Villacorta, 1989). A chuva pode influenciar tanto direta como indiretamente a infestação da broca; a ação direta ocorre durante o período de trânsito, atuando como sinal para a praga abandonar os frutos velhos presentes no chão e iniciar a infestação dos frutos da nova safra (Baker, 1984). Por outro lado, a ação da chuva manifesta-se indiretamente, atuando como um regulador nos processos de florescimento e formação dos frutos do cafeeiro, condicionando assim, o fornecimento de alimento para a broca. A umidade relativa, de modo geral, exerce influência direta sobre a broca, atuando sobre os frutos, principalmente os secos, que permanecem na planta, ou os caídos ao solo, após a colheita. O excesso de umidade provoca a podridão dos grãos do solo, desfavorecendo o desenvolvimento do inseto; por outro lado, a umidade muito baixa causa a seca do fruto, reduzindo a multiplicação da broca. Além disso, a umidade

relativa alta fornece condições para o desenvolvimento de patógenos que podem exercer um controle natural da praga.

2.3 Modelagem espaço-temporal

Em ecologia de insetos, a modelagem tem sido usada, principalmente, para estudar relações inseto-*habitat*, inseto-inseto ou ambas, em sistemas homogêneos (reais ou artificiais), e para examinar questões pertinentes aos efeitos da variabilidade espacial e da dispersão das espécies sobre a estabilidade e a persistência das populações (Hastings, 1990). Muitos modelos teóricos, para o estudo da dinâmica espacial de insetos, são formulados como mapas de látices ligados que representam sistemas dinâmicos com tempo discreto, espaço discreto, e uma classe contínua, na qual o grau de ligação entre subpopulações no sistema espacial é definido claramente por regras de dispersão ou migração.

O movimento espacial de insetos é modelado diferentemente, dependendo do tipo de sistema que esteja sendo considerado. Em sistemas contínuos (em tempo e espaço), é comum o uso das equações de reação-difusão para modelar o movimento espacial (Okubo, 1980; Rudd & Gandour, 1985), embora a matemática envolvida neste tipo de modelos seja complexa. Segundo Andersen (1991), quando os organismos estão distribuídos em aglomerados, as equações de reação-difusão não são a ferramenta mais apropriada. Uma alternativa é modelar o movimento espacial com regras de dispersão que especificam vizinhos ou conexões locais entre subpopulações, tais como o movimento da torre (ortogonal), o bispo (diagonal) e a rainha (ortogonal e diagonal) no jogo de xadrez (Sokal & Oden, 1978), ou com regras para ligar globalmente todas as subpopulações.

Brewster & Allen (1997) desenvolveram uma estrutura, baseada em equações íntegro-diferenciais, para criar um modelo determinístico espaço-temporal da dispersão de populações de insetos em tempo discreto usado como uma ferramenta, para estudar a dinâmica de insetos em sistemas regionais de produção agrícola. O modelo combina um submodelo de simulação da população com um mapa de recursos que descreve a estrutura (tipo e arranjo espacial) dos recursos dos insetos num *habitat* heterogêneo.

Além disso, o movimento espacial entre subpopulações no sistema é incorporado, explicitamente, no modelo.

Os modelos determinísticos, segundo Gibson & Austin (1996), são mais apropriados para representar a expansão de epidemias sobre grandes escalas espaciais, mas eles não são adequados para estudos sobre escalas menores, freqüentemente observadas em sistemas experimentais. Para este último propósito, há benefícios evidentes em considerar modelos que representem populações como coleções discretas de indivíduos, ao invés de contínuas. Além disso, os modelos deveriam incorporar a aparente aleatoriedade inerente a sistemas biológicos, e isto leva a considerar modelos epidemiológicos espaço-temporais estocásticos.

Há um considerável número de trabalhos na modelagem de processos estocásticos na área de Fitopatologia, e alguns na área de Entomologia (Gibson & Austin, 1996); mas, em geral, esses trabalhos tratam do comportamento a longo prazo ou longa escala de epidemias, e não podem ser facilmente aplicados às escalas espaciais e temporais sobre as quais os pesquisadores realmente observam as epidemias no campo. Em particular, tem havido poucos trabalhos direcionando o problema de ajuste de modelos espaço-temporais estocásticos a seqüências temporais de mapas de doenças.

Smyth et al. (1992) descrevem a estimação de parâmetros em um modelo baseado em cadeias de Markov, para estudar a expansão da antracnose em *Stylosanthes scabra*. Buckland & Elsten (1993) igualmente aplicaram métodos baseados em cadeias de Markov na modelagem da dinâmica espaço-temporal de espécies selvagens num ecossistema. Estes exemplos compartilham a suposição de que o estado de dois indivíduos (ou sítios) quaisquer, num ponto dado no tempo, é independente, condicional ao estado da população num ponto prévio no tempo. Esta suposição é válida se a amostragem é freqüente e o progresso da doença entre pontos no tempo correspondentemente limitada. Em alguns casos, porém, esta suposição não é válida, complicando o ajuste de modelos estocásticos espaço-temporais.

Gibson & Austin (1996) propuseram um método para ajustar modelos estocásticos espaço-temporais em tempo contínuo, a seqüências de mapas de doenças em sistemas agrícolas observados em tempos discretos, que dispensa a necessidade de

assumir independência no estado da doença entre indivíduos. Eles usaram um único parâmetro em seus modelos que caracterizava a relação entre pressão infectiva e distância. A estimação dos parâmetros foi feita via máxima verossimilhança. Segundo os autores, a técnica pode ser aplicada a modelos mais complexos envolvendo um maior número de parâmetros. Eles, igualmente, usaram mapas simulados de doenças cujo grau de agregação iguala ao de uma epidemia real, como critério para comparar modelos alternativos. Gibson (1997) estendeu esta abordagem usando métodos de Monte Carlo, implementados via cadeias de Markov (MCMC) para a estimação dos parâmetros.

2.3.1 Modelos geoestatísticos

O formato básico para dados geoestatísticos univariados é (x_i, y_i), $i = 1 \ldots n$, sendo que x_i identifica a posição espacial tipicamente em um plano bidimensional e y_i é uma medição escalar tomada na posição x_i. A variável de medição y_i pode ao menos em princípio estar localizada em qualquer lugar numa região de estudo A. Usualmente, é assumido que o delineamento de amostragem para as posições x_i é estocasticamente independente do processo que gera as medições y_i e que a região A é um subconjunto fixo de pontos em $\Re^2$.

A forma básica de um modelo geoestatístico é um processo estocástico de valor real $\{Y(x): x \in A \subset \Re^2\}$ o qual é tipicamente considerado como uma realização parcial de um processo estocástico $\{Y(x): x \in \Re^2\}$. Freqüentemente, o processo de medição Y_i pode ser registrado como uma versão ruidosa de uma variável aleatória subjacente $S(x_i)$, o valor na posição x_i do processo de interesse $\{S(x): x \in A \subset \Re^2\}$. $S(x)$ é chamado de sinal. Esse modelo básico pode ser estendido para um modelo com dois componentes: um processo estocástico $S(x)$, e um modelo estatístico para as medições, $\mathbf{y} = (y_1, ..., y_n)^T$ condicional a $\{S(x): x \in A \subset \Re^2\}$.

Um processo estocástico é considerado Gaussiano, se a distribuição conjunta de $S(x_1)$, ..., $S(x_n)$ é multivariada Gaussiana para qualquer inteiro n, e um conjunto de posições x_i. O processo é estacionário se, a esperança e a variância de $S(x)$ é a mesma para todo x, e a correlação entre $S(x_i)$ e $S(x_j)$ depende unicamente da distância entre as posições x_i e x_j. O processo é estacionário e isotrópico se, adicionalmente, esta

correlação depende unicamente de $u = || x_i - x_j ||$, a distância euclideana entre x_i e x_j, para qualquer par de inteiros $0 < i, j \leq n$.

Segundo Diggle & Ribeiro (2000), o modelo gaussiano, estacionário e isotrópico para um conjunto de dados (x_i, y_i), $i = 1 \ldots n$, é definido pelas seguintes suposições:

(i) $\{S(x): x \in A \subset \Re^2\}$ é um processo gaussiano estacionário com média μ, variância σ^2, função de covariância $\gamma(x_i, x_j) = \text{cov}\{S(x_i), S(x_j)\} = \sigma^2 \rho(u)$, e função de correlação $\rho(u) = \text{corr}\{S(x_i), S(x_j)\}$, em que $u = || x_i - x_j ||$;

(ii) condicionalmente a $\{S(x): x \in \Re^2\}$, os y_i são realizações mutuamente independentes dos Y_i, normalmente distribuídos com média condicional $E[Y_i \mid S(.)] = S(x_i)$ e variância condicional σ^2.

Uma forma equivalente de expressar este modelo é

$$Y_i = S(x_i) + Z_i, \quad i = 1 \ldots n,$$

sendo que Z_i são erros aleatórios mutuamente independentes distribuídos normalmente com média zero e variância σ^2.

Para especificar este modelo, é necessário especificar apenas os momentos de primeira e segunda ordem, ou seja, a função de média $\mu(x) = E[Y(x)]$ e a função de covariância $\gamma(x_i, x_j) = \text{cov}\{Y(x_i), Y(x_j)\}$.

Uma outra classe de modelos muito útil na prática são os processos gaussianos para os quais a média é variável, podendo depender de covariáveis e erros aleatórios, porém, com estrutura de covariância estacionária. Para tais situações, $Y(x) - \mu(x)$ é um processo estacionário gaussiano com média zero.

A matriz de covariâncias especificada através da função de correlação $\rho(u)$ deve ser positiva definida. Esta restrição assegura que para qualquer inteiro m, qualquer conjunto de posições x_i e qualquer constante real a_i, a combinação linear $\sum_{i=1}^{m} a_i S(x_i)$ terá uma variância não-negativa.

Uma das famílias paramétricas de correlação, dada sua flexibilidade, sugerida por Diggle & Ribeiro (2000), é a classe de funções de correlação de dois parâmetros proposta por Matérn (1986), cuja forma é:

$$\rho(u; \phi, \kappa) = \{ 2^{\kappa-1} \Gamma(\kappa) \}^{-1} (u / \phi)^{\kappa} K_{\kappa}(u / \phi)$$

em que $K_{\kappa}(\cdot)$ denota a função de Bessel modificada de terceira classe, com ordem dada pelo parâmetro κ. O parâmetro $\phi > 0$ determina a taxa para a qual a correlação tende a zero ao se aumentar u. O parâmetro $\kappa > 0$, é chamado de ordem do modelo Matérn e determina a suavidade analítica do sinal S(x).

Cressie & Huang (1999) introduziram novas famílias paramétricas de funções de covariância estacionárias não-separáveis, para processos aleatórios indexados em espaço e tempo.

Às vezes, o mecanismo de amostragem sugere o uso de uma distribuição de probabilidade não Gaussiana, condicional aos valores de uma quantidade de interesse desconhecida que varia espacialmente. Alguma flexibilidade pode ser obtida usando um parâmetro extra λ, definindo uma transformação Box-Cox como em Christensen et al. (2001). Este modelo, porém, tem limitações e os autores sugerem que seja usado como um ponto de partida para tratar dados não Gaussianos antes de tentar modelos mais complexos. Uma alternativa interessante, proposta por Diggle et al. (1998), e implementada por outros autores como Christensen et al. (2000), é o uso de modelos explicitamente não Gaussianos, como é o caso dos modelos lineares generalizados com efeitos aleatórios correlacionados espacialmente.

Os modelos lineares generalizados mistos (GLMM) como definidos por Breslow & Clayton (1993), são extensões dos modelos lineares generalizados (McCullagh & Nelder, 1989) que permitem fontes adicionais de variabilidade devida a efeitos aleatórios não observáveis. Um caso particular são os GLMM espaciais, em que os efeitos aleatórios são modelados por um campo gaussiano espacial.

Uma generalização da formulação condicional do modelo gaussiano, estendida para permitir informação de covariáveis é:

(i) $\{S(x): x \in \Re^2\}$ é um processo gaussiano estacionário com média zero, variância σ^2 e função de correlação $\rho(u, \phi)$;

(ii) $d_1(x)$, ..., $d_p(x)$ é um conjunto de p variáveis explicativas espacialmente referenciadas e

(iii) condicionalmente sobre $\{S(x): x \in \Re^2\}$, as variáveis aleatórias Y_1, ..., Y_n são mutuamente independentes.

Neste modelo $d_k(x)$ representa variáveis explicativas que contribuem para a variação espacial em Y; $S(x)$ representa a variação espacial não explicada em Y; e σ^2 representa a variação não espacial não explicada em Y.

Então, a função $E[Y_i \mid S(x_i)]$ que varia espacialmente só através do valor de $S(x)$ na posição x_i toma a forma $E[Y_i \mid S(x_i)] = M(x_i)$, sendo que a média condicional $M(x)$, está relacionada a $S(x)$ através de uma função de ligação g de modo que

$$g(M(x)) = \eta(x) = S(x) + \mathbf{d}(x)^T\boldsymbol{\beta}$$

sendo $\mathbf{d}(x)^T \in \Re^p$ um vetor de covariáveis associadas à posição x e $\boldsymbol{\beta} \in \Re^p$ um vetor de parâmetros de regressão.

A distribuição de $Y(x) \mid S$ tem uma densidade que depende unicamente da média condicional $M(x)$. Esta densidade pertence à família exponencial da forma

$$f(z, \mu) = \exp [z\, g_c(\mu) + b(z) - a(g_c(\mu))], \quad z \in \Omega$$

em que $\Omega \subseteq \Re$ é o suporte da densidade, μ é o parâmetro média e a, b, g_c são funções reais; g_c é chamada função de ligação canônica. Assume-se que a função de ligação é contínua, diferenciável e estritamente crescente. Estas condições são satisfeitas no caso especial em que $g = g_c$. A função de ligação não pode ser escolhida arbitrariamente, já que a amplitude de $g(M(x))$ deve estar na linha real inteira. Para a distribuição binomial a função de ligação canônica é $g(\mu) = \log[\mu / (N - \mu)]$.

Existem várias alternativas para se fazer inferência em GLMM espaciais. Os métodos mais freqüentemente usados são o de pseudo-verossimilhança, também chamada quasi-verossimilhança penalizada (PQL) ou verossimilhança-*h* (Breslow & Clayton, 1993), e a inferência Bayesiana (Diggle et al., 1998; Christensen et al., 2000). Segundo Clayton (1996), embora o método PQL tenha um desempenho bastante aceitável em muitas situações, geralmente, é bastante laborioso, já que envolve, em cada iteração, a inversa de uma matriz quadrada de tamanho igual ao número total de parâmetros. Por outro lado, os métodos baseados em inferência bayesiana, têm ganho espaço nos últimos anos, devido ao desenvolvimento de métodos computacionais de simulação, particularmente o caso dos métodos Monte Carlo implementados via cadeias de Markov, conhecidos popularmente como métodos MCMC (*Markov chain Monte Carlo*).

2.3.2 Modelos para a construção de mapas de risco de doenças

Tem havido ultimamente um crescimento muito rápido no número de trabalhos que combinam modelos lineares generalizados mistos e métodos MCMC para dados espaciais. Uma das principais áreas de aplicação é em epidemiologia humana, seguindo os trabalhos iniciais de Clayton & Kaldor (1987) e Besag et al. (1991) para mapear riscos de doenças e descrever sua variação no espaço ou simultaneamente em espaço e tempo. Aqui, os dados tomam a forma do número de casos de uma doença em particular e o correspondente tamanho da população, para um determinado número de regiões espaciais discretas.

2.3.2.1 Modelo clássico de riscos relativos de doenças

Em sua forma mais simples, os dados para estudos de variação geográfica de riscos de doenças consistem em pares (y_i, n_i), sendo n_i o número de indivíduos sob risco e yi o número de casos da doença na área i, $i = 1, ..., I$. A análise deste tipo de dados, usualmente, inicia-se supondo que Y_i segue uma distribuição binomial com parâmetros n_i e $\boldsymbol{p}_i$, sendo $\boldsymbol{p}_i$ a probabilidade de risco desconhecida da doença, na área i. Assume-se

que os Y_i's são mutuamente independentes (num contexto Bayesiano os Y_i's seriam condicionalmente independentes, dada a probabilidade de risco p_i).

Para o caso de doenças pouco freqüentes e não contagiosas (n_i grande e p_i pequena), o modelo binomial pode ser aproximado por um modelo Poisson, com parâmetro $m_i = E(Y_i) = n_i p_i$. Usando uma probabilidade de risco de referência, $p > 0$, o modelo pode ser rescrito como $Y_i \sim \text{Poisson}(E_i \lambda_i)$, sendo $E_i = n_i p$ o número esperado de casos da doença, obtido a partir da taxa de risco de referência p, e $\lambda_i = p_i / p$ o risco de aparecer um caso da doença na área i, relativo ao risco de referência p, conhecido como *risco relativo* em relação a p.

A escolha mais simples para a taxa de referência p é o risco global $\Sigma y_i / \Sigma n_i$ também chamado de *padronização interna*, devido a p ser obtido internamente a partir dos dados da própria região sob estudo. As taxas de referência também podem ser obtidas a partir de fontes de informação externas à área sob estudo, quando, então, ocorre uma *padronização externa*.

Pode ser facilmente demonstrado que o estimador de máxima verossimilhança de λi é a razão yi / Ei conhecida como *razão de morbidade padronizada* (RMP). Esta taxa, porém, é de difícil interpretação, já que a variância de y_i / E_i é igual a $\lambda_i E_i / (E_i)^2 = \lambda_i / E_i = \lambda_i / (n_i p)$ e desse modo inversamente proporcional às contagens populacionais n_i. Assim, os valores de RMP mais extremos são tipicamente encontrados nas áreas com as populações menores. Essa desvantagem das RMP como estimativas do risco relativo é mais óbvia para o caso de áreas sem casos da doença ($y_i = 0$), o que é bastante comum no estudo de doenças raras. Desse modo, faz-se necessário incorporar informação de outras áreas, $j \neq i$, para melhorar a estimativa de risco na área i como é descrito a seguir.

2.3.2.2 Modelo hierárquico bayesiano para mapear riscos de doenças

Considere I áreas geográficas, cada uma delas possuindo uma medida de risco ξ_i, $i = 1, ..., I$, que deve ser estimada. Segundo Assunção (2001), estes riscos podem ser absolutos, tais como taxas per capita ou por 1000 habitantes, ou alternativamente eles podem ser relativos a algum nível, geralmente aquele do valor esperado E_i que considera

a probabilidade de risco de referência para a região sob consideração. Os dados da região i correspondem aos números de casos y_i numa população de tamanho n_i. A inferência bayesiana para esses riscos é baseada na distribuição *a posteriori* do risco $\mathbf{x}$ dadas as observações $\mathbf{y}$, isto é,

$$f(\mathbf{x} \mid \mathbf{y}) \propto L(y_1, ..., y_I \mid \mathbf{x})\, f(\mathbf{x})$$

em que $L(y_1, ..., y_I \mid \mathbf{x})$ é a função de verossimilhança e $f(\mathbf{x})$ é a distribuição *a priori* do vetor de parâmetros $\mathbf{x} = (x_1, ..., x_I)^T$. Como foi descrito na seção anterior, geralmente, para o caso de doenças raras e não contagiosas, é assumido que, condicionalmente aos x_i's, os Y_i's são independentes com distribuição Poisson com média $n_i x_i$ no caso de taxas, ou $E_i x_i$ no caso de riscos relativos. Para este último caso, a função de verossimilhança de $x_1, ..., x_I$ é dada por

$$L(y_1, ..., y_I \mid \hat{\mathbf{i}}) = \prod_{i=1}^{I} \frac{(x_i E_i)^{y_i}}{y_i!} \exp(-x_i E_i).$$

O logaritmo do risco relativo, desconhecido para a i-ésima área, $h_i = \log(x_i)$, pode ser modelado como a soma de uma média global, denotada por m, que expressa o nível geral do logaritmo do risco relativo através do mapa, e um efeito específico por área, denotado por w_i, que representa a diferença entre o logaritmo do risco relativo para a área i e a média global. Deste modo, o modelo genérico para a i-ésima área pode ser escrito como

$$h_i = \log(x_i) = m + w_i. \qquad (1)$$

A equação (1) é um exemplo simples de um modelo linear generalizado misto, com uma estrutura de erro Poisson, uma função de ligação logarítmica, um efeito fixo (m) e um conjunto de efeitos de área aleatórios ($w_1, ..., w_I$). Para completar a

especificação do modelo é necessário somente descrever a distribuição *a priori* de $\mathbf{w}$, dados seus hiperparâmetros associados.

A forma mais simples *a priori* de modelar o logaritmo dos riscos relativos é determinar um valor da média geral, para o qual são deslocadas as estimativas específicas por área, sendo que este deslocamento depende da estabilidade intrínseca dos estimadores, e não da localização das áreas no mapa. Estes são chamados modelos de heterogeneidade não estruturada. Os dois modelos paramétricos mais estudados são, a distribuição gama para os riscos relativos, $\boldsymbol{x}_i$, e uma distribuição Gaussiana para o logaritmo dos riscos relativos, $\boldsymbol{h}_i$. Segundo Clayton & Bernardinelli, (1992) o primeiro leva a resultados algébricos mais simples, com expressões de forma fechada para as médias *a posteriori*, mas o segundo tem consideravelmente um potencial maior para extensão. Assim, uma *priori* Gaussiana comumente usada é dada por

$$P(\boldsymbol{w} \mid \boldsymbol{t}) \propto \left\{ -\frac{\boldsymbol{t}}{2} \sum_i (\boldsymbol{w}_i - \overline{\boldsymbol{w}})^2 \right\},$$

em que $\overline{\boldsymbol{w}}$ é a média aritmética de todos os logaritmos de taxas da doença e $\boldsymbol{t} = 1 / \boldsymbol{s}^2$ é um parâmetro que representa o inverso da variabilidade (precisão) dos logaritmos de taxas da doença. Note-se que log[P(ω | τ)] é proporcional à soma dos desvios quadrados de $\boldsymbol{w}_i$ em relação a sua média e pode ser interpretada como uma penalização por heterogeneidade não estruturada. Segundo Clayton & Bernardinelli (1992), esta função não é uma distribuição própria já que ela não varia em relação a $\overline{\boldsymbol{w}}$, sendo necessário adotar uma restrição linear (usualmente $\overline{\boldsymbol{w}} = 0$) ou omitir o efeito fixo $\boldsymbol{m}$ do modelo (1), em cujo caso $\overline{\boldsymbol{w}}$ tomará o papel de $\boldsymbol{m}$ no modelo.

Modelos *a priori* mais estruturados permitem incorporar interdependência entre os $\boldsymbol{x}_i$'s, isto é, considerar que o estimador de risco relativo em uma área dada é fortemente influenciado pelas estimativas de áreas geograficamente adjacentes, e unicamente de uma forma indireta pelas estimativas das outras áreas do mapa. Como um resultado disto, cada estimativa é deslocada para uma média local ao invés de uma

média global, com relações mais fortes para áreas geograficamente mais próximas. A classe de modelos baseada em campos aleatórios Markovianos como definidos em Besag (1974) tem sido a mais intensivamente estudada.

Os efeitos w_i, no modelo (1), podem ser considerados como um substituto para aquelas variáveis desconhecidas, ou não observadas, que afetam o risco. Se estas variáveis fossem realmente observadas, algumas delas não teriam uma estrutura geográfica definida, enquanto que outras visualizariam padrões geograficamente estruturados de variação. Se o primeiro tipo de variáveis fosse dominante, os w_i's verdadeiros exibiriam heterogeneidade não estruturada. Por outro lado, se o segundo tipo de variáveis fosse o predominante, os w_i's verdadeiros exibiriam heterogeneidade geograficamente estruturada, em que os valores de risco relacionados a um par de áreas vizinhas seriam, geralmente, mais parecidos do que para um par de áreas escolhidas arbitrariamente. A abordagem Bayesiana requer que seja especificada uma distribuição multivariada *a priori* para os w_i's que refletirá a crença *a priori* em relação ao tipo de heterogeneidade geográfica que eles representam. Uma hipótese de heterogeneidade não estruturada levará a uma suposição *a priori* de que os w_i's são independentes e identicamente distribuídos. Por outro lado, uma hipótese de heterogeneidade estruturada levará à incorporação dentro do modelo *a priori* de uma correlação espacial para os w_i's. Os modelos anteriores são chamados modelos simples, já que eles modelam unicamente a heterogeneidade estruturada ou a heterogeneidade não estruturada.

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

$$\eta_i = \log(\lambda_i) = \phi_i + \theta_i. \quad (2)$$

Segundo Assunção (2001), a primeira parte do modelo não possui uma estrutura espacial e representa efeitos de pequena escala, ou diferenças entre uma área e outra que não ultrapassam suas fronteiras geográficas. Estes efeitos poderiam, por exemplo, ser ações de saúde pública ou características socio-econômicas locais não compartilhadas pelas áreas vizinhas. Assim, por serem efeitos específicos de cada área, podem ser modelados como efeitos aleatórios independentes.

A segunda parte do modelo possui uma estrutura espacial e é definida de modo a considerar que áreas vizinhas são mais semelhantes em termos de seu risco relativo do que duas áreas arbitrariamente escolhidas na região. Portanto, este componente representa efeitos de longa escala, de alcance geográfico maior do que as fronteiras das áreas que estão sendo consideradas. Muitas variáveis observadas, ou não, que afetam o risco de uma doença em particular, possuem esta característica. Por exemplo, características ambientais, em geral, mudam suavemente no espaço e assim, áreas próximas tendem a ter valores similares para estas variáveis.

2.3.2.2.1 Escolha das distribuições *a priori*

Em termos de modelagem, é comum assumir que os elementos do primeiro componente da equação (2) possuem, *a priori*, distribuições normais independentes com média zero e variância σ_ϕ^2. Assim, $\boldsymbol{\phi} = (\phi_1,..., \phi_I)^T$ tem uma distribuição normal multivariada composta de termos independentes. Esta distribuição depende de um parâmetro de precisão $\tau_\phi = 1/\sigma_\phi^2$, que, em geral, é desconhecido. Este parâmetro controla o grau de dispersão dos efeitos aleatórios sem estrutura espacial, sendo inversamente relacionado à variabilidade não estruturada dos riscos relativos. Uma distribuição *a priori* alternativa para o componente ϕ_i é a *t*-Student com média zero, parâmetro de escala τ_ϕ e ν graus de liberdade desconhecidos (Pascutto et al., 2000). A vantagem desta distribuição é que ela tem uma cauda mais pesada do que a normal e assim a inferência é mais robusta em áreas com valores de risco extremos. A

distribuição *t*-Student já tinha sido sugerida por Besag & Higdon (1999) como uma alternativa para atenuar a vulnerabilidade de *prioris* gaussianas na presença de *outliers* na análise bayesiana de experimentos agrícolas de campo.

A distribuição *a priori* escolhida para o segundo componente $\mathbf{q} = (\theta_1, ..., \theta_I)^T$ que tem uma estrutura espacial é tipicamente um campo aleatório Markoviano Gaussiano. Esta distribuição, além de ter uma estrutura de vizinhança espacial, depende de um parâmetro desconhecido adicional τ_θ, inversamente relacionado à variabilidade dos θ_i's e que mede o grau de similaridade espacial entre estes componentes do risco relativo. Este tipo de *prioris* merece especial atenção e é descrito em detalhe a seguir.

2.3.2.2.1.1 Distribuições *a priori* baseadas em campos aleatórios Markovianos

Os efeitos aleatórios espacialmente estruturados de um modelo hierárquico Bayesiano seguem uma distribuição contínua. O modelo *a priori* mais popular para este tipo de efeitos é aquele em que a distribuição conjunta é normal multivariada. Em particular, as auto-regressões condicionais também conhecidas como modelos CAR (Besag, 1974), têm sido amplamente usadas na literatura. Estes modelos especificam a distribuição *a priori* condicional de θ_i como sendo normal, com média que depende da média dos θ_j, $j \neq i$, nas áreas da vizinhança da área i. A forma básica desta distribuição é dada pela expressão

$$\theta_i \mid \theta_j : j \neq i \sim \mathrm{N}\left(\mu_i + \sum_{j=1}^{I} c_{ij}(\theta_j - \mu_j),\ \sigma_i^2\right). \qquad (3)$$

Besag (1974), mediante o teorema da fatorização, demonstrou que a especificação condicional auto-gaussiana (3) implica que

$$\mathbf{q} \sim \mathrm{N}(\boldsymbol{\mu}, \tau\mathbf{S})$$

em que $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_I)^T$ e $\mathbf{S} = (\mathrm{I} - \gamma\mathbf{C})^{-1}\mathbf{M}$ é uma matriz de correlação n x n, simétrica positiva-definida; $\mathrm{C} \equiv (c_{ij})$ é uma matriz de pesos n x n cujo (i, j)-ésimo elemento, c_{ij}, reflete a associação espacial entre as áreas i e j (a influência de θ_j sobre a esperança de θ_i). Assume-se que a matriz de pesos $\mathbf{C}$ é simétrica para assegurar que $(\mathrm{I} - \gamma\mathbf{C})$ seja invertível; $\mathbf{M} \equiv \mathrm{diag}(m_1, ..., m_I)$ é uma matriz diagonal n x n conhecida e escolhida de modo que Σ seja simétrica e positiva definida; γ é um parâmetro que controla a magnitude da correlação espacial e $\boldsymbol{\nu} = (\sigma_1^2, \sigma_2^2, ..., \sigma_I^2)$ é um vetor de parâmetros de escala.

A escolha da matriz $\mathbf{S}$ é importante, já que ela indica como a associação espacial entra no modelo. $\mathbf{S}$ será simétrica se e somente se $c_{ij}m_{jj} = c_{ji}m_{ii}$, $i, j = 1, ..., I$. Para assegurar que $\mathbf{S}$ seja positiva definida, o parâmetro γ deve pertencer ao intervalo $(\gamma_{\min}, \gamma_{\max})$, em que $1/\gamma_{\min} < 0 < 1/\gamma_{\max}$ são o menor e o maior autovalores de $\mathbf{M}^{-1/2}\mathbf{C}\mathbf{M}^{1/2}$. Na prática, freqüentemente é esperada uma dependência espacial positiva, e desse modo é comum tomar γ no intervalo $(0, \gamma_{\max})$.

A maioria das abordagens, incorporando um modelo condicional para dependência espacial, iniciam especificando-se um conjunto de pesos espaciais para uso em (3). Estes pesos tradicionalmente estão associados a um conjunto de áreas vizinhas que contribuem com um peso positivo à esperança condicional de θ_i, com $c_{ij} = 0$ para as demais regiões e $c_{ii} = 0$. Há um grande número de escolhas para M, C e γ que satisfazem à condição de que Σ seja uma matriz de variâncias, simétrica e positiva-definida. Um modelo muito usado para associações espaciais é a auto-regressão condicional intrínseca gaussiana (Besag et al., 1991; Besag & Kooperberg, 1995), dada pela expressão

$$\theta_i \,|\, \boldsymbol{\theta}_{-i} \sim \mathrm{N}\left(\frac{\sum_{j \in \partial_i} c_{ij}\theta_j}{\sum_{j \in \partial_i} c_{ij}}, \frac{\tau^2}{\sum_{j \in \partial_i} c_{ij}} \right)$$

em que ∂_i é o conjunto de áreas adjacentes (vizinhas) à área i e $\boldsymbol{\theta}_{-i} = \{\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_I\}$.

Fazendo-se $c_{ij} = 1$ para $j \in \partial_i$, tem-se o modelo

$$\theta_i \mid \theta_{-i} \sim N\left(\bar{\theta}_{(i)}, \frac{\nu_\theta^2}{r_i}\right) \qquad (4)$$

em que $\bar{\theta}_{(i)} = \frac{\sum_{j \in \partial_i} \theta_j}{r_i}$ e r_i é o número de vizinhos da área *i*. Sob este modelo, o risco na área *i* é suavizado para um risco médio definido pelo conjunto de áreas na vizinhança da área *i*, com variância inversamente proporcional ao número de vizinhos. A comparação com o modelo (3) mostra que o modelo (4) corresponde à escolha de

$$\mu = 0, \ \alpha = \alpha_{\max} = 1, \ M_{ii} = \frac{1}{r_i} \ \text{ e } \ c_{ij} = \begin{cases} \frac{1}{r_i}, & \text{se as áreas } i \text{ e } j \text{ são vizinhas} \\ 0, & \text{em caso contrário.} \end{cases}$$

Este modelo não leva a uma matriz de precisão Σ positiva-definida. Para se ver isto, note-se que na *i*-ésima linha de **I - C**, tem-se só uma entrada com o valor 1 e r_i entradas com valores $-1/r_i$. Assim, a soma em todas as linhas de **I - C** é zero, indicando que a matriz Σ é singular (tem posto *n*-1) e portanto, não tem inversa. Devido a isto, o modelo (4) define uma distribuição *a priori* imprópria para os efeitos aleatórios espacialmente estruturados $\boldsymbol{\theta}$ em (2). A variância σ_θ^2 em (4), a qual somente é interpretável condicionalmente, não é mais proporcional à variância marginal, já que o modelo é não estacionário (é possível ter uma média arbitrária para cada θ_i) e, portanto, a distribuição conjunta de $\boldsymbol{\theta}$ não existe. Segundo Congdon (2001), a natureza imprópria desta distribuição surge devido ao fato de que ela especifica unicamente diferenças nos riscos da doença e não seus níveis. Porém, segundo Wakefield et al. (2000), uma grande vantagem do modelo não estacionário (4), é que a forma da dependência espacial pode variar através da região sob estudo.

Cressie & Chan (1989) propuseram uma escolha para a matriz de variância-covariância de um processo CAR, a qual pode ser usada como uma distribuição *a priori* própria para os efeitos aleatórios espacialmente estruturados do modelo (2) com as seguintes escolhas para **M**, **C** e g

$\gamma \in (\gamma_{min}, \gamma_{max})$, $M_{ii} = \dfrac{1}{E_i}$, o inverso do número esperado de casos na i-ésima área, e

$$c_{ij} = \begin{cases} \sqrt{\dfrac{E_j}{E_i}}, & \text{se as áreas } i \text{ e } j \text{ são vizinhas} \\ 0, & \text{em caso contrário.} \end{cases}$$

Ao contrário da auto-regressão intrínseca (4) que seleciona $g = g_{\max} = 1$, indicando uma forte suposição *a priori* em relação à dependência espacial, o modelo proposto por Cressie & Chan (1989) permite aos dados sugerir prováveis valores para g dentre aqueles para os quais Σ é positiva-definida já que, *a priori*, este parâmetro pode variar sobre qualquer valor de seu espaço paramétrico $(g_{\min}, g_{\max})$.

Uma dificuldade com a aproximação condicional é que, freqüentemente, não fica claro como escolher os pesos c_{ij} e a vizinhança ∂_i. Vários autores (Clayton & Kaldor, 1987; Besag et al., 1991; Waller et al., 1997, dentre outros) têm considerado as áreas i e j como vizinhas se elas compartilham um limite comum. Isto seria razoável no caso de todas as regiões serem de tamanho similar e arranjadas num padrão regular, mas em estudos de mapeamento de doenças, as regiões são freqüentemente subdivisões políticas, tendo pouco a ver com a etiologia das doenças ou com fatores de risco comum. Além disso, devido em parte à aglomeração da população em risco em núcleos urbanos e à mobilidade destas populações, as regiões podem variar amplamente tanto em área geográfica quanto em tamanho da população. De acordo com Conlon & Waller (1998), para alguns fatores de risco ambiental, a contiguidade regional pode não ser a definição mais satisfatória de vizinhança. Vários outros esquemas de vizinhança são possíveis (Cliff & Ord, 1981), embora tais formulações devessem ser consideradas à luz de uma restrição de simetria (isto é, $c_{ij}m_{jj} = c_{ji}m_{ii}$). Cressie & Chan (1989) consideram uma estrutura de vizinhança que depende da distância entre centróides de áreas. Eles

determinaram a distância dentro da qual duas áreas são consideradas como vizinhas através de uma análise exploratória usando o variograma experimental. Best et al. (1999) e Conlon & Waller (1998) consideraram esquemas de vizinhança baseados em funções paramétricas de distância; estas consideram pesos que decrescem à medida que se aumenta a distância entre centróides de áreas.

Como já foi ressaltado, a distribuição *a priori* conjunta para $\mathbf{q} = (\theta_1, ..., \theta_I)$ que corresponde ao modelo (4) é imprópria já que o nível médio para cada θ_i não está definido. A distribuição *a posteriori* para os θ_i's, porém, é identificável na presença de qualquer dado informativo y_i (Besag et al., 1995; Mollié, 2000). Alternativamente, Besag & Kooperberg (1995) propuseram uma reparametrização incluindo no modelo um termo de intercepto separado, α, representando o nível médio de risco, e impondo a restrição $\Sigma_i \theta_i = 0$ para assegurar que o modelo seja identificável. A distribuição *a priori* para este termo de intercepto é totalmente não informativa, ou seja, uma uniforme variando entre $-\infty$ e $+\infty$. MacNab & Dean (2000) fizeram extensões dos modelos CAR clássicos e apresentaram outro tipo de *prioris* auto-regressivas condicionais para uso na modelagem dos efeitos aleatórios espacialmente estruturados do risco de doenças.

Também é possível o uso de formas não gaussianas para especificar a distribuição deste componente espacialmente estruturado. Por exemplo, Besag et al. (1991) e Best et al. (1999) consideram a distribuição Laplaciana (exponencial dupla) a qual leva a um modelo baseado na mediana ao invés da média das taxas nas áreas da vizinhança. Segundo Wakefield et al. (2000), isto pode ser mais apropriado quando são esperadas discontinuidades em taxas da doença entre áreas.

Cada uma das *prioris* dos componentes $\mathbf{f} = (\phi_1,..., \phi_I)^T$ e $\mathbf{q} = (\theta_1, ..., \theta_I)^T$, depende dos parâmetros τ_ϕ e τ_θ. Estes parâmetros são chamados de hiperparâmetros e são modelados por outra distribuição chamada *hiperpriori*, já que ela se refere aos parâmetros da distribuição *a priori* dos parâmetros da função de verossimilhança. Deve-se tomar especial cuidado na especificação destas hiper*prioris*. O uso de *hiperprioris* impróprias uniformes não é possível nestes casos, já que isto daria como resultado distribuições *a posteriori* impróprias. De acordo com Richardson & Monfort (2000),

quando se usarem *prioris flats* é necessário especificar limites inferiores para as variâncias, como foi feito por Richardson et al. (1995). No caso de dados com variabilidade maior do que a esperada, ainda com algumas *prioris* próprias tais como a distribuição gama ou a χ^2, é necessário especificar limites inferiores para as variâncias para evitar problemas com a estimação (*trapping*) para valores próximos de zero. Tipicamente, tem sido escolhida uma distribuição Gama $\Gamma(\hat{I}, \hat{I})$ para o inverso da variância (precisão) dos efeitos aleatórios, com $\hat{I}$ sendo um valor pequeno tal como 10^{-2} ou 10^{-3} para ambos os hiperparâmetros. Kelsall & Wakefield (1999) destacaram, porém, que ainda um*a priori* vaga como esta pode ser altamente informativa. Em particular, estas *prioris* não são consistentes com níveis muito pequenos de variabilidade nos efeitos aleatórios. Como uma alternativa, estes autores sugerem o uso de uma *hiperpriori* Gama(0,5; 0,0005), já que, em muitos contextos esta dará um intervalo plausível para o inverso da variância dos riscos relativos através da região sob estudo. A média da distribuição gama representa a suposição *a priori* do valor do parâmetro de precisão t_q ou t_f, e sua variância reflete a incerteza em relação a esta crença *a priori.* Portanto, as estimativas Bayesianas dos parâmetros de interesse no modelo poderiam ser sensíveis à escolha dos parâmetros desta *hiperpriori.* Apesar disto, Mollié (2000) sustentou que as estimativas do risco relativo e a construção de mapas parece ser robusta a esta escolha. Assunção (2001) igualmente destacou que a escolha de outras hiper*prioris* bastante diferentes da Gama(0,001; 0,001), produziu, essencialmente, as mesmas estimativas do risco relativo num modelo espaço-temporal para mapeamento de taxas de doenças. Segundo Mollié (2000), se o objetivo principal da análise é obter estimativas específicas por área de riscos de doenças, a sensibilidade à escolha das hiper*prioris* não tem conseqüências grandes; porém, se o objetivo principal é quantificar e explicar o padrão subjacente da variação em risco de doenças através da região sob estudo, deve-se tomar cuidado na interpretação das estimativas em função das *prioris* escolhidas. Uma distribuição *a priori* própria alternativa para t_f e t_q sugerida por Bernardinelli et al. (1995) e Richardson & Monfort (2000) é a χ^2, com parâmetros de escala e graus de liberdade pré-especificados. Outras estratégias para a escolha destas

hiperprioris foram igualmente sugeridas por Bernardinelli et al. (1995) e Mollié (1996, 2000).

Os parâmetros $\sigma_f^2 = 1/\tau_f$ e $\sigma_q^2 = 1/\tau_q$ controlam a magnitude dos componentes $\mathbf{f}$ e $\mathbf{q}$ respectivamente, sendo que valores pequenos da razão σ_q^2/σ_f^2 refletem um domínio da heterogeneidade não estruturada espacialmente, enquanto que valores grandes indicam um domínio da variação estruturada espacialmente. Segundo Mollié (2000), porém, a calibração desta razão é complicada pelo fato que σ_q^2 somente é proporcional à variação espacial condicional do logaritmo dos riscos relativos, sendo que cada área tem uma constante de proporcionalidade diferente, dada pelo inverso de seu número de vizinhos. A variância marginal pode ser encontrada considerando a forma da distribuição conjunta para o log χ_i, $i = 1, ..., I$. Isto leva a uma covariância *a priori* marginal teórica dada pela expressão $\sigma_q^2\mathbf{I} + \sigma_f^2(\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}$. Como foi visto anteriormente, porém, a matriz $(\mathbf{I} - \mathbf{C})$ é singular, e desse modo a covariância marginal não existe. No entanto este autor sugeriu uma forma de aproximar a variabilidade marginal dos efeitos aleatórios, calculando as variâncias marginais empíricas

$$s_q^2 = \left(\frac{1}{n-1}\right)\sum_i (\hat{e}_i - \bar{e})^2 \qquad \text{e} \qquad s_f^2 = \left(\frac{1}{n-1}\right)\sum_i (f_i - \bar{f})^2.$$

Assim, a razão s_q^2/s_f^2 pode ser usada para estimar a importância relativa de cada componente para a variação total dos efeitos aleatórios.

O modelo (1) pode ser generalizado para permitir efeitos de covariáveis, substituindo $\boldsymbol{\mu}$ por $\mathbf{Z}\boldsymbol{\beta}$, sendo $\mathbf{Z}$ uma matriz de p covariáveis conhecidas, medidas em cada área e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ um vetor de parâmetros desconhecidos. Em alguns casos essas covariáveis podem também variar no espaço, e assim, os parâmetros β_i podem ser modelados considerando uma estrutura espacial similar à modelagem dos efeitos aleatórios θ_i como foi proposto por Assunção et al (2002).

Literatura recente sobre mapeamento de doenças discute também o uso de modelos com efeitos aleatórios multi-níveis que incorporam efeitos aleatórios

hierárquicos para explicar a variabilidade em diferentes níveis da hierarquia geográfica. Um exemplo deste tipo de modelos pode ser encontrado em Dean & MacNab (2001).

2.3.2.2.2 Implementação dos modelos hierárquicos bayesianos

Juntando os elementos anteriormente descritos, é possível especificar a distribuição de probabilidade *a posteriori* conjunta dos parâmetros e dos hiperparâmetros. Esta distribuição é proporcional ao produto das distribuições dos três níveis hierárquicos considerados, isto é,

$$(\boldsymbol{x}, \boldsymbol{f}, \grave{e} \mid \mathbf{y}) \propto \left(\prod_{i=1}^{I} \frac{(\boldsymbol{x}_i E_i)^{y_i}}{y_i!} \exp(-\boldsymbol{x}_i E_i) \right) f(\boldsymbol{f}_1, ..., \boldsymbol{f}_I \mid \hat{o}_f) f(\grave{e}_1, ..., \grave{e}_I \mid \hat{o}_{\grave{e}}) f(\hat{o}_f) f(\hat{o}_{\grave{e}}). \quad (5)$$

Com base nesta distribuição *a posteriori* conjunta, é possível fazer qualquer inferência desejada a respeito do logaritmo do risco relativo. O cálculo da distribuição marginal *a posteriori* de **x** dado **y** implica integrar *a posteriori* conjunta (5) em relação a **x**. Esta integração, porém, geralmente, é analiticamente intratável. Nestes casos, o uso de métodos computacionalmente intensivos tais como os métodos MCMC, fornecem uma alternativa para obter esta densidade marginal. O mais popular deles é o amostrador Gibbs, algoritmo introduzido na estatística a partir do artigo de Geman & Geman (1984). Mediante este método, ao invés de calcular ou aproximar diretamente a densidade marginal de **x**, é possível gerar uma amostra que segue uma distribuição da densidade marginal requerida, sem ter que precisar do cálculo dessa marginal. A simulação de uma amostra, suficientemente grande, permite o cálculo da média, da variância e de outras características da densidade *a posteriori* marginal de interesse para um grau de exatidão desejado. Por exemplo, em um contexto de mapeamento de doenças, é possível calcular um intervalo de credibilidade bayesiano para o risco relativo em cada ponto do mapa, ou calcular a probabilidade de que o logaritmo do risco relativo numa área em particular exceda um nível dado, contando o número de valores simulados

que são maiores do que esse nível. Detalhes deste método podem ser encontrados por exemplo em Casella & George (1992) e Gilks et al. (1996). Embora os métodos MCMC permitam a estimação dos parâmetros de interesse, o tamanho finito da simulação introduzirá um grau de erro de aproximação conhecido como erro padrão de Monte Carlo. Amostras grandes e aproximadamente independentes (isto é, com autocorrelações baixas entre valores amostrados consecutivamente) terão erros de Monte Carlo relativamente baixos. Segundo Wakefield et al. (2000), porém, infelizmente os modelos usados em contextos de mapeamento podem exibir altas correlações entre parâmetros do modelo e incluir termos que são fracamente identificados (isto é, identificados somente por um*a priori* vaga), o que pode gerar amostras altamente correlacionadas e desse modo as simulações via MCMC devem ser efetuadas para um número grande de iterações buscando gerar uma amostra com uma precisão adequada para fazer a inferência *a posteriori.*

2.3.2.2.3 Modelos hierárquicos bayesianos espaço-temporais

Mais recentemente o interesse tem sido centrado na análise de taxas de doenças em espaço e tempo (Bernardinelli et al., 1995b; Waller et al., 1997, 1998; Knorr-Held & Besag, 1998; Xia & Carlin, 1998; Pickle, 2000; Sun et al., 2000; Assunção et al., 2001). Em sua forma mais simples, os dados para tais estudos consistem da observação do número de indivíduos sob risco, e o número de casos ou mortes para cada combinação espaço-tempo. Knorr-Held (2000) propôs uma estrutura unificada para a análise de incidência, ou dados de mortalidade em espaço e tempo que, além dos efeitos principais espaciais e temporais, leva em consideração as interações espaço-tempo. Este autor introduziu quatro tipos de interação com diferentes suposições *a priori* a respeito da inter-relação entre os parâmetros espaciais e temporais. A análise é iniciada com um modelo só de efeitos principais separáveis em espaço e tempo, em que n_{it} denota o número de indivíduos em risco na localidade i (i = 1, ..., I), no ano t (t = 1, ..., T). É assumido que o número de casos ou mortes Y_{it}, durante o ano t, segue uma distribuição binomial com parâmetros n_{it} e $\boldsymbol{p}_{it}$. A probabilidade de sucesso $\boldsymbol{p}_{it}$ é modelada num

contexto de modelos lineares generalizados, usando uma função de ligação logística para a binomial, e um preditor linear η_{it} que se decompõe aditivamente em efeitos dependentes em tempo e espaço. Assim, para o modelo só com efeitos principais é assumido que

$$\eta_{it} = \ln\{ \pi_{it} / (1 - \pi_{it}) \} = \alpha_t + \gamma_t + \theta_i + \phi_i , \qquad (6)$$

sendo α_t e γ_t efeitos temporais, representando características não especificadas do ano t que, respectivamente, visualizam, ou não, uma estrutura temporal *a priori.* Similarmente, θ_i e ϕ_i representam características não especificadas da localidade i, que, respectivamente, visualizam, ou não, estrutura espacial. Distribuições *a priori* Gaussianas multivariadas com média zero e matriz de precisão $k\boldsymbol{K}$ são atribuídas aos quatro vetores de parâmetros $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_T)^T$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_T)^T$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_I)^T$ e $\boldsymbol{\phi} = (\phi_1, ..., \phi_I)^T$ de η_{it}, em que k é um escalar desconhecido e $\boldsymbol{K}$ é uma matriz estrutural conhecida, como definida em Clayton (1996). A matriz $\boldsymbol{K}$ é diferente para cada vetor de parâmetros já que ela tenta descrever diferentes suposições sobre a inter-relação entre parâmetros dentro de cada vetor. Na presença de interação espaço-tempo, esta formulação é expandida, adicionando-se parâmetros de interação δ_{it} a (6), na forma de produtos de um dos dois efeitos principais espaciais com um dos dois efeitos principais temporais.

Bernardinelli et al (1995) propuseram um modelo de interação espaço-tempo para a análise de variação geográfica de taxas de doenças que modela o logaritmo das taxas como uma função linear do tempo, com coeficientes específicos por área e dando uma distribuição *a priori* que considera áreas vizinhas como tendo perfis de risco similares. Uma parametrização diferente deste modelo foi adotada por Assunção et al (2001) para mapear e predizer taxas de leishmaniose visceral em Belo Horizonte (MG). Eles propuseram um modelo de tendência polinomial de segundo grau específico por área que captura a aceleração e desaceleração das taxas da doença através do tempo, tendo a forma

$$q_{it} = \ln(x_{it}) = a_i + b_i(t-1) + g_i(t-1)^2$$

em que a_i representa o logaritmo da taxa da doença na área i no primeiro tempo e $b_i+g_i(2t\text{-}1)$ representa o incremento no logaritmo da taxa na área i ao passar de um tempo t para um tempo t+1. *Prioris* auto-regressivas condicionais (CAR) foram atribuídas para os três vetores de parâmetros **a** , **b** e **g**.

MacNab & Dean (2001) propuseram o uso de modelos aditivos generalizados mistos (GAMM) para a análise espaço-temporal de riscos de doenças. Estes modelos incluem efeitos aleatórios espacialmente correlacionados e componentes temporais aleatórios e fixos. Os autores usaram uma aproximação de suavização auto-regressiva local através da dimensão espacial, mas adicionaram uma suavização utilizando *splines*-B de baixa ordem sobre a dimensão temporal. Eles usaram uma suavização *spline* de efeitos fixos para desmascarar tendências gerais para toda a região considerada e *splines* de efeitos aleatórios para isolar tendências de pequena área. Van der Linde et al. (1995) tinham registrado previamente o uso de *splines* na análise de taxas de mortalidade num contexto espaço-tempo. Fahrmeir & Lang (2001) fizeram aplicações dos GAMM com efeitos aleatórios espaciais e covariáveis múltiplas, usando inferência Bayesiana.

Os trabalhos citados até aqui referem-se a casos de doenças não infecciosas tais como o câncer ou a leishmaniose, em que são observadas contagens de indivíduos em pequenas áreas sobre uma série de períodos longos de tempo (geralmente, um ano). Estes estudos geralmente têm por objetivo encontrar dependência sobre uma ou mais covariáveis que poderiam orientar sobre políticas públicas de saúde a serem seguidas, para prevenir tais doenças. A dependência espaço-temporal neste tipo de modelos é importante para uma inferência eficiente sobre os coeficientes de regressão das covariáveis. Em contraste, quando a doença é infeciosa, os períodos entre observações são, usualmente, curtos (por exemplo, uma semana). Aqui a doença é muito mais dinâmica, e o interesse é encontrar como as contagens num tempo t-1 influenciarão as contagens num tempo t. Neste caso as covariáveis ainda são importantes para explicar

heterogeneidade, mas a ênfase está centrada nos parâmetros que governam a dinâmica temporal. Cressie & Mugglin (2000) propuseram um modelo bayesiano para caracterizar heterogeneidade espaço-temporal, incorporando ferramentas de modelagem bayesiana, usadas no caso de doenças não infeciosas. O modelo supõe que a intensidade da doença numa área pequena num tempo *t* depende linearmente da mesma área, de seus vizinhos mais próximos, e de seus segundos vizinhos mais próximos, todos num tempo *t*-1. O modelo tem poder preditivo para determinar necessidades de assistência à medida que a epidemia se desenvolve.

Kaiser et al. (2002) ressaltaram as limitações na modelagem da dependência espacial através de campos aleatórios Markovianos Gaussianos para o caso de dados binários ou de contagem, e especificamente em modelos de mistura, em que há um processo espacial latente que atua para produzir o padrão espacial observado. Eles apresentaram uma metodologia que permite a construção de modelos de mistura sem restringir a distribuição de mistura espacial a ser gaussiana. Esta metodologia é baseada em distribuições condicionais da família exponencial e é ilustrada com um exemplo usando a distribuição beta-binomial.

2.4 Critérios para seleção de modelos

No contexto de modelagem clássica, a comparação de modelos comumente é feita definindo-se uma medida de ajuste, tipicamente a estatística *deviance* e, uma medida de complexidade, o número de parâmetros *livres* no modelo. Como o incremento na complexidade dos modelos vai acompanhado de um ajuste melhor, os modelos usualmente são comparados, procurando-se um equilíbrio dessas duas quantidades. Seguindo o trabalho pionero de Akaike (1973), os métodos propostos para comparar modelos freqüentemente têm estado baseados na minimização de uma medida de perda esperada sobre uma replicação do conjunto original de dados.

Infelizmente, em modelos hierárquicos complexos, nos quais os parâmetros geralmente, superam o número de observações, métodos tradicionais não podem ser aplicados diretamente. Por exemplo, os testes de razão de verossimilhança, em geral, possibilitam a escolha entre modelos somente no caso aninhado, onde há uma hipótese

nula não ambígua (Gelfand, 1996), sendo a seleção baseada em uma aproximação χ^2 assintótica para a distribuição da diferença de *deviances* o que pode não ser confiável para tamanhos amostrais pequenos. Além disso, tal aproximação freqüentemente não será aplicável para modelos não regulares (no sentido de que o número de parâmetros tende ao infinito à medida que se aumenta o número de observações). Mesmo quando as aproximações usuais assintóticas são válidas, Gelfand (1996) sustenta que o teste de razão de verossimilhança é inconsistente, já que quando o tamanho amostral tende ao infinito, a probabilidade de que o modelo superparametrizado seja selecionado, dado que o modelo reduzido é verdadeiro, não se aproxima de zero. A razão de verossimilhança dá muito peso a modelos de dimensões maiores, sugerindo a necessidade de impor uma redução ou penalização para dimensão sobre a função de verossimilhança.

O fator de Bayes tem sido amplamente usado na comparação de modelos Bayesianos, mas segundo Gelfand (1996), ele carece de interpretação no caso de modelos com *prioris* impróprias, freqüentemente usadas em especificações hierárquicas complexas, além de ser difícil de calcular para modelos com conjuntos grandes de dados. Algumas formulações sugeridas para corrigir este problema incluem o *fator de Bayes intrínseco* (Berger & Pericchi, 1996) e o *fator de Bayes fracionário* (O'Hagan, 1995). Estas dificuldades têm levado ao desenvolvimento recente de critérios de escolha de modelos Bayesianos alternativos, tais como o critério de informação da *deviance* (DIC) proposto por Spiegelhalter et al. (1998, 2002) e um critério baseado na minimização de uma função de perda quadrática preditiva proposto por Gelfand & Ghosh (1998), descritos a seguir.

2.4.1 Critério de Gelfand & Ghosh (1998)

Estes autores propuseram um critério de verossimilhança penalizada surgindo sob a distribuição preditiva *a posteriori*, que pode ser definido da seguinte forma: sejam $\mathbf{y}_{obs}$ um vetor cujos componentes são valores observados da variável de interesse Y, $\mathbf{y}_{rep}$ uma replicação de $\mathbf{y}_{obs}$ para um determinado modelo m ($\mathbf{y}_{rep}$ tem a mesma distribuição que $\mathbf{y}_{obs}$) e a uma ação a ser tomada.

Uma função de perda univariada $L(y, a)$ para o l-ésimo componente de $\mathbf{y}_{rep}$ e a ação a é definida por

$$L(y_{l,rep}, a_l ; y_{l,obs}) = L(y_{l,rep}, a_l) + kL(y_{l,obs}, a_l), \quad k \geq 0, \tag{7}$$

em que a_l pode ser vista como uma ação de compromiso, já que (7) recompensa os modelos tanto pela proximidade de a_l a $y_{l,rep}$, quanto a $y_{l,obs}$; k indica o peso dado a cada componente da função (7), sendo que $L(y_{l,rep}, a_l)$ captura a precisão das estimativas e $L(y_{l,obs}, a_l)$ avalia a qualidade do ajuste.

O critério de avaliação de modelos baseado na perda *a posteriori* preditiva é dado pela minimização em a da esperança de função de perda $L(y_{rep}, a; y_{obs})$ em $y_{l,rep}$ condicionalmente a y_{obs} e m. Dessa forma, define-se

$$\begin{aligned} D_k(m) &= \sum_{l=1}^{n} \min_{a_l} E_{y_{l,rep}|y_{obs},m} L(y_{l,rep}, a_l ; y_{obs}) \\ &= \sum_{l=1}^{n} \min_{a_l} \left\{ E_{y_{l,rep}|y_{obs},m} L(y_{l,rep}, a_l) + kL(y_{l,obs}, a_l) \right\}. \end{aligned} \tag{8}$$

Usando-se a função de perda quadrática $L(y, a) = (y - a)^2$, a equação para o l-ésimo termo em (8), com um a_l fixo, pode ser reescrita como

$$\sigma_l^{2(m)} + (a_l - m^{(m)})^2 + k(a_l - y_{l,obs})^2$$

sendo

$$\mu_l^{(m)} = \mathrm{E}(y_{l,rep} \mid y_{obs}, m) \text{ e } \sigma_l^{2(m)} = \mathrm{var}(y_{l,rep} \mid y_{obs}, m).$$

O valor de a_l que minimiza essa função é dado por

$$a_l = \frac{(m_l^m + k y_{l,obs})}{(k+1)}.$$

Substituindo-se este valor de a_l em (8) obtém-se

$$D_k(m) = \sum_{l=1}^{n} \sigma_l^{2(m)} + \frac{k}{k+1} \sum_{l=1}^{n} (\hat{\mu}_l^{(m)} - y_{l,\mathrm{obs}})^2 = P(m) + \frac{k}{k+1} G(m)$$

com

$$G(m) = \sum_{l=1}^{n} (\mu_l^{(m)} - y_{l,obs})^2 \qquad e \qquad P(m) = \sum_{l=1}^{n} \sigma_l^{2(m)}$$

sendo $G(m)$, uma medida de qualidade de ajuste e $P(m)$, um termo de penalidade. Para modelos muito simples, tanto $G(m)$ quanto $P(m)$ tendem a ter valores altos. À medida que a complexidade do modelo aumenta, essas duas estatísticas decrescem, sendo que $P(m)$ voltará a crescer quando o modelo passar a apresentar uma quantidade exagerada de parâmetros. Dessa forma, $P(m)$ punirá modelos que serão favorecidos pela estatística $G(m)$.

A constante k, geralmente, pouco influencia a escolha do modelo, como foi mostrado por Gelfand & Ghosh (1998). Um caso extremo é fazer $k \to \infty$, em cujo caso tem-se

$$D(m) = \lim_{k \to \infty} D_k(m) = P(m) + G(m).$$

Valores pequenos de *D(m)* indicam um melhor ajuste do modelo. Este critério foi aplicado por Xia & Carlin (1998) para comparar modelos hierárquicos espaço-temporais na área de mapeamento de doenças.

2.4.2 Critério de informação da deviance (DIC)

Recentemente, Spiegelhalter et al. (1998, 2002) propuseram uma generalização do critério de informação de Akaike, baseado na distribuição *a posteriori* da estatística da *deviance*

$$D(\mathbf{q}) = -2 \log p(\mathbf{y}|\mathbf{q}) + 2 \log f(\mathbf{y})$$

em que $p(\mathbf{y}|\mathbf{q})$ a função de verossimilhança para o vetor de dados observados **y**, dado o vetor de parâmetros **q** e $f(\mathbf{y})$ alguma função de padronização dos dados sobre si mesmos e desse modo não tendo impacto sobre a seleção de modelos. Sob esta aproximação, o ajuste do modelo é resumido pela esperança *a posteriori* da *deviance*

$$\overline{D} = E_{\theta|y}[D]$$

enquanto que a complexidade de um modelo é capturada pelo número efetivo de parâmetros P_D, definido como a *deviance* esperada menos a *deviance* avaliada nas esperanças *a posteriori*

$$P_D = E_{q|y}[D] - D(E_{q|y}[\grave{e}]) = \overline{D} - D(\bar{\grave{e}}).$$

Finalmente, os modelos podem ser comparados usando o *critério de informação da deviance* definido como

$$DIC = \overline{D} + P_D = 2\overline{D} - D(\bar{\grave{e}})$$

com valores pequenos de DIC indicando um melhor ajuste do modelo. Zhu & Carlin (2000), aplicaram este critério na comparação de modelos hierárquicos espaço-temporais usados na área de mapeamento de risco de doenças.

Outros critérios preditivos para a avaliação e escolha de modelos hierárquicos Bayesianos espaço-temporais num contexto de mapeamento de riscos de doenças podem ser encontrados em Waller et al. (1997) e Knorr-Held (2000).

2.5 Modelos para dados ausentes

Segundo Congdon (2001), todos os modelos para estimação de dados ausentes estão, implícita ou explicitamente, baseados na expansão dos dados observados para explicar os dados latentes ausentes. Desta forma, existe uma conexão natural com modelos que incluem variáveis sujeitas a não resposta, em que a função de verossimilhança do conjunto de dados completo consiste de um modelo de amostragem para os resultados observados juntamente com um modelo para o mecanismo de não resposta ou mecanismo dos dados faltantes (Little & Rubin, 1986).

Vários tipos de abordagens têm sido aplicados a problemas de dados ausentes (uma completa revisão pode ser encontrada em Little & Rubin, 1986), mas, segundo Congdon (2001), certas restrições aplicam-se a todos eles. Por exemplo, a simples omissão dos dados ausentes da análise, somente leva a inferências válidas se os dados são ausentes completamente ao acaso, isto é, os valores desses dados ausentes são uma simples amostra aleatória de todos os valores dos dados. Uma suposição menos restritiva é a de ausência ao acaso, sob a qual a probabilidade de que uma observação esteja ausente depende dos dados observados mas não dos dados ausentes. De modo geral, considere $\mathbf{Y} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{aus}\}$ como sendo o conjunto completo de dados observados e ausentes, e $R = \{R_i\}$ um indicador binário tal que $R_i = 1$, se o dado foi registrado ou $R_i = 0$, se o dado é faltante, para o indivíduo i, $i = 1, ..., n$. Assume-se que R é observado completamente. Este indicador é considerado como uma variável aleatória em adição às variáveis observadas Y_{obs}. Assim, a distribuição conjunta dos indicadores de resposta R e dos dados Y tem a forma

$$f(R, Y \mid \theta_R\theta_Y) = f_1(Y \mid \theta_Y)\, f_2(R \mid Y, \theta_R)$$

em que f_1 é a função de densidade de Y com vetor de parâmetros θ_Y e f_2 é a função de probabilidade para o mecanismo de resposta com vetor de parâmetros θ_R. Assim, o mecanismo de resposta para a geração de R como expresso no modelo, com vetor de parâmetros θ_R, pode ser influenciado pelos resultados Y, sejam eles observados ou ausentes. A distribuição condicional de R, dado o conjunto de dados completo Y, indexado pelo vetor de parâmetros desconhecido θ_R, descreve o mecanismo de dados ausentes. Esse mecanismo será considerado como aleatório ao acaso se

$$f_2(R \mid Y, \theta_R) = f(R|Y_{obs}, Y_{aus}, \theta_R) = f(R \mid Y_{obs}, \theta_R).$$

Desse modo, a distribuição do mecanismo de dados ausentes pode depender de outros valores observados (incluindo covariáveis completamente observadas), além dos parâmetros θ_R, mas não depende dos valores ausentes. Segundo Gelman et al (1995), este caso é o mais comum de aparecer na prática e o fato de ter suposições menos severas faz possível a inferência sobre os parâmetros de interesse sem precisar modelar o mecanismo dos dados ausentes.

2.5.1 Imputação múltipla

Esta técnica estatística proposta originalmente por Rubin (1977) e descrita em detalhe em Rubin (1987), está baseada na substituição de cada valor ausente ou deficiente com $m \geq 2$ valores aceitáveis, representando uma distribuição de possibilidades (nossa incerteza sobre qual valor imputar). Assim, com as m imputações para cada dado ausente é possível criar m conjuntos de dados completos e cada um desses conjuntos é analisado, usando-se procedimentos padrões para conjuntos de dados completos, tal como se os dados imputados fossem os dados reais. Num contexto Bayesiano, essas imputações são obtidas via a técnica de predição Bayesiana usual, tratando os dados ausentes como parâmetros extras a serem estimados. Assim, no caso de dados ausentes ao acaso, a densidade preditiva de Y_{aus} é

$$P(y_{aus} \mid y_{obs}) = \int p(y_{aus} \mid y_{obs}, \boldsymbol{q}_Y)\, p(\boldsymbol{q}_Y \mid y_{obs}) d\boldsymbol{q}_Y$$

em que $\mathbf{q}_Y$ é o vetor de parâmetros do modelo usado para estimar os dados ausentes $\mathbf{y}_{aus}$. Assim, as imputações podem ser criadas amostrando-se primeiro da densidade *a posteriori* de θ, para obter $\mathbf{q}^*$ e depois amostrando da distribuição preditiva $P(\mathbf{y}_{aus}|\mathbf{y}_{obs}, \mathbf{q}^*)$. A análise dos *m* conjuntos de dados completos, usando-se métodos padrões fornecem estimativas diferentes de $\boldsymbol{q}(\hat{\boldsymbol{q}}_1, ..., \hat{\boldsymbol{q}}_m)$, com variâncias associadas $U_1, ..., U_m$, respectivamente. O passo seguinte é combinar essas estimativas para obter inferências finais que reflitam apropriadamente a incerteza associada aos dados ausentes, sob o modelo especificado. A estimativa combinada para cada $\boldsymbol{q}$ é dada por

$$\bar{\boldsymbol{q}} = \sum_{i=1}^{m} \frac{\hat{\boldsymbol{q}}_i}{m}.$$

A variabilidade associada a esta estimativa tem dois componentes: a variância dentro das imputações,

$$W = \sum_{i=1}^{m} \frac{U_i}{m}$$

e a variância entre imputações,

$$B = \frac{\sum_{i=1}^{m} (\boldsymbol{q}_i - \bar{\boldsymbol{q}})^2}{m-1}.$$

Assim, a variância total estimada para a estimativa combinada $\bar{\boldsymbol{q}}$ é

$$T = W + \left(\frac{m+1}{m}\right)B$$

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

Quando θ é um escalar, estimativas por intervalo e níveis de significância podem ser obtidos usando-se uma distribuição de referência t-Student com $v = (m - 1)(1 + r^{-1})^2$ graus de liberdade, sendo que r é o incremento relativo na variância devido aos dados ausentes, isto é,

$$r = \left(1 + \frac{1}{m}\right)\frac{B}{\bar{U}}.$$

Assim, uma estimativa por intervalo para θ, com um coeficiente de confiança de $100(1 - \alpha)\%$, é dada por

$$\bar{\theta} \pm t_v(\alpha / 2)T^{1/2}.$$

A fração de informação sobre θ ausente devido aos dados faltantes é dada por

$$\gamma = \frac{r + 2/(v+3)}{r+1}.$$

Extensões desta metodologia e aplicações recentes podem ser encontradas para a área de saúde pública em Zhou et al (2001) e sob um contexto bayesiano em Landrum & Becker (2001) e Wu & Wu (2001).

2.6 Modelos de misturas

Embora muitos fenômenos permitam uma modelagem probabilística através de distribuições clássicas tais como a normal, gama, Poisson, binomial etc, a estrutura probabilística de alguns fenômenos observados pode ser complicada demais para ser acomodada por estas formas simples. Entre as soluções estatísticas para este problema estão escolher uma abordagem não paramétrica ou usar distribuições parametrizadas mais elaboradas. Esta última opção está baseada em densidades padrões $f(x|\boldsymbol{q})$, usadas como uma base funcional para aproximar a densidade verdadeira $g(x)$ da amostra da seguinte forma

$$g(x) \cong \hat{g}(x) = \sum_{i=1}^{k} p_i f(x|\boldsymbol{q}_i) \tag{9}$$

em que $p_1 + ... + p_k = 1$. O lado direito da equação (9) é chamado de distribuição de mixtura finita e as várias distribuições $f(x \mid \boldsymbol{q}_i)$ são os componentes da mistura. É comum assumir que os componentes da mistura são todos da mesma família paramétrica, com diferentes vetores de parâmetros $\boldsymbol{q}_i$.

Um caso particular de distribuições de mistura são os modelos inflacionados de zeros, muito úteis quando, por exemplo, a classe zero de um conjunto de dados é inflacionada pela inclusão de indivíduos pertencendo a um grupo de “não sucetíveis” ou “não infestados”. Cohen (1966) considerou a forma geral deste tipo de distribuição de mistura tendo a expressão

$$P(x; p, \boldsymbol{q}) = \begin{cases} (1-p) + pP_1(0;\boldsymbol{q}) & se \quad x = 0 \\ pP_1(x;\boldsymbol{q}) & se \quad x > 0 \end{cases}$$

em que P1(x;θ) é qualquer distribuição discreta definida sobre o domínio $x > 0$ com parâmetros $\mathbf{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_k)^T$.

Cobrindo aspectos de ajuste e inferência, Ridout et al. (1998) apresentam uma revisão de literatura e discutem uma metodologia geral para modelar dados de contagem inflacionados de zeros, com ênfase em aplicações em agricultura, enquanto que Vieira et al. (2000) apresentam alguns modelos para dados de proporções inflacionados de zeros com aplicações a ensaios de controle biológico. Outras aplicações recentes dos modelos Poisson e binomial inflacionados de zeros podem ser encontrados, por exemplo, em Hall (2000) e sob um contexto bayesiano em Ghosh et al. (2001).

3 MATERIAL E MÉTODOS

3.1 Material

A motivação para este trabalho surgiu a partir de um conjunto real de dados, proveniente de um levantamento da infestação da broca ao longo de quatro anos (1995-1998) numa cultura de café, desde o início do seu ciclo produtivo. Os dados fazem parte de um experimento realizado no *Centro Nacional de Investigaciones de Café* (Cenicafé) na Colômbia, como parte de um projeto que pretende estudar a distribuição do ataque da praga, e sugerir possíveis esquemas de amostragem. A cultura tinha 2214 plantas de café (*Coffea arabica* var. Colômbia), distribuídas numa área de aproximadamente 0,5ha, localizada na estação experimental "La Catalina", no município de Pereira, Colômbia, a 1350 metros acima do nível do mar, com uma temperatura média de 21,6°C, precipitação pluviométrica de 1978 mm/ano, e insolação de 1606 horas/ano. A Figura 1 mostra uma representação esquemática da área experimental. O lote apresentava uma declividade entre 40% e 80%, típica de muitos cafezais da região cafeeira central colombiana, não tendo limites com outras culturas de café. A cultura foi selecionada nove meses após o plantio no campo quando apresentava suas primeiras florações. O espaçamento entre plantas foi de 1,5m x 1,5m. As práticas culturais realizadas na área experimental foram: duas fertilizações ao ano, segundo análises de solos, controle oportuno de plantas daninhas e a colheita permanente de frutos maduros, sobremaduros e secos. Durante toda a duração do experimento foram registradas diariamente as variáveis temperaturas máxima, mínima e média, insolação, umidade relativa e precipitação pluviométrica em uma estação meteorológica localizada cerca de 200 metros da área experimental. A informação começou a ser obtida em julho de 1995, três meses depois do registro da primeira floração importante na cultura. Iniciou-se com

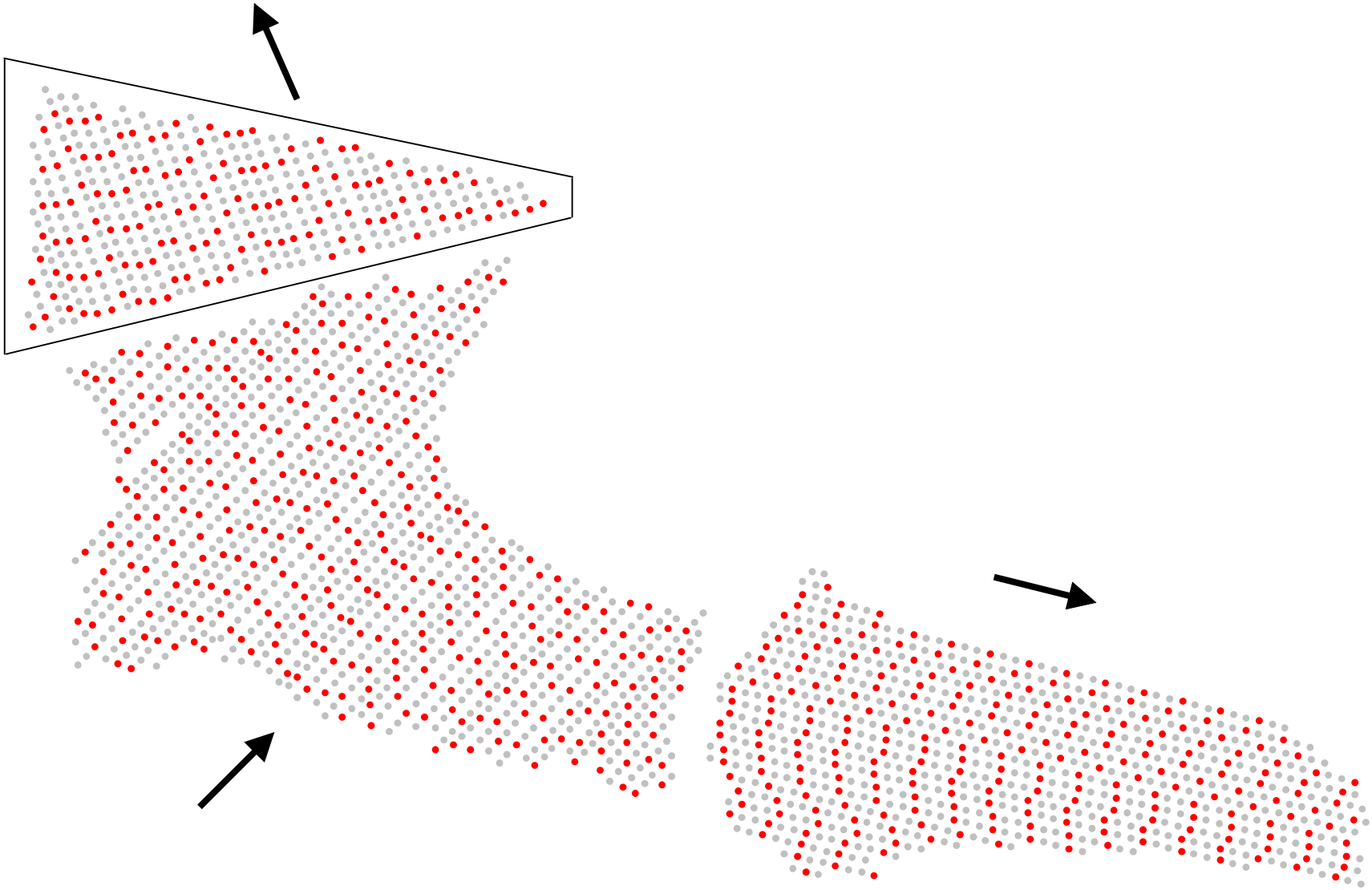

Figura 1 – Representação esquemática da área experimental. A região encerrada num polígono indica a área selecionada para análise. Pontos vermelhos representam as plantas avaliadas entre jul/96 e set/97. Flechas indicam a direção dos sulcos.

uma inspeção mensal em cada planta do lote, observando-se a presença ou ausência da broca. No caso de se encontrar pelo menos um fruto com broca numa planta, procedia-se à contagem de todos os frutos sãos e infestados de cada ramo, em toda a planta; em caso contrário, a planta simplesmente era registrada como não infestada (0% de infestação) e não era realizada a contagem do total de frutos dessa planta. Este primeiro sistema de avaliação foi feito durante os primeiros dez meses. A partir de julho de 1996, devido ao aumento do número de frutos por planta e da proporção de plantas infestadas no lote, as avaliações continuaram sendo feitas mensalmente de forma sistemática, mas somente a cada três plantas, até setembro de 1997 (ver a distribuição das plantas avaliadas na Figura 1); e depois sistematicamente só a cada cinco plantas, até dezembro de 1998. A localização (coordenadas X-Y) das 2214 plantas do lote foi referenciada num plano cartesiano a partir de uma origem arbitrária.

Alguns aspectos a considerar em relação à área experimental, e ao tipo de organismo a se estudar são:

(i) a infestação da broca numa planta não é independente da infestação em outras plantas, devido ao caráter agregado da praga, seja pelas características intrínsecas da espécie, ou em resposta à heterogeneidade ambiental, causada por diferenças na topografia (declive), um possível gradiente de fertilidade, diferenças de microclima etc;

(ii) a dinâmica de produção de frutos nas plantas para a região em estudo faz com que, durante o ano todo, haja frutos em diferentes estágios de desenvolvimento. É possível, então, para todo mês, encontrar plantas com frutos maduros, infestados pela broca, ou não, que serão colhidos durante esse mês, diminuindo ou aumentando o nível de infestação, e frutos que começam sua formação e ainda não são atraentes para o ataque da praga, embora a proporção destes tipos de frutos em relação ao total de frutos produzidos seja variável ao longo do ano, dependendo de certas condições ambientais tais como a distribuição de chuvas, que regula o padrão de florações na cultura;

(iii) ligado ao item anterior está o grau de preferência da broca por certo tipo de frutos; esta preferência vai aumentando com a idade do fruto, à medida que este se torna mais consistente (frutos com mais de 120 dias são mais susceptíveis de serem atacados);

(iv) o número de frutos produzidos varia muito de uma planta para outra como resultado de fatores intrínsecos à planta de café (qualidade das sementes), ou extrínsecos (desuniformidade na realização de práticas culturais, gradientes de fertilidade do solo, diferenças de microclima no interior da parcela etc.). Isso pode dar lugar ao aparecimento de valores extremos na proporção de frutos com broca por planta, principalmente no caso de plantas com poucos frutos;

(v) durante todo o período de observação, mas principalmente nos primeiros sete meses, foi registrada uma grande quantidade de plantas com frutos não infestados pela broca, adicionando consequentemente uma grande quantidade de valores iguais a zero aos registros de proporção de frutos com broca por planta.

Os dados disponíveis podem ser divididos em dois subconjuntos com características diferentes e que, portanto, deverão ter estratégias diferentes de análise. Um primeiro subconjunto de dados (primeiros dez meses) corresponde à fase inicial da infestação, na qual a praga se dispersa a partir de focos iniciais de infestação até colonizar quase a totalidade da área experimental. Esse foi o período em que se registrou o nível de infestação em todas as plantas. Já no segundo subconjunto de dados, a praga está presente na maioria das plantas e é de interesse estudar sua dinâmica através do tempo e em função de covariáveis associadas.

Devido à extensão e à complexidade do tema em estudo e às limitações computacionais e de tempo, esta dissertação considera unicamente a análise do primeiro subconjunto de dados (período inicial), selecionando uma subárea de 392 plantas dentre as 2214 disponíveis (ver área marcada na Figura 1). Essa subárea, porém, é representativa do que aconteceu na parcela completa nesse período.

3.2 Métodos

Uma vez definida a informação a ser utilizada, o passo seguinte foi especificar um modelo básico para os dados. A especificação deste modelo iniciou-se com a identificação da informação disponível. Assim, sejam n_{it} e y_{it}, respectivamente, o número total observado de frutos e o número observado de frutos infestados pela broca na planta i, $i = 1, ..., 392$ no tempo t, $t = 1, ..., 10$. Devido ao fato de n ser finito e algumas vezes estar próximo de zero e existirem taxas de infestação por planta variando entre 0% e 100%, foi assumido que o número de frutos infestados Y_{it} seguia uma distribuição binomial com parâmetros n_{it} e π_{it}, cuja função de verossimilhança é dada por:

$$L(\boldsymbol{\theta} \mid y_{[1,1]},...,y_{[392,10]}) = \prod_{i=1}^{392}\prod_{t=1}^{10}\binom{n_{it}}{y_{it}}\pi_{it}^{\,y_{it}}(1-\pi_{it})^{n_{it}-y_{it}}.$$

A probabilidade de risco desconhecida, π_{it}, foi modelada num contexto de modelos lineares generalizados mistos com uma função de ligação logística, que é a ligação canônica para a distribuição binomial, e um preditor linear η_{it} que se decompõe aditivamente em efeitos fixos e aleatórios segundo o modelo específico que esteja sendo usado. A abordagem Bayesiana foi seguida para a estimação dos parâmetros de todos os modelos considerados nesta dissertação.

3.2.1 Estimação dos dados faltantes

Como foi descrito anteriormente, nas plantas não infestadas (que no início da infestação são a maioria), não foi feita a contagem do número total de frutos, mas estes valores são necessários para implementar o modelo baseado na distribuição binomial, precisando, portanto, serem estimados. O método escolhido para a estimação desses dados faltantes foi o de imputação múltipla (Rubin, 1987), já decrito na revisão de literatura. A escolha deste método foi devido à sua simplicidade de implementação e ao fato de permitir que a estimação dos dados ausentes seja feita separadamente, sendo

possível, depois, o uso de métodos padrão para análise de conjuntos de dados completos na estimação dos parâmetros de interesse do modelo que esteja sendo considerado, o que simplifica bastante o processo.

Informação sobre o número total de frutos das plantas que foram observadas no campo permitiu identificar uma tendência crescente para esta variável ao longo do tempo durante os primeiros 10 meses, conforme mostra a Figura 2.

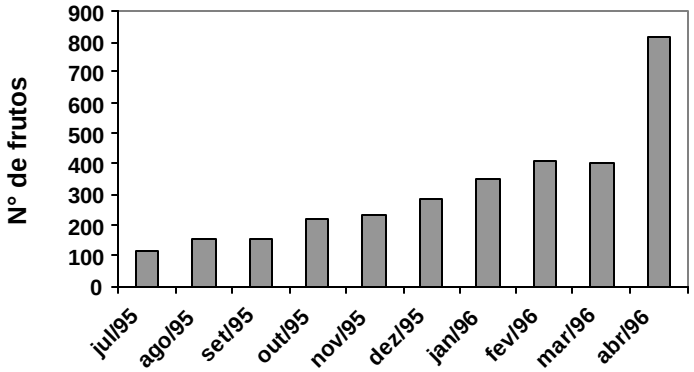

Figura 2 – Média do número total de frutos por planta no período jul/95 - abr/96 baseada nas contagens do total de frutos das plantas infestadas pela broca nesse período.

Portanto, para a modelagem dessas contagens, foi adotado um modelo de crescimento da forma

$$N_{it} \sim \text{Poisson}(\mu_{it}) \text{ com } \log(\mu_{it}) = \alpha_i + \beta_{it} \, , \; i = 1, \ldots, 392; \;\; t = 1, \ldots, 10 \qquad (10)$$

sendo que α_i representa a média geral do logaritmo do número total de frutos da planta i e β_i é o parâmetro relacionado com o tempo e que determina se o número total de frutos da planta i, em cada tempo t, está aumentando ou diminuindo.

Foram atribuídas distribuições para **a** e **b** que representam o conhecimento *a priori* a respeito desses parâmetros. Para se ter uma idéia sobre a existência, ou não, de correlação espacial entre as contagens do número de total de frutos por planta, foi

calculado um índice de autocorrelação de Moran (Moran, 1948) para esta variável nos dois últimos meses (março e abril de 1996), pois nesses meses existia informação do total de frutos para 58% e 70% das plantas respectivamente. Os resultados não indicaram evidência de correlação espacial significativa para essa variável e, portanto, não foi considerada dependência espacial para nenhum dos parâmetros do modelo (10). Foi, então, admitido que *a priori*

$$\alpha \sim \text{normal}(\lambda_\alpha, \tau_\alpha) \quad \text{e} \quad \beta \sim \text{normal}(\lambda_\beta, \tau_\beta)$$

para $\lambda_\alpha = 4{,}6$; $\tau_\alpha = 1{,}6$; $\lambda_\beta = 0{,}1$ e $\tau_\beta = 83$. Os parâmetros τ_α e τ_β correspondem às precisões (inverso das variâncias) de cada distribuição.

Os parâmetros destas *prioris* representam valores plausíveis que dariam uma estimativa razoável do número total de frutos e foram obtidos com base no conhecimento prévio do número de frutos que seria esperado para plantas dessa idade. Assim, por exemplo, para o intercepto α, foi assumido que um valor plausível para representar a média geral do número de frutos nesse período seria igual a 100 frutos (isto é, com ln(100) = 4,6), mas que este valor poderia flutuar entre um mínimo de cinco frutos e um máximo de 2500 frutos; desse modo, em escala logarítmica, a amplitude entre a média e o valor máximo, 7,8 - 4,6 = 3,2, seria aproximadamente igual a quatro vezes o desvio padrão (3,2 = 4σ), o que dá uma precisão $\tau = 1/\sigma^2 = 1{,}6$. Um raciocínio similar fornece os valores para os parâmetros da *priori* normal para β.

O conhecimento *a priori* sobre os parâmetros do modelo foi atualizado pela amostra n_{it} em cada área i e tempo t, através do modelo (10) para gerar um conhecimento *a posteriori* sobre esses parâmetros incluindo os dados faltantes que num contexto Bayesiano são tratados como parâmetros extras e, portanto, estimados junto com α e β. Esse conhecimento é expresso pela distribuição de probabilidade *a posteriori* conjunta

$$P(\boldsymbol{a},\boldsymbol{b},n_{\text{ausentes}} \mid n_{\text{observados}}) \propto \prod_{i=1}^{392}\prod_{t=1}^{10}\left\{\frac{(\hat{\imath}_{it})^{n_{it}}\exp(-\grave{\imath}_{it})}{n_{it}!}\right\} \exp\left\{-\frac{\hat{o}_{\acute{a}}}{2}\sum_{i=1}^{392}(\acute{a}_i-\ddot{e}_{\acute{a}})^2\right\}$$

$$\exp\left\{-\frac{\hat{o}_{\hat{a}}}{2}\sum_{i=1}^{392}(\hat{a}_i-\ddot{e}_{\hat{a}})^2\right\}.$$

O ajuste do modelo foi feito via métodos de simulação MCMC, implementados no *software* Winbugs versão 1.3 (Spiegelhalter et al., 2000). Foi gerada uma única cadeia do amostrador Gibbs, com um ciclo de pré-convergência (*burn-in*) de 5000 iterações, seguidas de 25000 iterações, das quais foram guardadas somente 5000 (uma a cada 5) para o cálculo das estatísticas *a posteriori* de interesse e para testar a convergência das simulações a qual foi verificada seguindo os critérios de Geweke (1992), Heidelberger & Welch (1983) e Raftery & Lewis (1992), usando o programa CODA versão 0.3 (Best et al., 1996).

Seguindo a metodologia de imputação múltipla e considerando que era necessário estimar mais de 50% dos *n*'s, o modelo (10) foi implementado dez vezes, usando diferentes conjuntos de valores iniciais para assim formar $m = 10$ conjuntos de valores imputados que representam uma distribuição de valores plausíveis do número total de frutos em cada planta, em cada tempo. Os valores imputados sempre corresponderam ao valor da última iteração do amostrador Gibbs e não da média *a posteriori* para permitir variabilidade de amostragem que reflete nossa incerteza sobre os valores ausentes a serem estimados.

3.2.2 Análise espacial

Após a estimação dos *n*'s ausentes, deu-se início ao estudo da variação da infestação da broca, considerando-se em principio somente o componente espacial. Para tal efeito foi escolhido um dos 10 meses (março/96) para fazer uma análise de sensibilidade das estimativas do risco de infestação à escolha de diferentes distribuições *a priori* para os parâmetros e os hiperparâmetros e à escolha de diferentes esquemas de vizinhança para o modelo

$$Y_i \mid \boldsymbol{p}_i \sim \text{Binomial}(n_i, \boldsymbol{p}_i) \text{ com } \text{logit}(\boldsymbol{p}_i) = \boldsymbol{h}_i = \boldsymbol{a} + \boldsymbol{q}_i + \boldsymbol{f}_i, \; i = 1, \ldots, 392 \qquad (11)$$

em que n_i e y_i representam, respectivamente, as contagens do número total de frutos e do número de frutos infestados pela broca na planta i; $\boldsymbol{a}$ representa o logaritmo do nível geral do risco relativo de infestação na área experimental. O modelo considera efeitos aleatórios para cada área, constituídos pela soma de um componente espacialmente estruturado, $\boldsymbol{q}_i$ e um componente de heterogeneidade não estruturada, $\boldsymbol{f}_i$ como definidos na equação (2). Estes dois efeitos representam características não especificadas da planta i que visualizam, ou não, estrutura espacial, respectivamente, e podem ser interpretados como substitutos de covariáveis não medidas. Foram avaliadas igualmente diferentes combinações desse modelo, incluindo só efeitos aleatórios não estruturados, estruturados espacialmente e a combinação de ambos os efeitos. Os modelos que foram avaliados estão resumidos na Tabela 1.

Para estudar a influência da escolha das hiper*prioris* para $\boldsymbol{t}_q$ e $\boldsymbol{t}_f$ sobre as estimativas do risco de infestação, foram escolhidas quatro distribuições gama com diferentes médias e variâncias, representando diferentes graus de conhecimento *a priori* sobre o valor dos parâmetros $\boldsymbol{t}_q$ e $\boldsymbol{t}_f$ tal como ilustra a Tabela 2, porém, sem chegar a serem muito informativas, conforme pode ser visto na Figura 3, refletindo a incerteza sobre esses parâmetros.

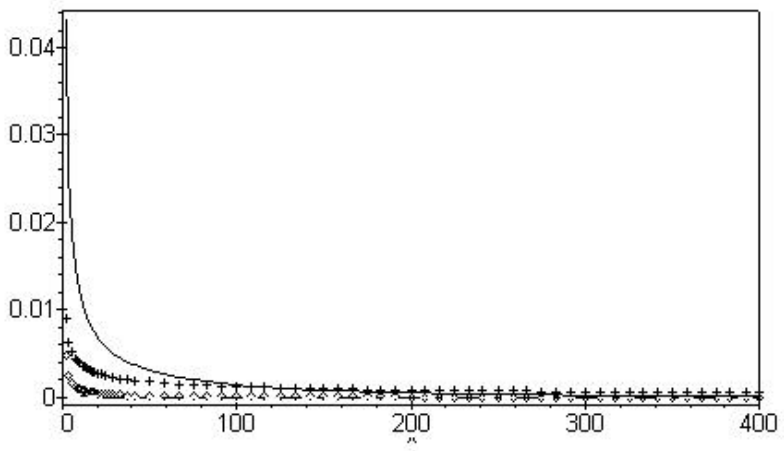

Figura 3 – Distribuições *a priori* gama para os parâmetros $\boldsymbol{t}_q$ e $\boldsymbol{t}_f$.

Tabela 1. Modelos, distribuições *a priori* e esquemas de vizinhança avaliados na análise espacial.

	Modelo	Esquema de vizinhança	Distribuições *a priori*			
			ϕ	θ	τ_θ	τ_ϕ
(1)	$\eta_i = \alpha + \theta_i + \phi_i$	2ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(2)	$\eta_i = \alpha + \theta_i + \phi_i$	2ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,5; 0,0005)	Gama(0,5; 0,0005)
(3)	$\eta_i = \alpha + \theta_i$	2ª ordem	-	CAR(τ_θ)	Gama(0,001; 0,001)	-
(4)	$\eta_i = \phi_i$	2ª ordem	Normal(0, τ_ϕ)	-	-	Gama(0,001; 0,001)
(5)	$\eta_i = \alpha + \theta_i + \phi_i$	2ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,25; 0,005)	Gama(0,25; 0,005)
(6)	$\eta_i = \alpha + \theta_i + \phi_i$	2ª ordem	t(0, τ_ϕ, 3)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(7)	$\eta_i = \alpha + \theta_i + \phi_i$	2ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,01; 0,001)	Gama(0,25; 0,005)
(8)	$\eta_i = \alpha + \theta_i + \phi_i$	1ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(9)	$\eta_i = \alpha + \theta_i + \phi_i$	4ª ordem	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(10)	$\eta_i = \alpha + \theta_i + \phi_i$	3 metros	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(11)	$\eta_i = \alpha + \theta_i + \phi_i$	5 metros	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(12)	$\eta_i = \alpha + \theta_i + \phi_i$	7 metros	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)
(13)	$\eta_i = \alpha + \theta_i + \phi_i$	10 metros	Normal(0, τ_ϕ)	CAR(τ_θ)	Gama(0,001; 0,001)	Gama(0,001; 0,001)

A distribuição *a priori* para α em todos os modelos foi $\alpha \sim$ Uniforme($-\infty$, $+\infty$).
CAR(τ_θ) corresponde a uma auto-regressão condicional intrínseca Gaussiana como definida na Equação (4), com parâmetro de precisão τ_θ.

Tabela 2. Médias e variâncias *a priori* e parâmetros para as quatro combinações de hiper*prioris* gama(*a*, *b*) consideradas para os hiperparâmetros do modelo 11.

Hiper*priori*	Componente não estruturado (τ_ϕ)			Componente com estrutura espacial (τ_θ)			Variância*	Razão de médias a *priori***
	média	*a*	*b*	média	*a*	*B*		
1	1	0,001	0,001	1	0,001	0,001	1000	1
2	50	0,25	0,005	50	0,25	0,005	10000	1
3	1000	0,5	0,0005	1000	0,5	0,0005	2,0E+6	1
4	10	0,01	0,001	50	0,25	0,005	10000	5

* Variância comum às *prioris* dos dois efeitos aleatórios.
** Razão entre as médias *a priori* para t_q e t_f

Dado que as plantas encontravam-se regularmente espaçadas na área experimental, foram inicialmente considerados esquemas de vizinhança definidos para látices regulares na representação da estrutura de correlação espacial dos dados. Foram avaliados os sistemas de vizinhança de primeira, segunda e quarta ordem, como descritos em Besag (1974) e ilustrados na Figura 4. No primeiro caso, a vizinhança para o sítio *i* é definida pelos pares de plantas horizontal e verticalmente adjacentes à planta *i*. O sistema de segunda ordem, em adição aos quatro vizinhos mais próximos do esquema de primeira ordem, inclui os vizinhos lateralmente adjacentes e assim por diante. Foram avaliados também esquemas de vizinhança baseados em distância, definidos por raios de 3, 5, 7 e 10 metros. Estes sistemas de vizinhança foram usados na especificação das auto-regressões condicionais Gaussianas descritas na equação (4) e que constituem as distribuições *a priori* para os efeitos aleatórios espacialmente estruturados $\boldsymbol{q}$ do modelo (11). Assim, a distribuição *a posteriori* conjunta para o modelo (11) tem a forma

$$P(\grave{\mathbf{e}}, \boldsymbol{f}, t_q, t_f \mid \mathbf{y}) \propto \prod_{i=1}^{392} \left\{ \binom{n_i}{y_i} \boldsymbol{p}_i^{\,y_i} (1-\boldsymbol{p}_i)^{n_i - y_i} \right\} t_f^{-392/2} \exp\left\{ -\frac{t_f}{2} \sum_{i=1}^{392} (\boldsymbol{f}_i - \bar{\boldsymbol{f}})^2 \right\}$$

$$t_q^{-392/2} \exp\left\{ -\frac{t_q}{2} \sum_{i \sim j} (\boldsymbol{q}_i - \boldsymbol{q}_j)^2 \right\}$$

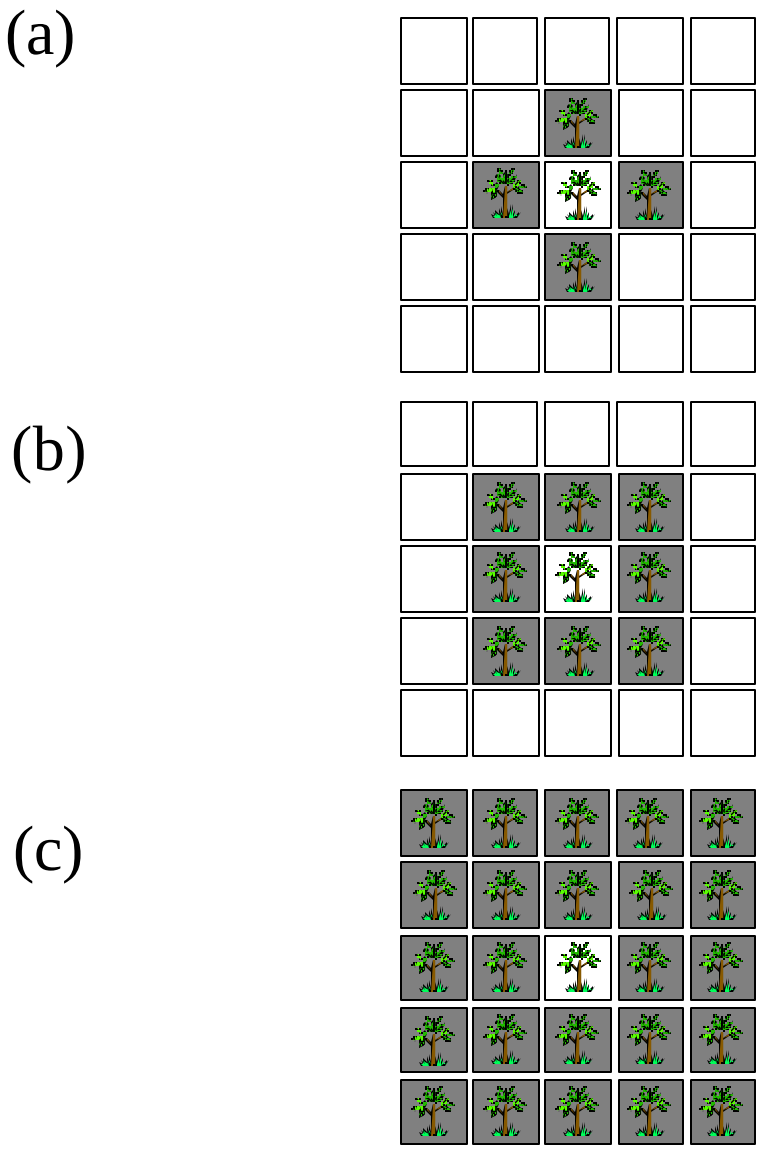

Figura 4 – Representação esquemática de sistemas de vizinhança sobre um látice regular: (a) de primeira ordem, (b) de segunda ordem, (c) de quarta ordem.

em que i~j indica que as plantas *i* e *j* são vizinhas. Cada um dos modelos propostos foi implementado dez vezes, usando-se os dez conjuntos de valores imputados para os *n*'s ausentes obtidos previamente. O ajuste foi feito de forma similar aos modelos para a estimação dos *n*'s, gerando uma única cadeia do amostrador Gibbs, descartando as primeiras 5000 iterações, seguidas de 25000 iterações, das quais somente foram guardadas 5000 (uma a cada 5) para o cálculo das estatísticas *a posteriori* de interesse. As estimativas *a posteriori* combinadas para os parâmetros de interesse foram obtidas usando a média aritmética das dez repetições de cada modelo. A convergência das cadeias foi testada, usando-se os mesmos critérios do modelo (10). Os diferentes modelos foram comparados com base no critério de Gelfand & Ghosh (1998) como foi descrito na revisão de literatura.

Posteriormente, foram ajustados modelos separadamente para cada um dos 10 tempos, seguindo o modelo 3 da Tabela 1, para se ter uma idéia de como era a variação dos parâmetros através do tempo, o que sugeriria possíveis estratégias de modelagem espaço-tempo. O ajuste destes modelos foi feito de forma semelhante à usada para os modelos avaliados na análise de sensibilidade.

3.2.3 Análise espaço-temporal

Durante os sete primeiros meses de avaliações (julho/95 - janeiro/96), a infestação por planta manteve-se em níveis baixos, e com uma grande quantidade de plantas ainda não infestadas como mostram os dados originais apresentados em forma de mapas nas Figuras 5 e 6, não sendo, portanto, de muito interesse para estudar a colonização ou a disseminação da praga na parcela. Isso somado ao fato das limitações computacionais para ajustar modelos com conjuntos grandes de dados em um ambiente *Windows*, levou a limitar a análise espaço-temporal aos últimos quatro meses (de janeiro/96 a abril/96) que constituíam o período mais interessante para estudar a dispersão da praga.

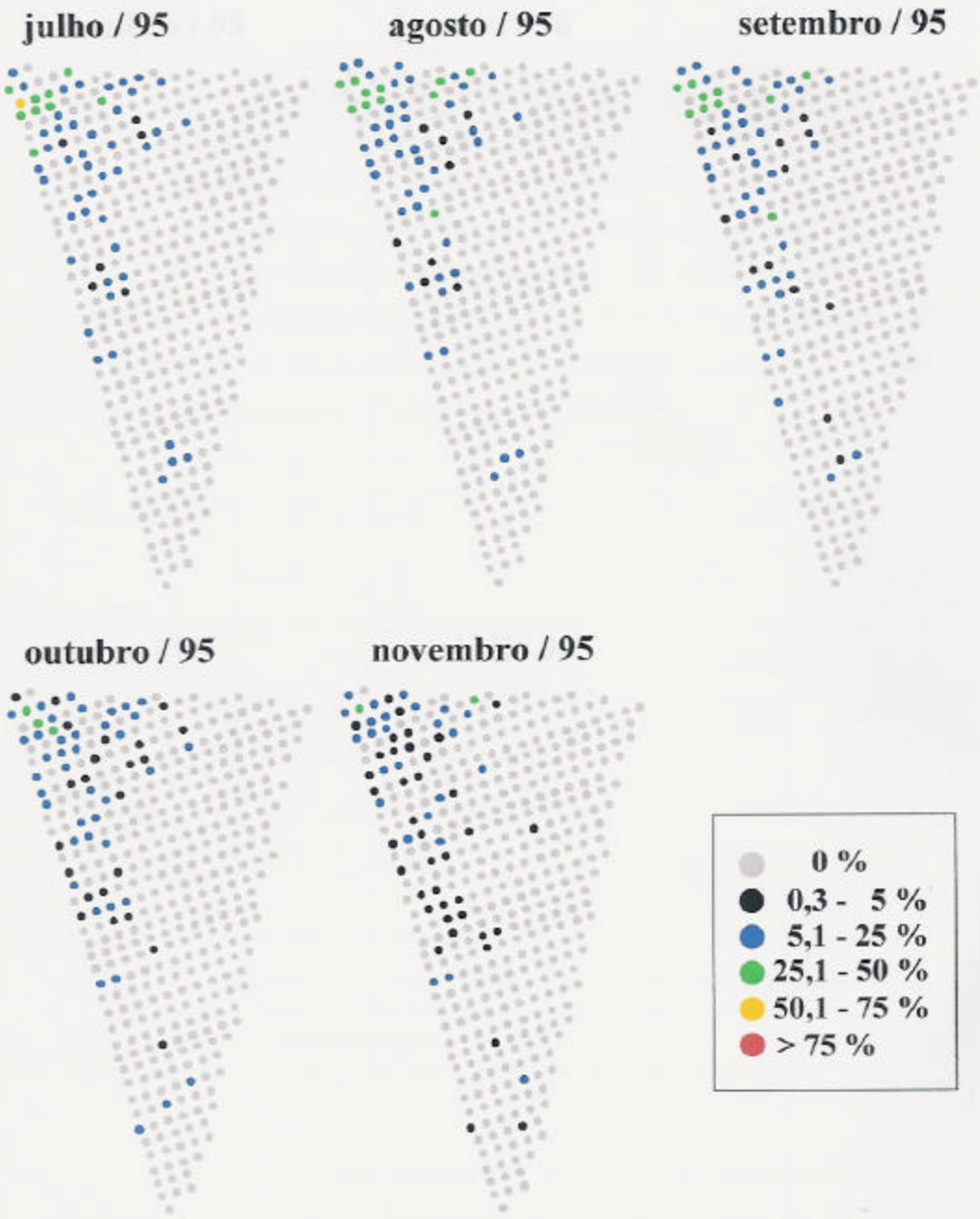

Figura 5 – Infestação da broca no período julho/95 – novembro/95

Figura 6 – Infestação da broca no período dezembro95 – abril96

Assim, foi assumido que $Y_{it} \sim$ binomial (n_{it} , π_{it}), com $i = 1, ..., 392$ e $t = 1, ..., 4$ e que logit (π_{it}) = η_{it}. Para a modelagem deste preditor linear foram considerados polinômios de primeiro e segundo grau que modelam o aumento ou decréscimo nas taxas de infestação através do tempo, segundo uma tendência linear ou quadrática. Os modelos e distribuições *a priori* avaliados estão resumidos na Tabela 3. Note-se que cada efeito aleatório com dependência espacial nos diferentes modelos tem associado um efeito fixo com uma distribuição *a priori* "flat"; esta parametrização é necessária devido ao fato de a *priori* CAR ser imprópria (ver revisão de literatura para detalhes). O esquema de vizinhança adotado para as *prioris* dos efeitos aleatórios com dependência espacial foi o de segunda ordem como ilustrado na Figura 4.

Tabela 3. Modelos e distribuições *a priori* avaliados na análise espaço-temporal.

Modelo*		Distribuições *a priori***		
		δ	γ	ν
(1)	$\eta_{it} = \delta_i + \gamma_i t$	Normal(0, τ_δ)	Normal(0, τ_γ)	-
(2)	$\eta_{it} = \delta_i + (\xi + \gamma_i)t$	Normal(0, τ_δ)	CAR(τ_γ)	-
(3)	$\eta_{it} = (\omega + \delta_i) + \gamma_i t$	CAR(τ_δ)**	Normal(0, τ_γ)	-
(4)	$\eta_{it} = (\omega + \delta_i) + (\xi + \gamma_i)t$	CAR(τ_δ)	CAR(τ_γ)	-
(5)	$\eta_{it} = \delta_i + \gamma_i t + \nu_i t^2$	Normal(0, τ_δ)	Normal(0, τ_γ)	Normal(0, τ_ν)
(6)	$\eta_{it} = (\omega + \delta_i) + \gamma_i t + \nu_i t^2$	CAR(τ_δ)	Normal(0, τ_γ)	Normal(0, τ_ν)
(7)	$\eta_{it} = \delta_i + (\xi + \gamma_i)t + (\varepsilon + \nu_i)t^2$	Normal(0, τ_δ)	CAR(τ_γ)	CAR(τ_ν)
(8)	$\eta_{it} = (\omega + \delta_i) + (\xi + \gamma_i)t + (\varepsilon + \nu_i)t^2$	CAR(τ_δ)	CAR(τ_γ)	CAR(τ_ν)

* A distribuição *a priori* para x , v e e em todos os modelos foi uma uniforme ($-\infty$, $+\infty$).

** CAR(τ) corresponde a uma auto-regressão condicional intrínseca Gaussiana como definida na equação (4), com parâmetro de precisão τ.

*** τ_δ, τ_γ e τ_ν têm distribuições gama (0,001; 0,001) em todos os modelos.

O ajuste e a comparação dos modelos foi feita de forma semelhante à usada para os modelos da seção anterior, descartando as primeiras 5000 iterações do amostrador Gibbs e guardando uma a cada 5 das 20000 iterações restantes.

3.2.4 Modelos de mistura

O risco de infestação da broca também foi modelado, usando-se um modelo inflacionado de zeros baseado na distribuição binomial, o qual constitui um modelo de mistura com dois componentes. Isto foi feito considerando inicialmente só o espaço e depois em espaço e tempo. Para o primeiro caso, seja Y a variável aleatória 'número de frutos com broca em n frutos', com observações (y_i , n_i), i = 1, 2,..., 392 e com y_i / n_i representando a proporção de frutos com broca. Considere-se igualmente a variável indicadora binária Z = {Z_i}, assumindo os valores Z_i = 1 se a planta i não esta infestada pela broca (y_i = 0) ou Z_i = 0 se a planta i tem algum grau de infestação (y_i = 1, 2, ..., n_i). Embora teoricamente todas as plantas de café com frutos numa área não infestada estejam em condições de serem atacadas pelo inseto, na prática, fatores tais como o caráter agregado da praga, pequenas diferenças em microclima no interior da cultura ou mesmo gradientes de fertilidade, plantas nas bordas da cultura limitando com culturas já infestadas, etc., fazem com que algumas plantas sejam mais atraentes para a broca no início da infestação do que outras. Esta situação, porém, vai ficando menos evidente à medida que a praga começa a se reproduzir e a colonizar a cultura. Assim, é razoável considerar um modelo de mistura para modelar o início da infestação no qual uma proporção p das plantas permanece não infestada enquanto que a proporção restante 1 – p tem algum grau de infestação, sendo que o número de frutos com broca nessas plantas segue uma distribuição binomial com parâmetros n_i e δ_i. Assim, assumindo que Z_i ~ Bernoulli(p), Y tem uma distribuição binomial inflacionada de zeros, dada por:

$$\Pr(Y_i = y_i) = \begin{cases} p + (1-p)(1-\delta_i)^{n_i} & y = 0 \\ (1-p)\binom{n_i}{y_i}\delta_i^{\,y_i}(1-\delta_i)^{n_i - y_i} & y = 1, 2, \ldots, n_i \end{cases}$$

com $0 \leq p < 1$. Note-se que quando $p = 0$, este modelo é equivalente à distribuição binomial padrão. A função de verossimilhança para este modelo de mistura é dada pela expressão

$$L(p, \boldsymbol{\delta} \mid \mathbf{Y}) = \prod_{i=1}^{N} \left[p + (1-p)(1-\pi_i)^{n_i} \right]^{Z_i} \left[(1-p) \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \right]^{1-Z_i}.$$

De forma semelhante aos modelos anteriores, foi usada uma função de ligação logística para a binomial, com logit(π_i) = δ_i . O parâmetro do preditor linear δ_i foi modelado, inicialmente, sem considerar dependência espacial *a priori.* Assim, foi assumido que $\delta_i \sim$ normal (μ, τ) com *hiperprioris* de parâmetros conhecidos $\mu \sim$ normal (0; 1,0E-6) e $\tau \sim$ gama (0,001; 0,001), porém, sendo pouco informativas. Um segundo caso foi a modelagem do preditor linear considerando dependência espacial, isto é, assumindo que plantas infestadas próximas entre si tendem a ter riscos de infestação semelhantes que variam suavemente na vizinhança de cada planta. Nesse caso, logit (π_i) = $\alpha + \gamma_i$ em que $\alpha \sim$ uniforme(-∞, +∞) e $\gamma_i \sim$ CAR(τ), com $\tau \sim$ gama(0,001; 0,001).

Para o parâmetro *p* foi atribuída uma distribuição *a priori* independente beta (*a*,*b*), com hiperparâmetros *a* e *b* conhecidos. O efeito da escolha desta *priori* sobre a classificação das observações em cada um dos componentes da mistura, foi avaliado usando-se três *prioris* beta diferentes; duas delas sendo pouco informativas e uma altamente informativa (ver Tabela 4).

Tabela 4. Médias e variâncias *a priori* e parâmetros das *prioris* beta avaliadas para *p*

Priori	Parâmetros da distribuição beta		Média	Variância
	a	*b*	*a priori*	*a priori*
1	1	1	0,5	0,08
2	0,5	0,5	0,5	0,12
3	30	5	0,85	0,0034

Assim, a *priori* beta (1,1) que é equivalente a uma uniforme (0,1), atribui a *priori* probabilidades iguais de uma planta pertencer a qualquer uma das duas categorias (não infestada ou com algum grau de infestação), representando um conhecimento vago sobre o parâmetro *p*. A diferença em relação à *priori* beta (0,5; 0,5), que também tem uma média igual a 0,5, é que esta última atribui um peso maior a valores próximos de zero ou um como pode ser visto na Figura 7 e assim, poderia favorecer a predominância de um certo componente da mistura. Os parâmetros da *priori* informativa beta (30, 5) foram selecionados para favorecer um valor de *p* alto, já que é sabido a *priori* que no início da infestação há uma grande quantidade de plantas ainda não infestadas. Assim, foi assumido que em média 85% das plantas no início da infestação podiam não estar infestadas (p = 0,85), mas que esta proporção média podia variar num intervalo entre 0,70 e 1,0. Desse modo, 0,15 (a diferença entre a média e qualquer dos extremos do intervalo) pode ser considerado como aproximadamente equivalente a dois desvios padrões ao redor da média, dando uma variância para o parâmetro *p* de 0,0034. A função de densidade de uma distribuição beta com parâmetros *a* e *b* tem média e variância dadas pelas expressões $m = a / (a + b)$ e $s^2 = m(1 - m) / (a + b + 1)$. Assim, os parâmetros *a* e *b* podem ser calculados a partir das expressões: $a = m[m(1 - m) / s^2 - 1]$ e $b = a[1 - m] / m$, dando os valores da *priori* informativa $a = 30$ e $b = 5$.

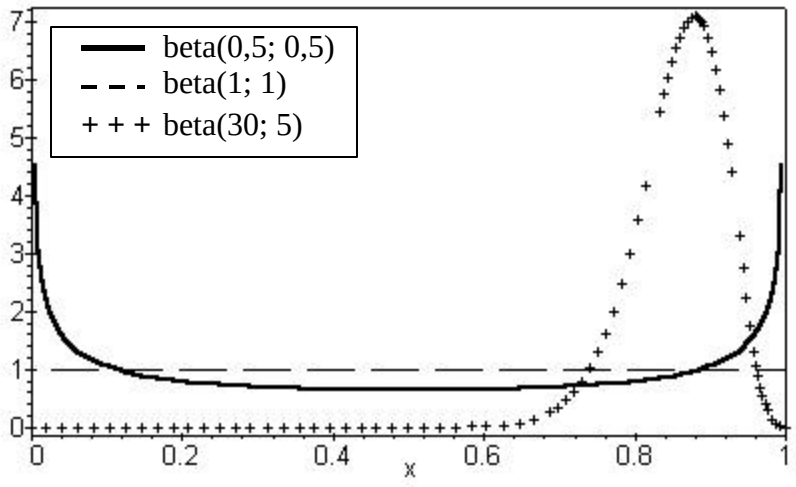

Figura 7 – Distribuições *a priori* para o parâmetro *p*.

O segundo modelo de mistura considerado assume que $Y_i \sim$ binomial(n_i, π_i), em que cada Y_i pode pertencer a um de dois grupos, T_i, $i = 1, 2$. Considera-se que uma fração desconhecida p das observações pertence ao grupo $T_i = 2$, enquanto que $1 - p$ das observações estão no grupo $T_i = 1$. Assumindo uma função de ligação logística para a binomial tem-se que logit (π_i) = δ_i em que o preditor linear δ_i é modelado como uma mistura de duas distribuições normais. Assim, $\delta_i \sim$ normal (μ_{Ti}, τ_i) e ($T_i - 1$) ~ Bernoulli(p), $i = 1,2$.

Quando se usa este modelo, existe a possibilidade que em alguma iteração todos os dados irão para um só dos componentes da mistura. Para superar este problema, Robert (1996) sugere a re-parametrização $\mu_2 = \mu_1 + \lambda$, com $\lambda > 0$.

Prioris independentes pouco informativas são dadas para os parâmetros deste modelo assim: $\mu_1 \sim$ normal (0; 1,0E-6), $\lambda \sim$ normal (0; 1,0E-6), $\tau_i \sim$ gama (0,001; 0,001). Para o parâmetro p foram consideradas as mesmas distribuições *a priori* beta (a,b) avaliadas no modelo de mistura anterior.

Para o ajuste de cada um destes modelos de mistura foi gerada uma cadeia de 45000 iterações, das quais foram descartadas as primeiras 5000 e guardadas uma a cada 20 das 40000 restantes para formar uma amostra final de 2000 iterações, usadas para o cálculo das estatísticas *a posteriori* de interesse. A convergência das simulações foi testada usando os mesmos critérios dos modelos anteriores. De forma semelhante aos casos anteriores, cada modelo foi implementado 10 vezes para obter estimativas combinadas dos parâmetros de interesse baseadas nos 10 conjuntos de valores imputados gerados previamente. Estes modelos foram implementados separadamente para os meses de janeiro, fevereiro, março e abril de 1996.

Posteriormente, o primeiro modelo de mistura foi estendido para modelar a infestação da broca em espaço e tempo para o período janeiro/96 – abril/96, assumindo que o parâmetro p varia no tempo, ou seja, $p_t \sim$ beta (a_t, b_t), $t = 1, 2, 3, 4$, com hiper-parâmetros a_t e b_t conhecidos e iguais a um. Dado que, $Y_{it} \sim$ binomial (n_{it}, π_{it}) para $y = 1, 2, ..., n_{it}$, a probabilidade π_{it} foi modelada como: logit (π_{it}) = $\delta + \gamma$ t , com $\delta \sim$ normal (μ_δ, τ_δ), $\gamma \sim$ normal (μ_γ, τ_γ) e *hiperprioris* de parâmetros conhecidos $\mu_\delta \sim$ normal (0;

1,0E-6) t_d ~ gama (0,001; 0,001), m_g ~ normal (0; 1,0E-6) e t_g ~ gama (0,001; 0,001). Para o ajuste desse modelo foi gerada uma cadeia de 20000 iterações do amostrador Gibbs, descartando as primeiras 5000 e guardando uma a cada 15 das 15000 restantes para formar uma amostra final de 1000 iterações. A convergência das simulações foi testada seguindo os mesmos critérios dos modelos anteriores.

4 RESULTADOS E DISCUSSÃO

Todos os modelos propostos na metodologia, a exceção do segundo modelo de mistura descrito na seção 3.2.4, foram satisfatoriamente ajustados usando o software WinBUGS e o número de iterações considerado foi suficiente em todos os casos para atingir a convergência das cadeias do amostrador Gibbs. A partir dos modelos ajustados, foram obtidos mapas das médias *a posteriori* dos riscos de infestação da broca. Dado que, nos dados originais a menor taxa de infestação para o período julho de 1995 até abril de 1996 foi de 0,3%, foi assumido nos mapas das médias *a posteriori* dos riscos de infestação para os modelos avaliados que, taxas abaixo desse valor eram equivalentes a 0% de infestação no mapa das taxas brutas.

4.1 Análise espacial

A partir do ajuste dos modelos foi possível obter estimativas *a posteriori* das variâncias marginais empíricas para os efeitos aleatórios não estruturados, ϕ e estruturados espacialmente, θ, as quais são apresentadas na Tabela 5 junto com a razão das médias *a posteriori* dessas variâncias (s_θ^2/s_ϕ^2) para os dois tipos de efeitos. Isto foi feito para os 13 modelos descritos na Tabela 1, avaliados em março de 1996. A razão entre as médias *a posteriori* dessas duas variâncias foi maior do que um para todos os modelos, indicando um domínio da variabilidade espacialmente estruturada nos modelos. Em particular, esta dominância foi mais forte à medida que a ordem do esquema de vizinhança se incrementava, passando de 1,11 para o modelo com vizinhança de primeira ordem, até 3,48 no modelo com vizinhança de quarta ordem. Isto foi igualmente válido para os modelos com esquemas de vizinhança baseados em distância, passando a relação entre estas variâncias de 1,48 no modelo com vizinhança

baseada num raio de 3 metros, até 4,93 para o modelo com um raio de vizinhança de 10 metros. Isso, porém, não foi refletido nas estimativas do risco de infestação, já que estas foram similares dentro de um mesmo modelo para os diferentes esquemas de vizinhança (ver Figuras 8 e 9), sendo que somente seis plantas mudaram de categoria ao mudar a ordem do esquema de vizinhança. Já considerando somente o esquema de vizinhança de segunda ordem, os valores da razão s_θ^2/s_ϕ^2 foram, em geral, similares para modelos com diferentes distribuições *a priori* sobre seus parâmetros e hiperparâmetros, variando esta razão entre 1,35 e 1,38 com exceção do modelo 6 (com distribuição *a priori t* para o efeito aleatório sem estrutura espacial) que teve uma razão s_θ^2/s_ϕ^2 de 1,13. De forma semelhante, a escolha das *prioris* para os hiperparâmetros das distribuições dos efeitos aleatórios, também não teve influência sobre as estimativas dos riscos de infestação da praga, como pode ser visto na Figura 10.

Tabela 5. Estimativas pontual e por intervalo da variância marginal empírica para os efeitos aleatórios **q** e **f** e razão entre as estimativas das variâncias.

Modelo	S_θ^2 (estruturada espacialmente)			s_ϕ^2 (não estruturada)			Razão
		IC 95%			IC 95%		
	Média	LI	LS	Média	LI	LS	s_θ^2/s_ϕ^2
1	5,029	2,986	8,375	3,676	1,857	5,334	1,368
2	4,934	2,898	8,233	3,643	1,793	5,315	1,354
3	10,081	8,174	12,489	-	-	-	-
4	-	-	-	21,918	17,037	28,281	-
5	4,998	2,935	8,381	3,703	1,850	5,399	1,350
6	5,794	3,667	8,910	5,148	2,271	10,963	1,125
7	5,077	2,999	8,497	3,682	1,853	5,370	1,379
8	4,666	2,927	7,329	4,194	2,639	5,591	1,113
9	6,496	4,013	9,787	1,867	0,542	4,103	3,479
10	5,224	3,085	8,735	3,524	1,628	5,270	1,482
11	6,480	3,735	9,874	1,995	0,611	4,316	3,247
12	6,931	4,072	10,045	1,464	0,424	3,751	4,734
13	6,856	3,873	9,916	1,389	0,380	3,845	4,935

IC 95% = intervalo de credibilidade ao 95%
LI = limite inferior
LS = limite superior

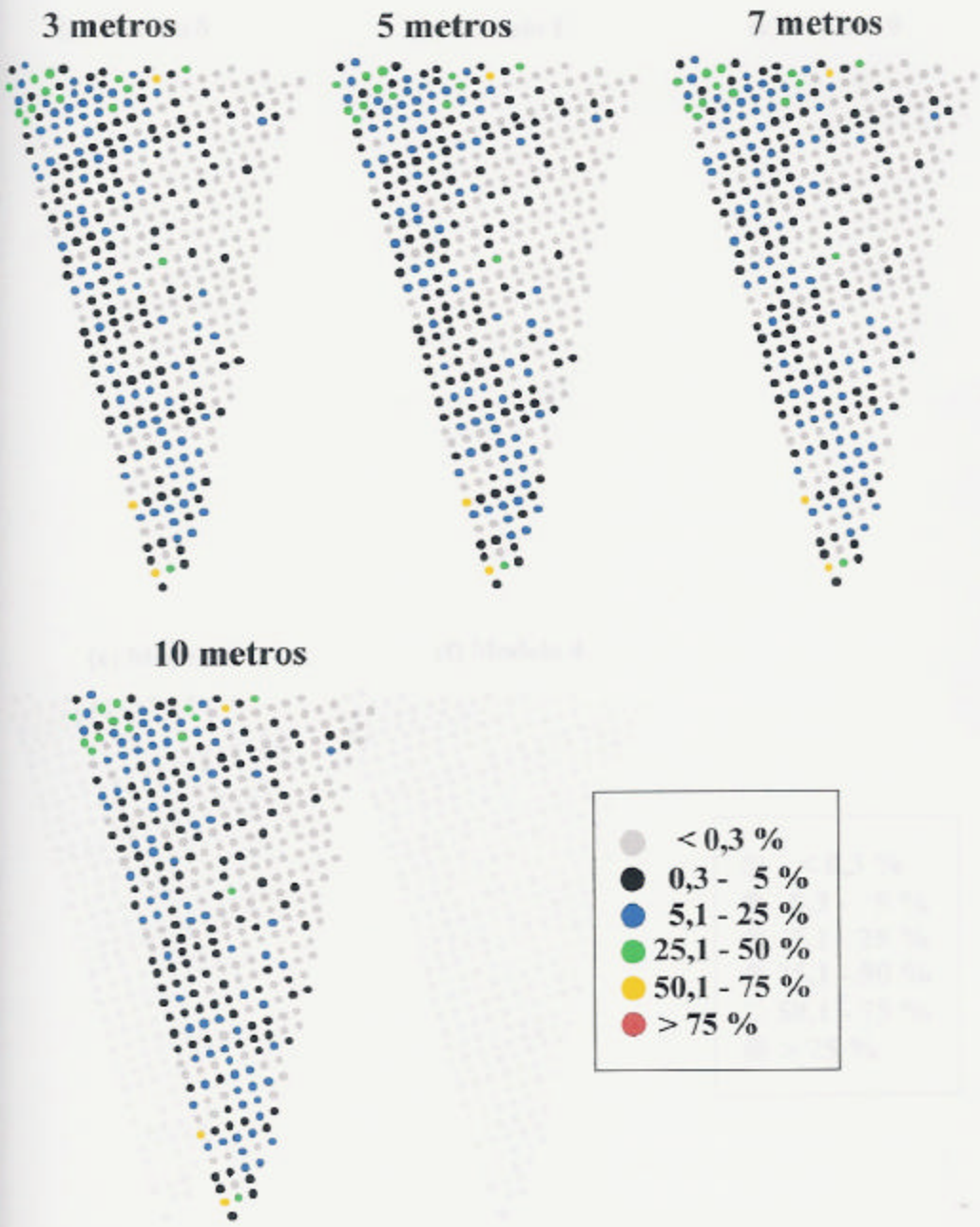

Figura 8 – Médias *a posteriori* do risco de infestação para os modelos com estrutura de vizinhança baseada em distância

Figura 9 – Médias *a posteriori* do risco de infestação da broca para os modelos com vizinhança baseada em adjacência: (a) 1ª ordem, (b) 2ª ordem, (c) 4ª ordem. (e) e (f) correspondem aos modelos (3) e (4) da Tabela 1.

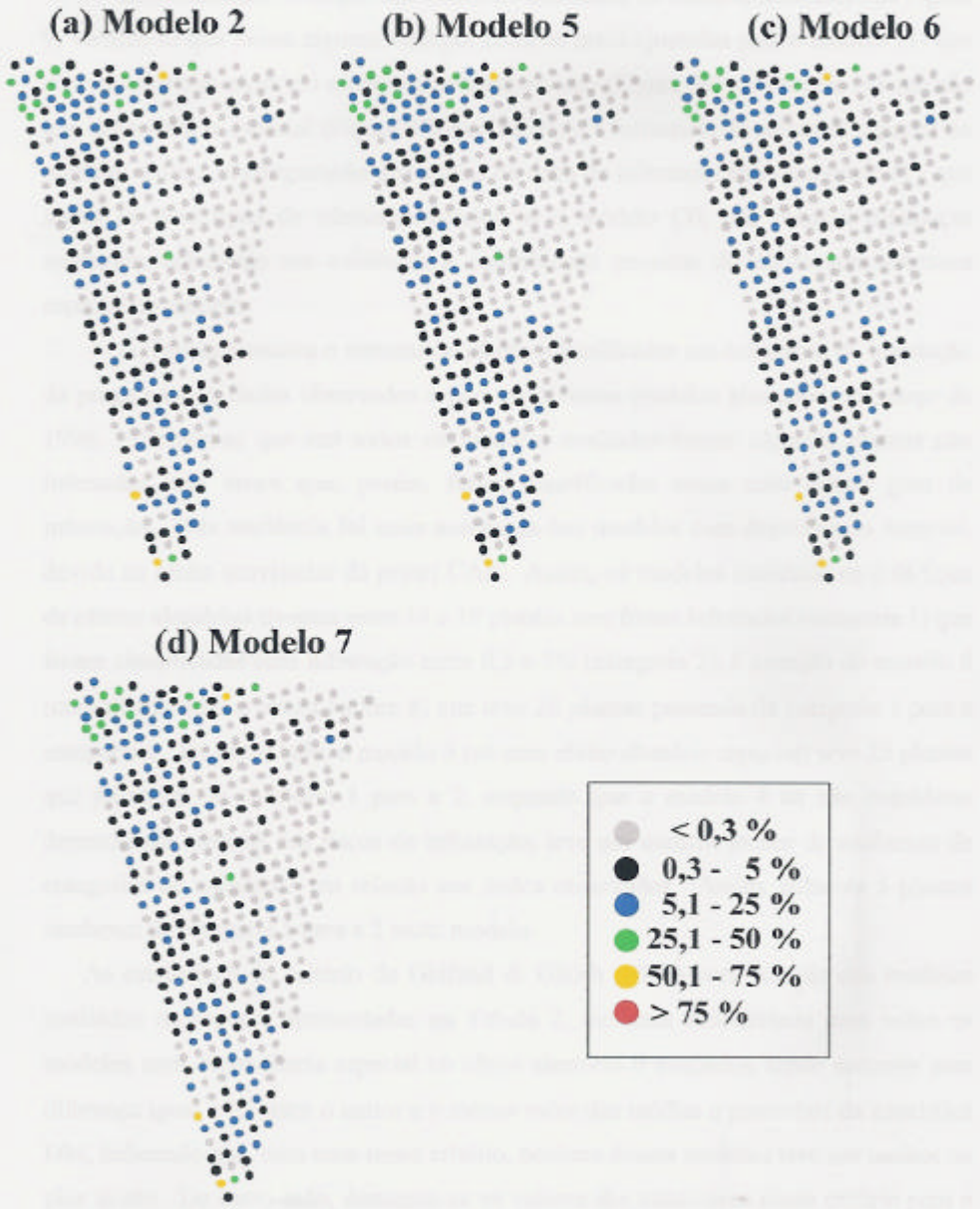

Figura 10 – Médias *a posteriori* do risco de infestação da broca nos modelos com diferentes *prioris* e *hiperprioris* para os efeitos aleatórios ϕ e θ.

Inspecionando os mapas dos riscos de infestação da broca apresentados na Figura 9, verifica-se que existe alguma variação entre as taxas ajustadas para o modelo (3) que só tem um efeito aleatório espacialmente estruturado (Figura 9e) e o modelo (4) que não tem dependência espacial (Figura 9f), confirmando a influencia da estrutura espacial no modelo. Porém, a comparação dos mapas de risco da infestação entre o modelo (1), que inclui os dois tipos de efeitos aleatórios, e o modelo (3), não revelou diferenças aparentes, colocando em evidência a contribuição pequena do efeito sem estrutura espacial no modelo.

A Tabela 6 mostra o número de plantas classificadas em categorias de infestação da praga para os dados observados e para os diversos modelos ajustados em março de 1996. Verifica-se que em todos os modelos avaliados houve algumas plantas não infestadas pela broca que, porém, foram classificadas como com algum grau de infestação. Esta tendência foi mais acentuada nos modelos com dependência espacial, devido ao efeito suavizador da *priori* CAR. Assim, os modelos incluindo os dois tipos de efeitos aleatórios tiveram entre 14 e 19 plantas sem frutos infestados (categoria 1) que foram classificadas com infestação entre 0,3 e 5% (categoria 2), a exceção do modelo 6 (com distribuição *a priori t* sobre ϕ) que teve 28 plantas passando da categoria 1 para a categoria 2. De outro lado, o modelo 3 (só com efeito aleatório espacial) teve 25 plantas que passaram da categoria 1 para a 2, enquanto que o modelo 4 ao não considerar dependência espacial nos riscos de infestação, teve um número menor de mudanças de categorias de infestação em relação aos dados observados. Assim, somente 5 plantas mudaram de categoria 1 para a 2 neste modelo.

As estatísticas do critério de Gelfand & Ghosh (1998) para seleção dos modelos avaliados no espaço, apresentadas na Tabela 7, mostram similaridade para todos os modelos com dependência espacial no efeito aleatório θ avaliados, tendo somente uma diferença igual a 25 entre o maior e o menor valor das médias *a posteriori* da estatística DM, indicando que, com base nesse critério, nenhum desses modelos teve um melhor ou pior ajuste. De outro lado, destacam-se os valores das estatísticas desse critério para o modelo (4) que não considera dependência espacial em seu parâmetro ϕ, tendo uma

maior média *a posteriori* para DM com uma diferença de 47 unidades acima do maior valor de DM para os outros 12 modelos. Este modelo também foi o mais penalizado.

Levando em consideração os resultados obtidos, o modelo (3) foi escolhido como o modelo básico para mapear os riscos de infestação no espaço durante os outros nove meses, cujas médias *a posteriori* são representadas graficamente nas Figuras 11 e 12. A Tabela 8 apresenta as estimativas *a posteriori* dos parâmetros desse modelo para os meses de julho de 1995 até abril de 1996. Pode ser visto que as estimativas dos parâmetros para o período julho de 1995 a janeiro de 1996 são muito similares, mas diferem das estimativas para os meses de fevereiro/96 até abril/96, período em que a infestação é muito mais dinâmica e seus níveis se incrementam rapidamente. Isso se viu refletido nos mapas das médias *a posteriori* dos riscos de infestação das Figuras 11 e 12.

Tabela 6. Classificação do número de plantas por categorias de infestação para os dados observados e os modelos avaliados na análise espacial em março/96

Modelo	Categoria de infestação				
	< 0,3%	0,3 - 5%	5,1 - 25%	25,1 - 50%	50,1 - 75%
1	156	137	82	14	3
2	156	137	82	14	3
3	150	142	83	14	3
4	170	117	87	15	3
5	156	137	82	14	3
6	147	146	82	14	3
7	155	138	82	14	3
8	156	137	82	14	3
9	159	134	82	14	3
10	155	138	82	14	3
11	155	138	82	14	3
12	159	134	82	14	3
13	161	132	82	14	3
Obs*	175	113	87	14	3

* dados observados

Tabela 7. Estatísticas do critério de Gelfand & Ghosh (1998) para seleção dos modelos avaliados em março de 1996

Modelo	DM	PM	GM
1	8747	8695	52
2	8737	8684	53
3	8749	8687	62
4	8799	8793	6
5	8738	8686	52
6	8741	8687	54
7	8734	8683	51
8	8727	8668	59
9	8752	8697	55
10	8745	8693	52
11	8752	8704	48
12	8745	8700	45
13	8747	8706	41

DM = *Deviance* preditiva esperada (DM = PM + GM)
PM = termo de penalização
GM = Medida de bondade de ajuste;

Tabela 8. Estimativas *a posteriori* dos parâmetros do modelo CAR.

Mês	Parâmetro	Média	Desvio padrão	Erro de Monte Carlo	Quantil 2.50%	mediana	Quantil 97.50%
jul/95	DM	1560	392,49	5,5937	943,1	1507,3	2472,5
ago/95	DM	1778	428,02	6,2152	1087,6	1727,6	2770,3
set/95	DM	1865	434,05	6,1357	1161,9	1814,2	2855
out/95	DM	1958	466,75	6,3672	1202,7	1904,2	3022,5
nov/95	DM	1381	347,92	4,9065	826,9	1338,9	2178,4
dez/95	DM	1107	275,22	4,0502	668,2	1073,4	1738,8
jan/96	DM	1735	392,05	5,4852	1090	1695,3	2616,1
fev/96	DM	5701	1012,55	14,4920	3966,3	5608,7	7928,9
mar/96	DM	8756	1320,20	17,2210	6456,1	8665,8	11602
abr/96	DM	14185	1859,50	25,3210	10920	14048	18165
jul/95	α	-11,332	0,9109	0,018862	-13,317	-11,26	-9,7624
ago/95	α	-11,709	0,9507	0,020415	-13,778	-11,628	-10,0664
set/95	α	-11,871	0,9675	0,020178	-14,008	-11,788	-10,215
out/95	α	-11,081	0,8024	0,015637	-12,826	-11,015	-9,6998
nov/95	α	-10,359	0,6705	0,013166	-11,819	-10,302	-9,1973
dez/95	α	-11,524	0,8707	0,019192	-13,43	-11,449	-10,0367
jan/96	α	-11,344	0,8148	0,017429	-13,123	-11,276	-9,9397
fev/96	α	-8,946	0,4885	0,007984	-9,993	-8,9129	-8,0853
mar/96	α	-5,326	0,1167	0,001631	-5,5672	-5,3217	-5,1093
abr/96	α	-4,275	0,0682	0,000944	-4,4146	-4,2735	-4,1471
jul/95	σ_q	9,274	1,0562	0,020616	7,4429	9,1979	11,581
ago/95	σ_q	9,546	1,0803	0,021575	7,6772	9,4582	11,909
set/95	σ_q	10,130	1,1226	0,021378	8,1866	10,0436	12,605
out/95	σ_q	9,005	0,9742	0,018653	7,3215	8,9283	11,122
nov/95	σ_q	7,900	0,8578	0,015734	6,4035	7,8312	9,7606
dez/95	σ_q	8,415	0,9881	0,020101	6,7147	8,3398	10,564
jan/96	σ_q	8,531	0,9374	0,018122	6,9054	8,4579	10,567
fev/96	σ_q	8,894	0,7297	0,011294	7,5939	8,8545	10,445
mar/96	σ_q	6,019	0,3557	0,004860	5,3671	6,0044	6,7575
abr/96	σ_q	5,595	0,2919	0,004071	5,0561	5,5846	6,2001

DM = *Deviance* preditiva esperada

α = intercepto do modelo (3)

σ_q = desvio padrão do parâmetro θ. $\sigma_q = (1/\tau_q)^{1/2}$

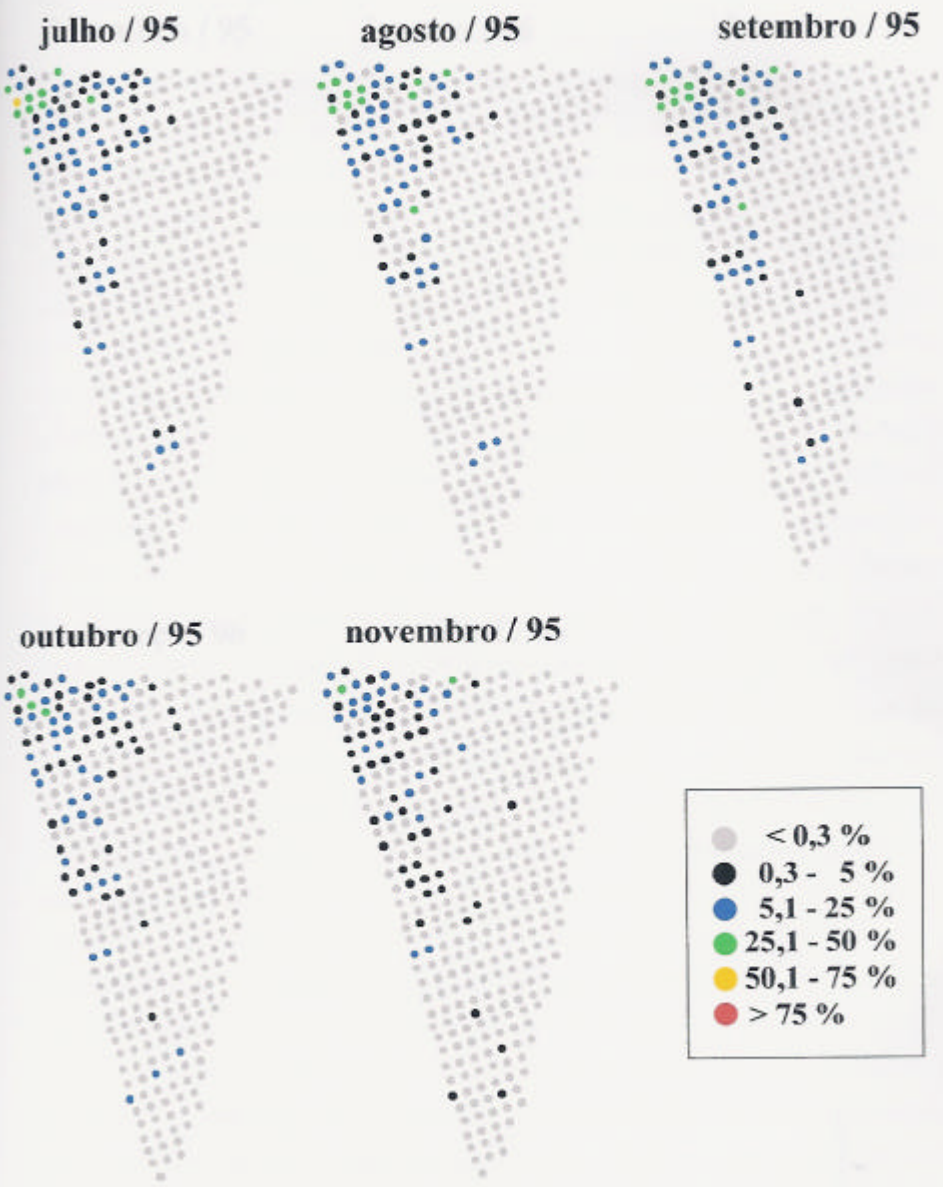

Figura 11 – Médias *a posteriori* do risco de infestação da broca para o modelo 3 (CAR) para os meses de julho/95 a novembro/95

Figura 12 – Médias *a posteriori* do risco de infestação da broca para o modelo 3 (CAR) para os meses de dezembro/95 a abril/96

4.2 Análise espaço-temporal

Os valores obtidos para as médias *a posteriori* dos riscos de infestação da broca para os modelos que levaram em consideração os fatores espaço e tempo (ver Tabela 3), são apresentados em forma de mapas nas Figuras 13 a 16 e Figuras 17 a 20, respectivamente, para modelos com tendências de crescimento linear e quadrática.

O parâmetro γ cujas estimativas estão apresentadas na Figura 21 para os modelos com tendência linear mede o grau de aumento ou decréscimo no nível de infestação de cada planta ao longo do tempo, permitindo identificar áreas de crescimento da epidemia. Percebe-se uma diferença na tendência de crescimento da infestação entre os modelos que não consideram estrutura espacial para o parâmetro γ (modelos 1 e 3) e aqueles modelos com dependência espacial sobre esse parâmetro (modelos 2 e 4). Assim, a magnitude de crescimento da infestação ao longo do tempo para plantas com valores de γ positivos, foi maior nos modelos 1 e 3 do que nos modelos 2 e 4, sendo que, no primeiro caso, 130 e 122 plantas, respectivamente, tiveram valores de $\hat{g} > 1$, enquanto que, no segundo caso, somente 3 e 6 plantas, respectivamente, tiveram valores de $\hat{g} > 1$. Para os modelos 1 e 3, embora se note alguma tendência geográfica, em que plantas localizadas na região noroeste têm um crescimento na infestação menor ($\hat{g} \leq 1$) do que plantas localizadas na região nordeste do mapa, o padrão espacial não é muito bem definido, alternando plantas com valores de $\hat{g}$ positivos e negativos em toda a área. Já nos modelos com estrutura espacial sobre γ, nota-se uma clara tendência de crescimento nas regiões nordeste e centro-oeste do mapa, e de decréscimo na magnitude da infestação para a região noroeste. Este comportamento é mais coerente com o que mostram os dados originais (ver Figura 6), dado que, a infestação da broca começou na região sul e foi avançando até colonizar toda a área em abril de 1996.

Já nos modelos com tendência de crescimento quadrática, são os parâmetros γ e ν os que determinam a magnitude do aumento ou decréscimo da infestação ao longo do tempo, sendo o parâmetro γ quem determina a velocidade de crescimento da infestação, enquanto que o parâmetro ν determina a aceleração do processo. Representações gráficas das médias *a posteriori* para esses parâmetros são apresentadas nas Figuras 22 e

Figura 13 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (1)

Figura 14 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (2)

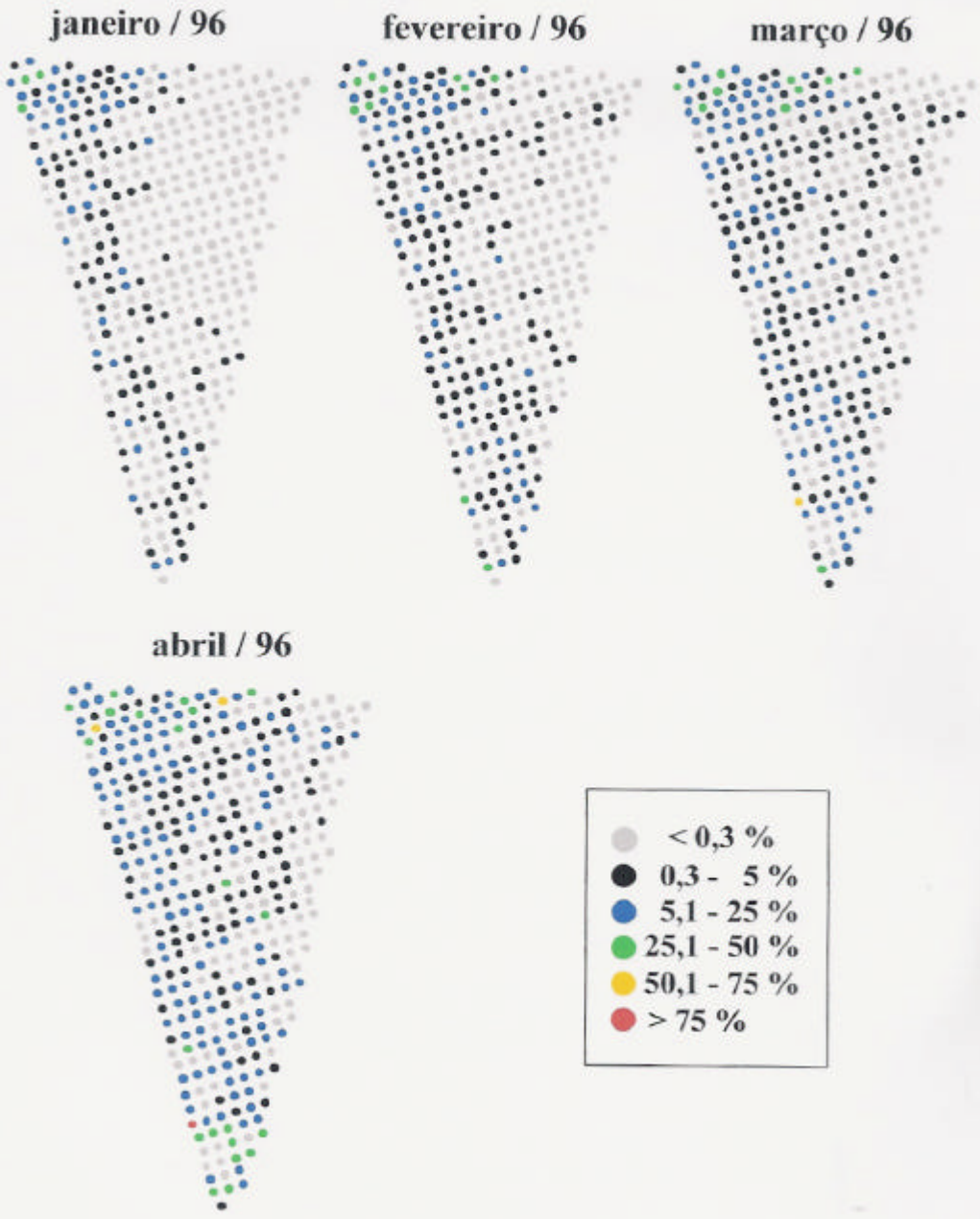

Figura 15 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (3)

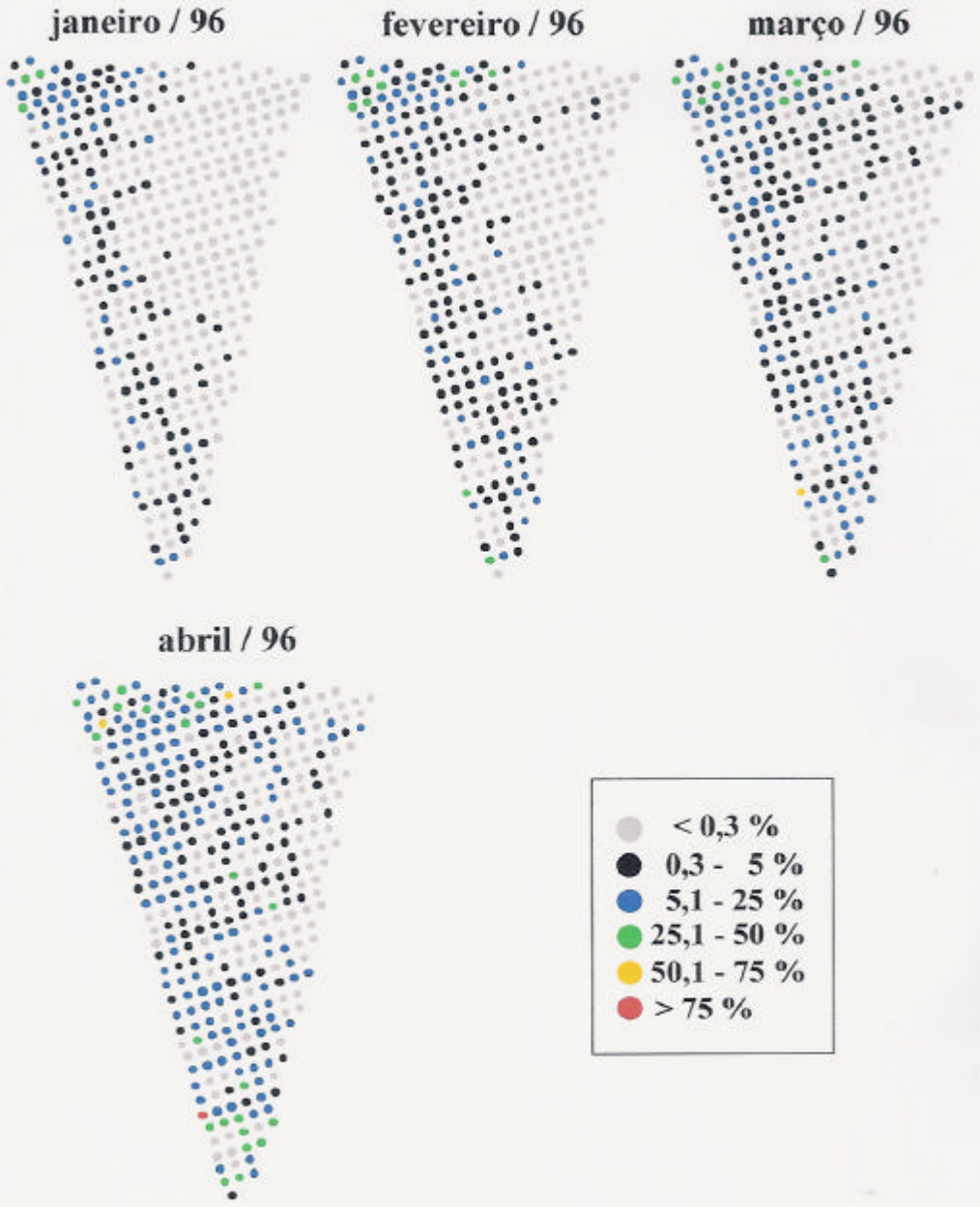

Figura 16 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (4)

Figura 17 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (5)

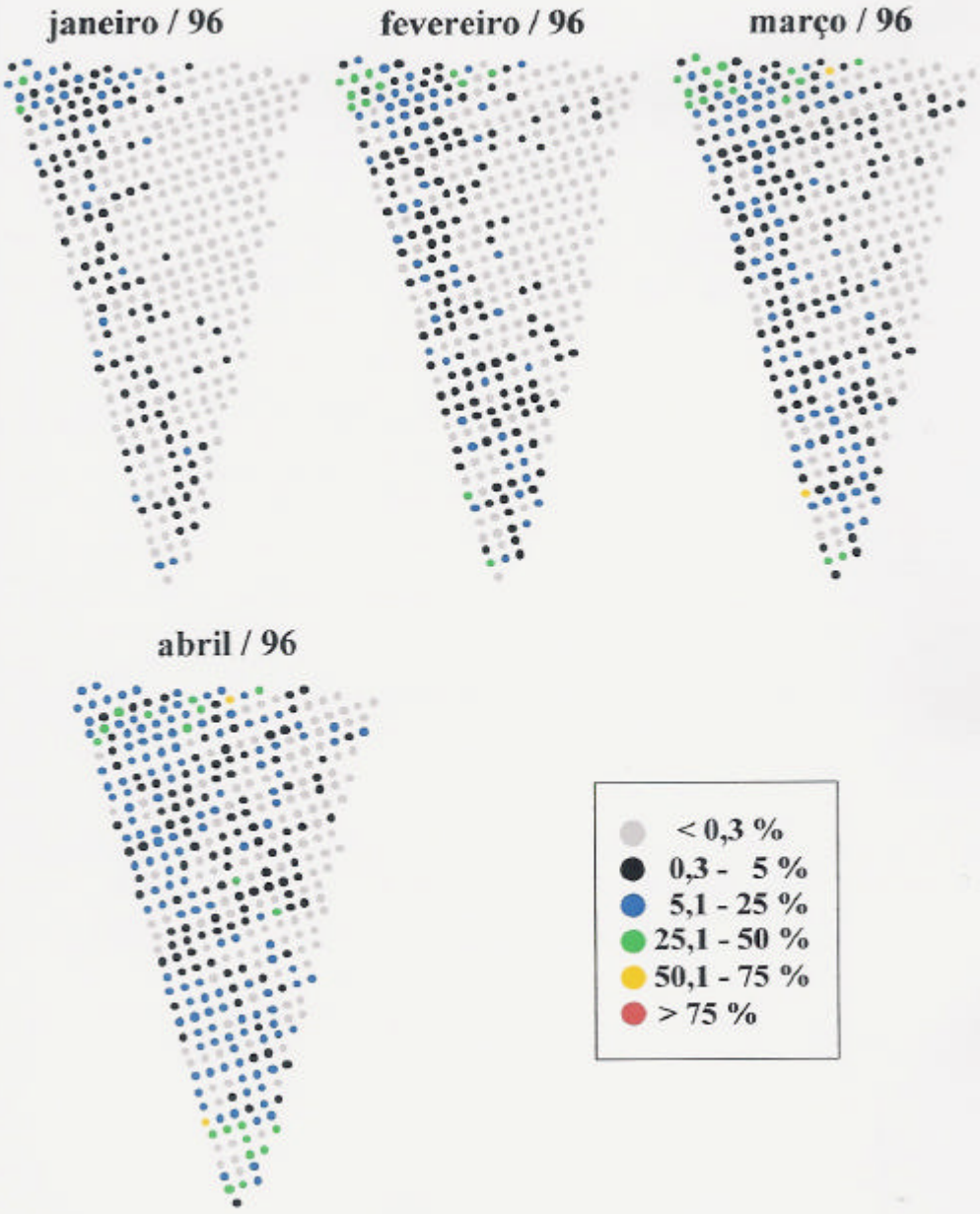

Figura 18 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (6)

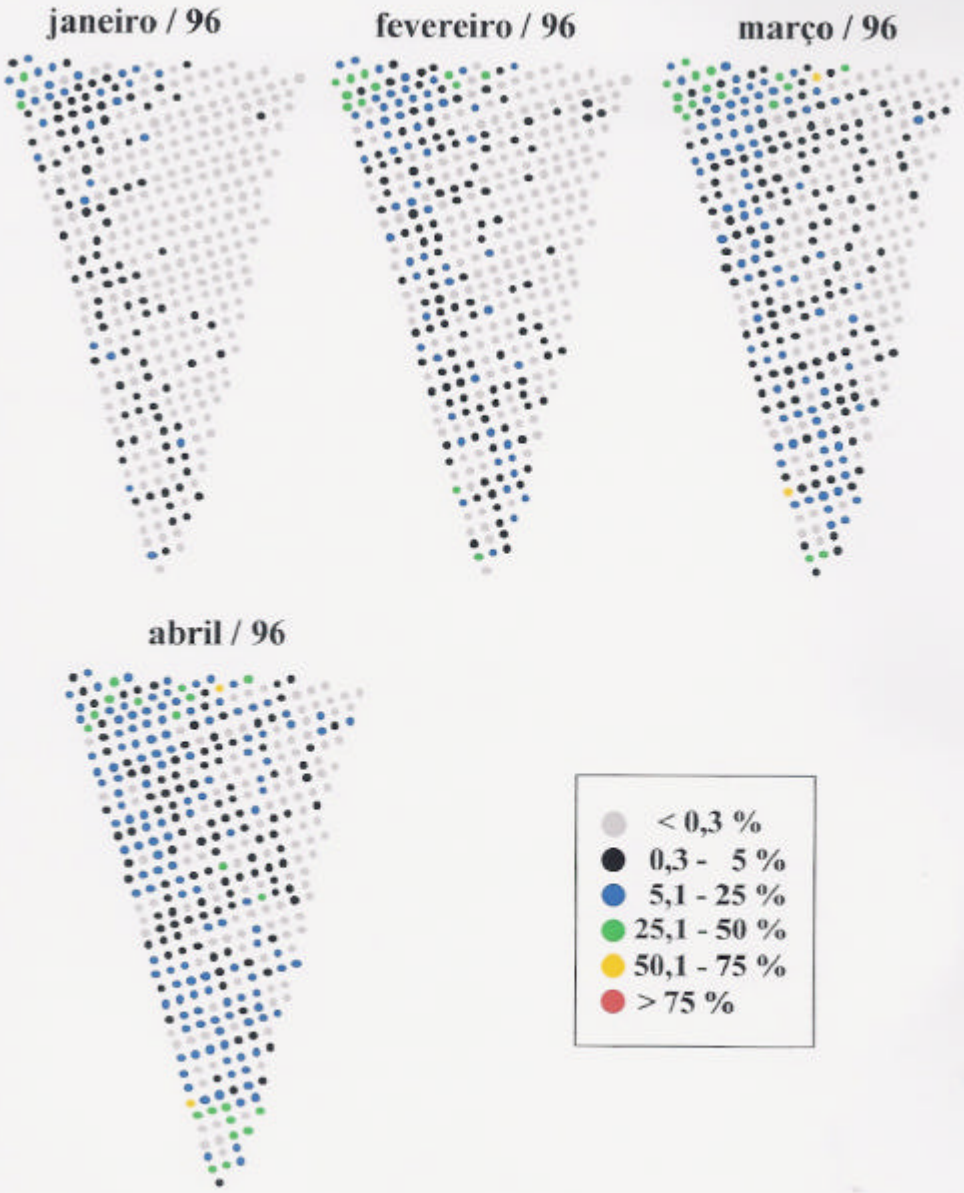

Figura 19 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (7)

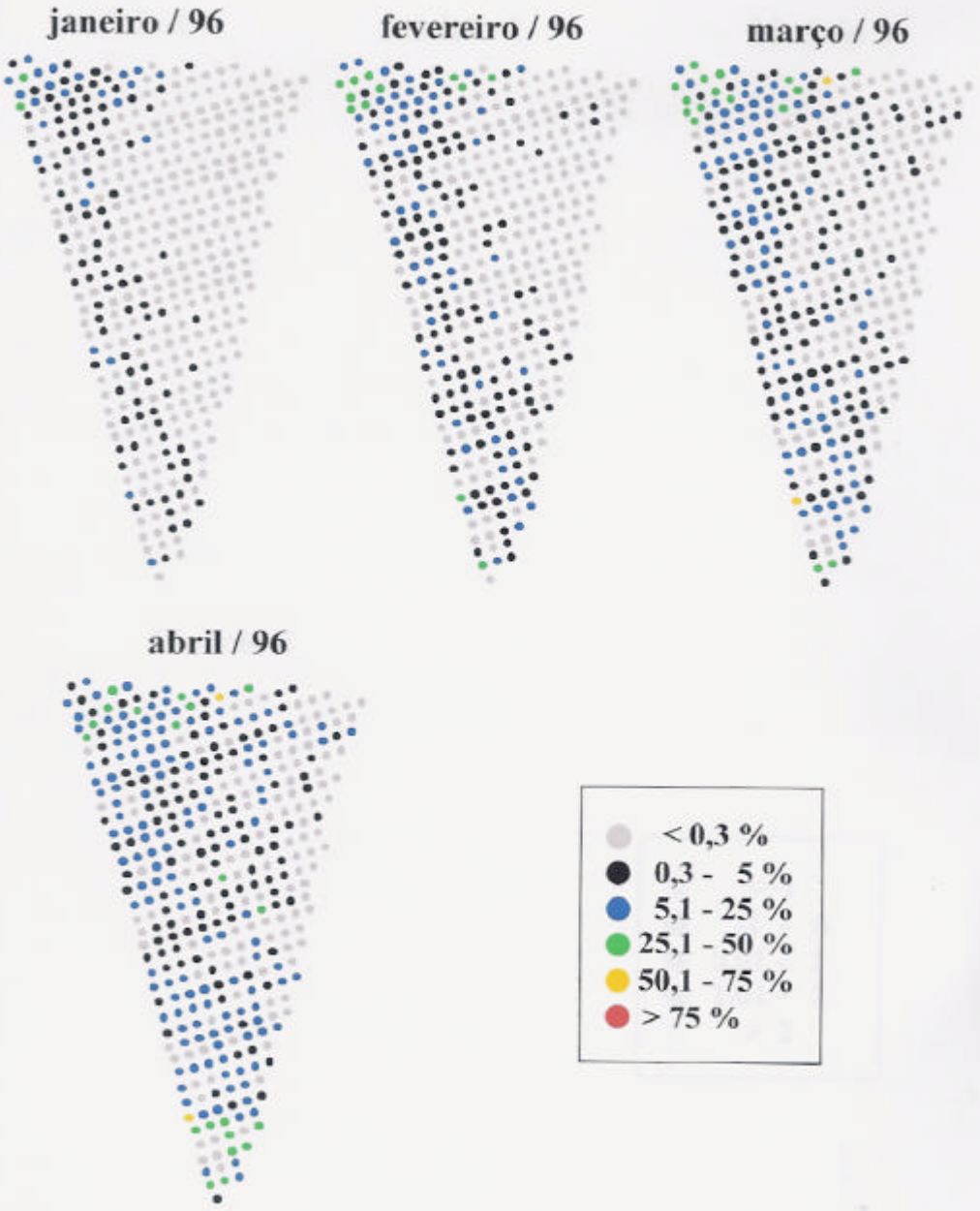

Figura 20 – Médias *a posteriori* dos riscos de infestação da broca para o modelo espaço-tempo (8)

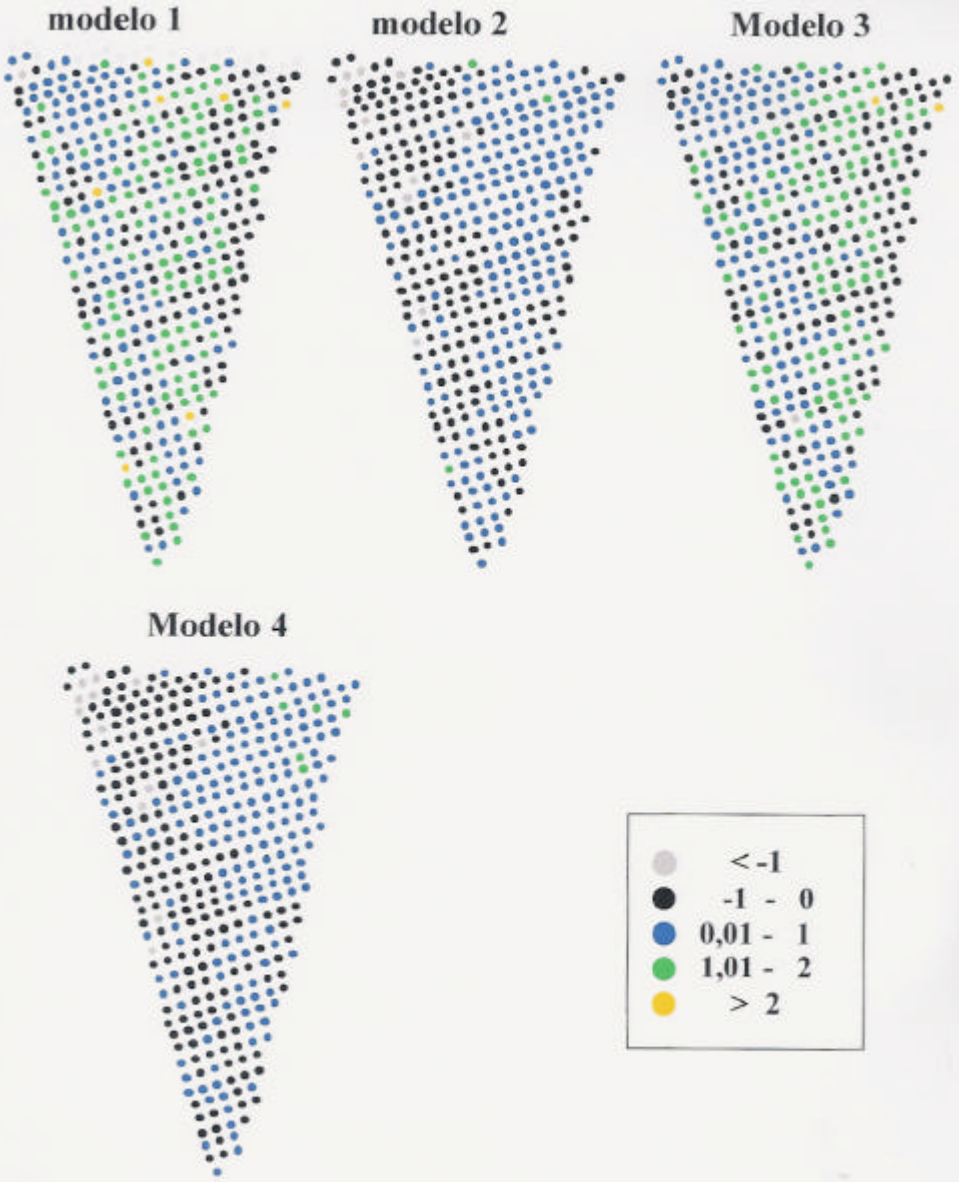

Figura 21 – Médias *a posteriori* do parâmetro γ para os modelos espaço-tempo com tendência de crescimento linear

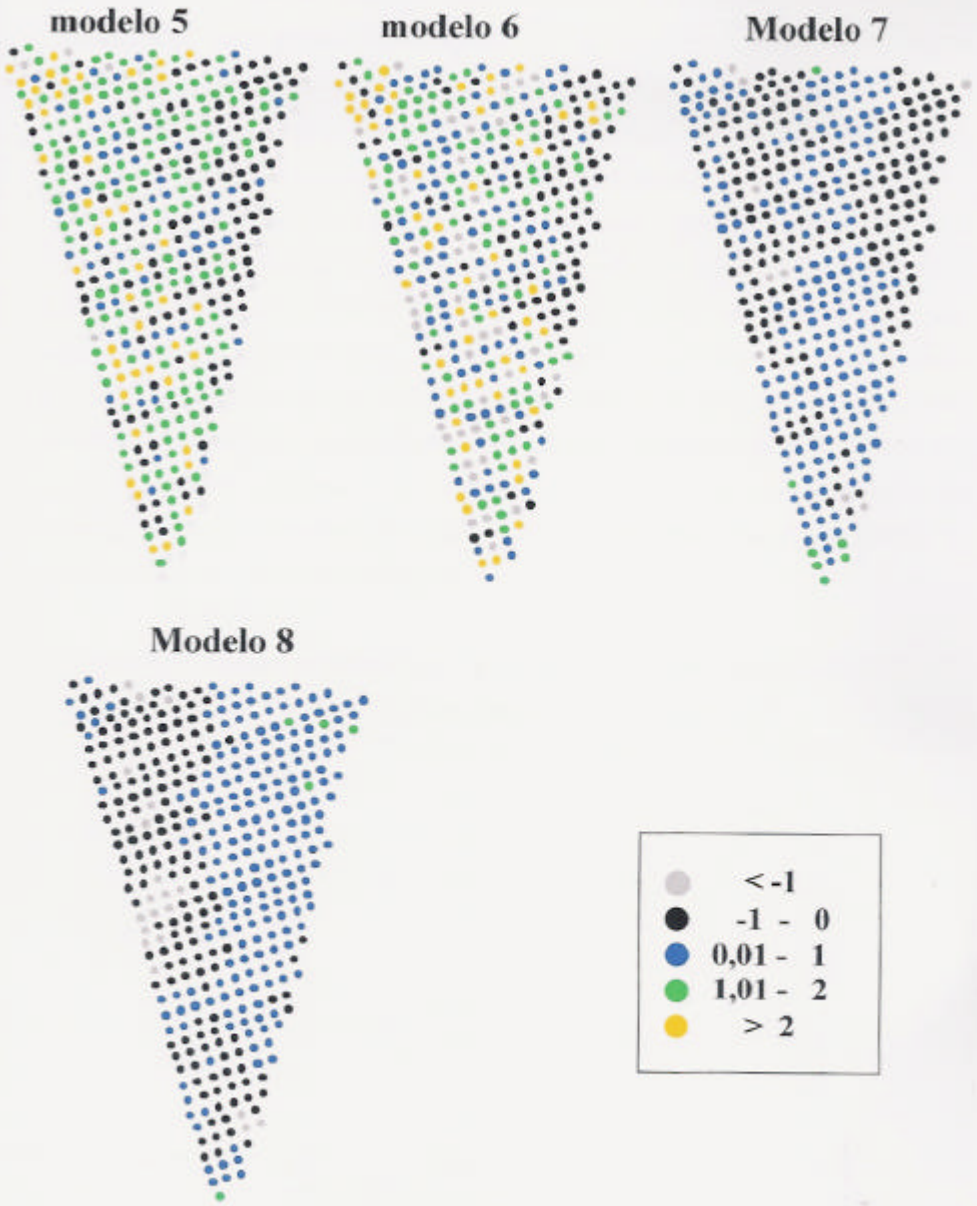

Figura 22 – Médias *a posteriori* do parâmetro γ para os modelos espaço-tempo com tendência de crescimento quadrática

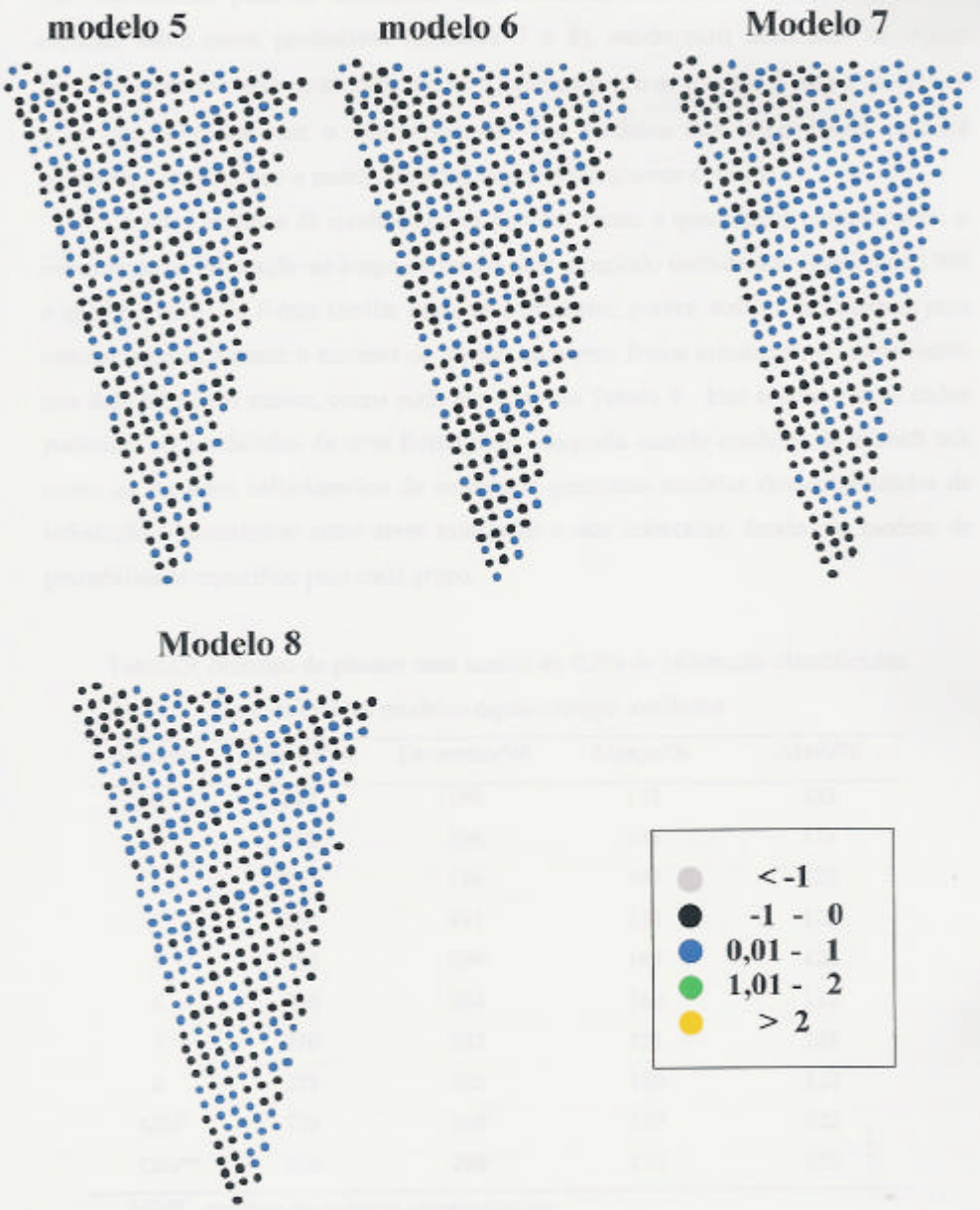

Figura 23 – Médias *a posteriori* do parâmetro ν para os modelos espaço-tempo com tendência de crescimento quadrática

23. Novamente pode se diferenciar uma tendência nos modelos com dependência espacial sobre esses parâmetros (modelos 7 e 8), sendo mais acentuada na região nordeste e centro-oeste no modelo 8 para o parâmetro γ, e no modelo 7 para o parâmetro ν, o que contrasta com o comportamento dos modelos sem dependência espacial (modelos 5 e 6) em que o padrão espacial não está claramente definido.

Ambos os tipos de modelos (com preditor linear e quadrático) representaram o incremento na infestação ao longo do tempo para o período considerado (janeiro de 1996 a abril de 1996) de forma similar aos dados originais; porém, todos eles falharam para estimar adequadamente o excesso de plantas com zero frutos infestados, principalmente nos dois primeiros meses, como pode ser visto na Tabela 9. Isso sugere que os dados poderiam ser modelados de uma forma mais adequada usando modelos de mistura tais como os modelos inflacionados de zeros que permitem modelar descontinuidades de infestação e discriminar entre áreas infestadas e não infestadas, dando um modelo de probabilidade específico para cada grupo.

Tabela 9. Número de plantas com menos de 0,3% de infestação classificadas nos diferentes modelos espaço-tempo avaliados

Modelo	Janeiro/96	Fevereiro/96	Março/96	Abril/96
1	247	196	152	123
2	252	196	148	121
3	250	186	147	123
4	254	193	154	122
5	254	209	167	122
6	260	204	160	122
7	270	207	151	121
8	273	205	150	122
MM*	329	269	150	122
Obs**	329	280	175	120

MM* modelo de mistura espaço-tempo
Obs** dados observados

As estatísticas do critério de Gelfand & Ghosh (1998) para comparação entre os modelos espaço-tempo avaliados estão apresentadas na Tabela 10. Pode-se notar um melhor ajuste com base nesse critério para os modelos com tendência de crescimento quadrática (modelos 4 a 8), em relação aos modelos com tendência de crescimento linear (modelos 1 a 4). Já dentro de cada tipo de modelos, não houve diferença no ajuste entre modelos com e sem dependência espacial no efeito aleatório γ para os modelos com crescimento linear, enquanto que, alguma diferença no ajuste pode ser notada em favor dos modelos sem dependência espacial nos parâmetros que determinam o crescimento da infestação para o caso quadrático.

Tabela 10. Estatísticas do critério de Gelfand & Ghosh (1998) para seleção dos modelos avaliados em espaço e tempo

MODELO	DM	PM	GM
1	75775	24164	51611
2	76158	24103	52055
3	76247	24034	52213
4	77496	24763	52732
5	44281	26573	17708
6	44526	26274	18252
7	50933	27672	23261
8	52019	27700	24319
MM*	68515	24811	43704

DM = *Deviance* preditiva esperada (DM = PM + GM)
PM = termo de penalização; GM = Medida de bondade de ajuste
MM* modelo de mistura espaço-tempo

O melhor ajuste dos modelos quadráticos em relação aos lineares foi verificado também com base nos gráficos de logit (π_{it}) vs. tempo apresentados nas Figuras 24 a 28 para 40 das 392 plantas da área sob estudo. As plantas selecionadas tiveram frutos infestados durante todo o período de estudo, não sendo necessário, portanto, estimar nenhum valor sobre elas. A comparação do logit ($\hat{p}_{it}$) para os dados observados (em que $\hat{p}_{it} = y_{it} / n_{it}$) com o logit ($\hat{p}_{it}$) para os diferentes modelos mostra uma tendência dos

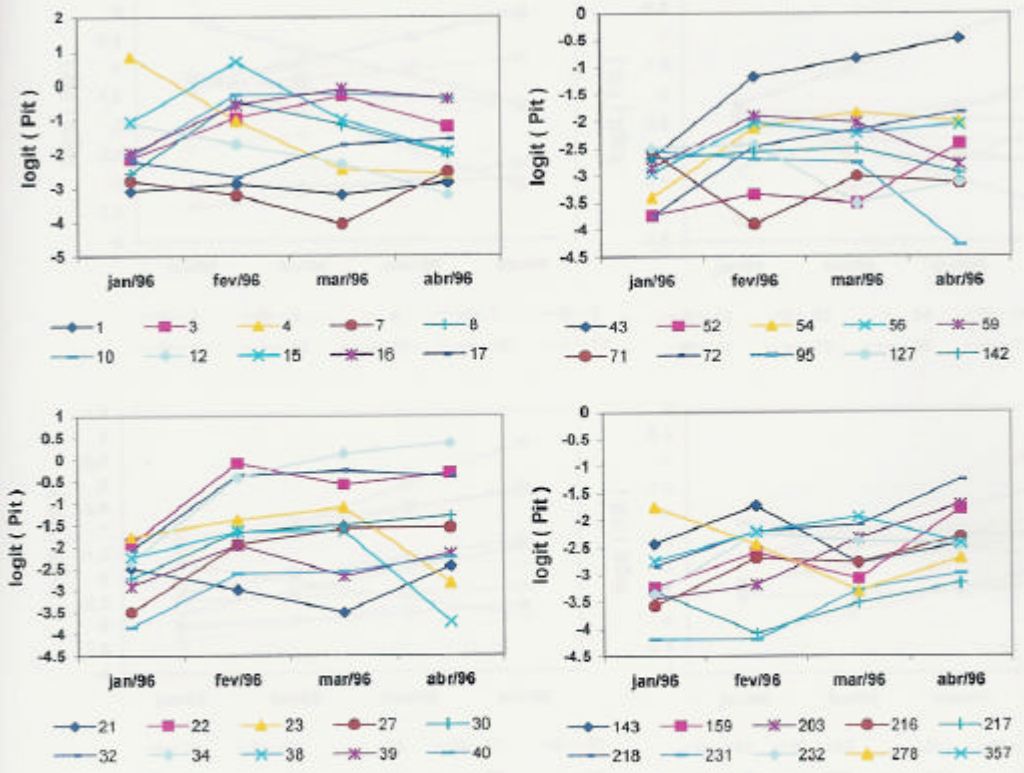

Figura 24 - Gráfico de logit (pit) vs tempo para 40 plantas da área sob estudo (dados observados).

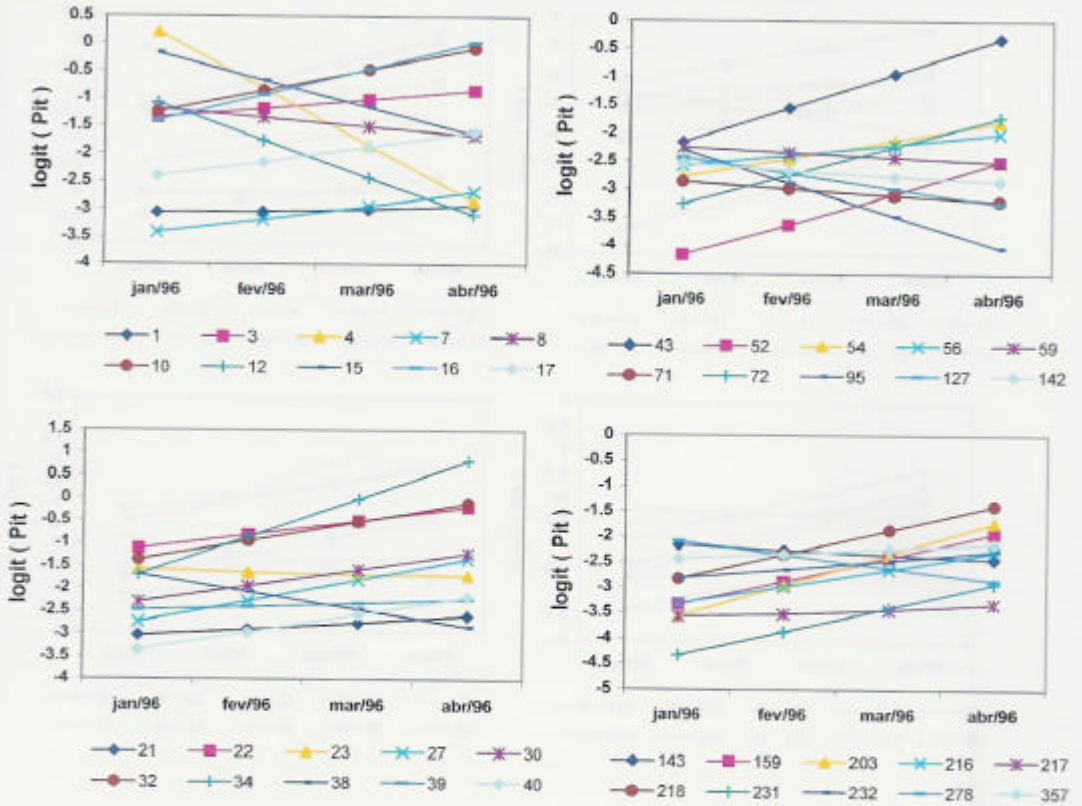

Figura 25 - Gráfico de logit (pit) vs tempo para 40 plantas da área sob estudo (modelo espaço-tempo 1).

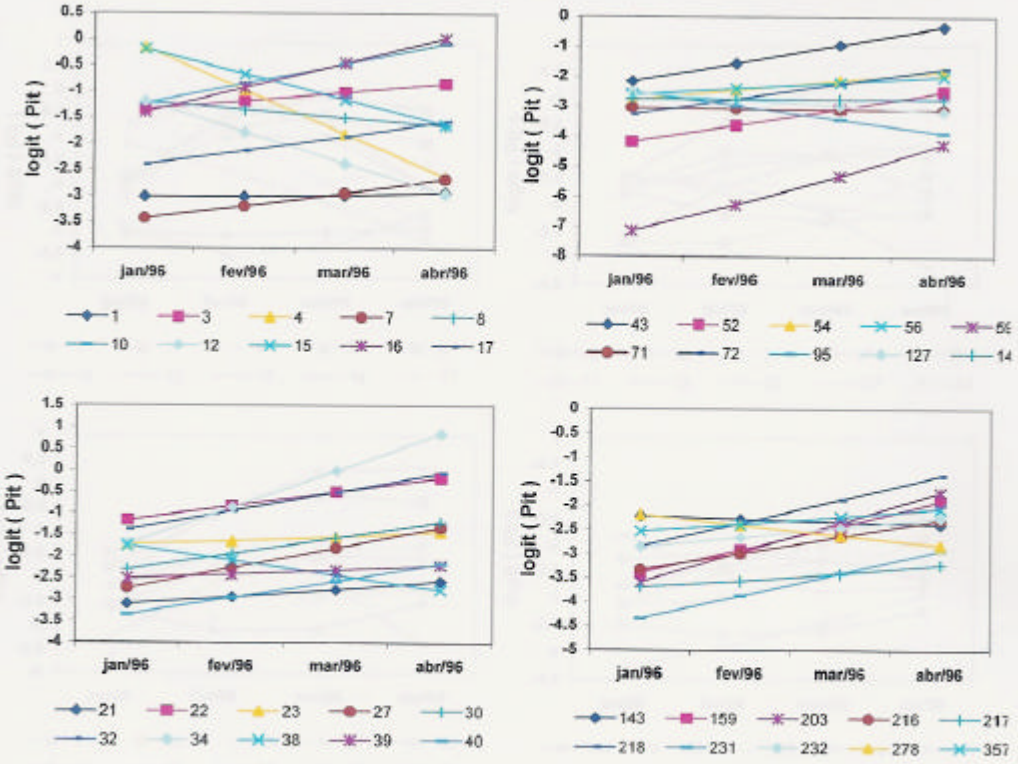

Figura 26 - Gráfico de logit (pit) vs tempo para 40 plantas da área sob estudo (modelo espaço-tempo 2)

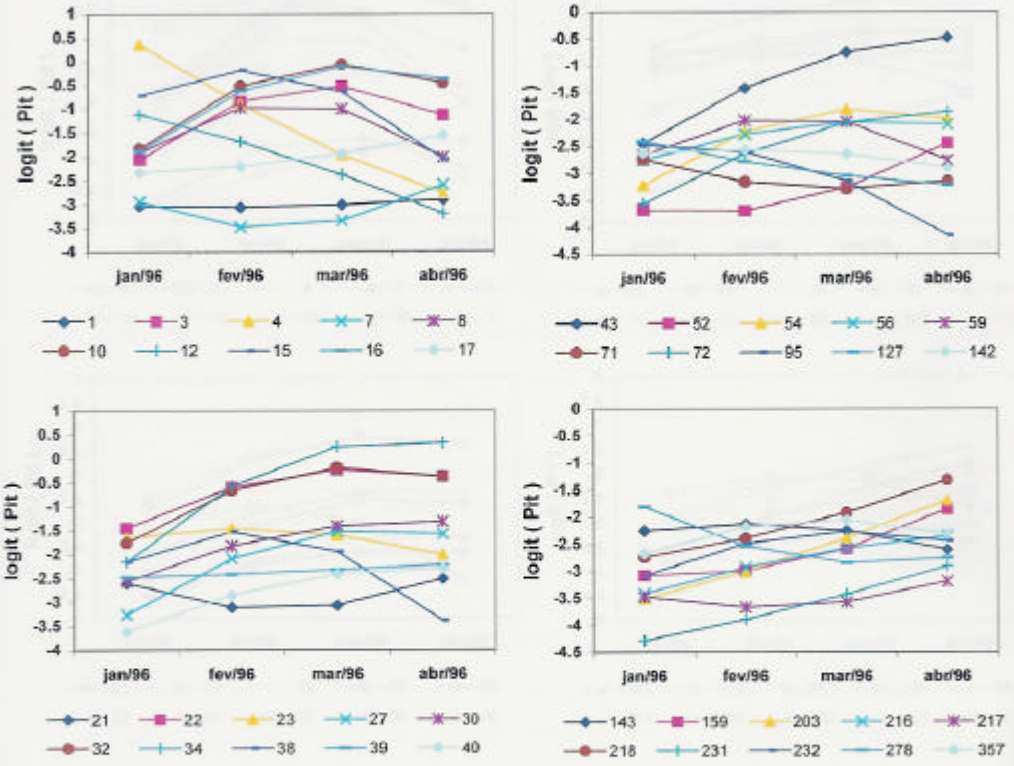

Figura 27 - Gráfico de logit (pit) vs tempo para 40 plantas da área sob estudo (modelo espaço-tempo 5)

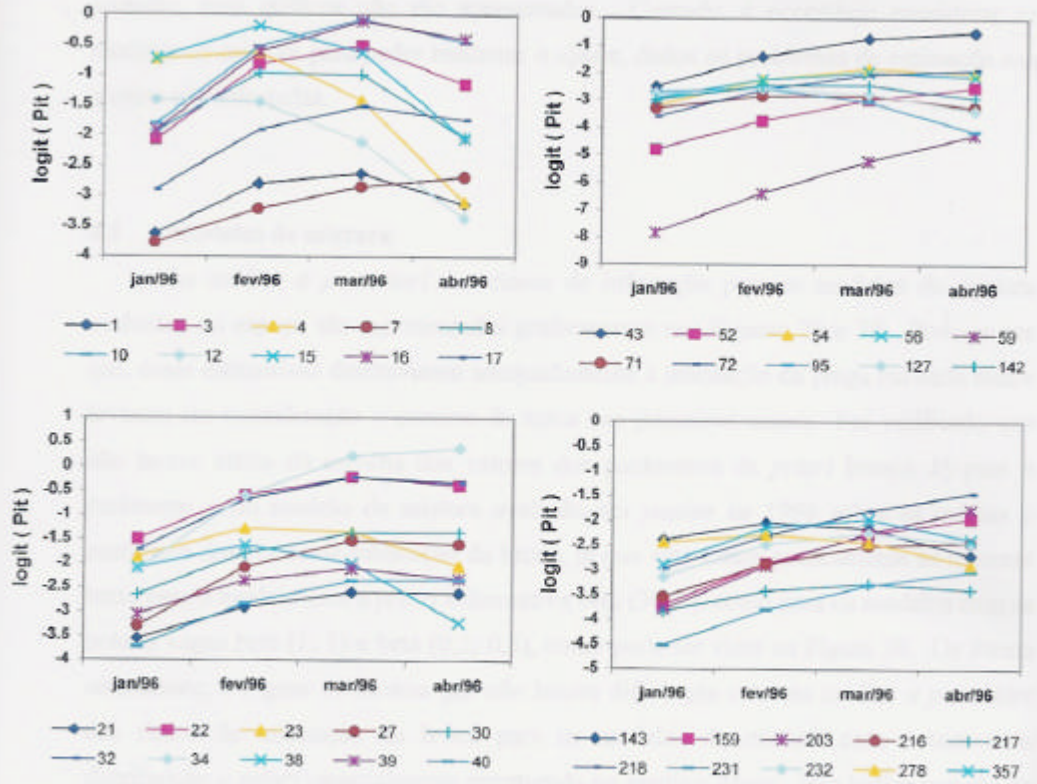

Figura 28 - Gráfico de logit (pit) vs tempo para 40 plantas da área sob estudo (modelo espaço-tempo 7)

modelos com crescimento quadrático mais parecida com o comportamento observado nos dados originais. A tendência do logit ($\hat{p}_{it}$) ao longo do tempo para os modelos espaço-tempo 3, 4, 6 e 8 foi muito parecida com a registrada nos modelos 1, 2, 5 e 7 e portanto, seus gráficos não são apresentados. Contudo, é necessário considerar os modelos de mistura para poder melhorar o ajuste, dados os problemas de estimação nas plantas não infestadas.

4.3 Modelos de mistura

As médias *a posteriori* dos riscos de infestação para os modelos de mistura avaliados no espaço são representadas graficamente nas Figuras 29 e 30. Pode-se ver que, essas estimativas descreveram adequadamente a infestação da praga em cada mês e levaram em consideração o excesso de zeros nos primeiros meses. Foi verificado que não houve efeito da escolha dos valores dos parâmetros da *priori* beta(a, b) para o parâmetro p do modelo de mistura avaliado em janeiro de 1996 sobre as médias *a posteriori* dos riscos de infestação da broca, já que elas foram praticamente as mesmas tanto para o modelo com a *priori* informativa beta (30; 5), como para os modelos com as *prioris* vagas beta (1; 1) e beta (0,5; 0,5), como pode ser visto na Figura 30. De forma semelhante, a Figura 30 mostra que não houve diferenças entre as médias *a posteriori* dos riscos de infestação da broca para os modelos de mistura com e sem uma distribuição *a priori* espacialmente estruturada no preditor linear. Isso indica que são os dados e não a informação *a priori*, que estão direcionando o ajuste do modelo nesse caso.

O segundo modelo de mistura considerado na metodologia teve problemas de convergência devido à alta correlação entre amostras para os parâmetros m_1 e τ_1, o que não foi possível corrigir mesmo após de 55000 iterações com um *burn-in* de 5000 e guardando uma a cada 25 iterações. Para poder ajustar adequadamente este modelo seria necessário fazer algum tipo de re-parametrização no preditor linear, trabalho que está fora dos objetivos desta dissertação mas que poderia ser abordado em trabalhos futuros. Portanto, esse modelo não foi implementado.

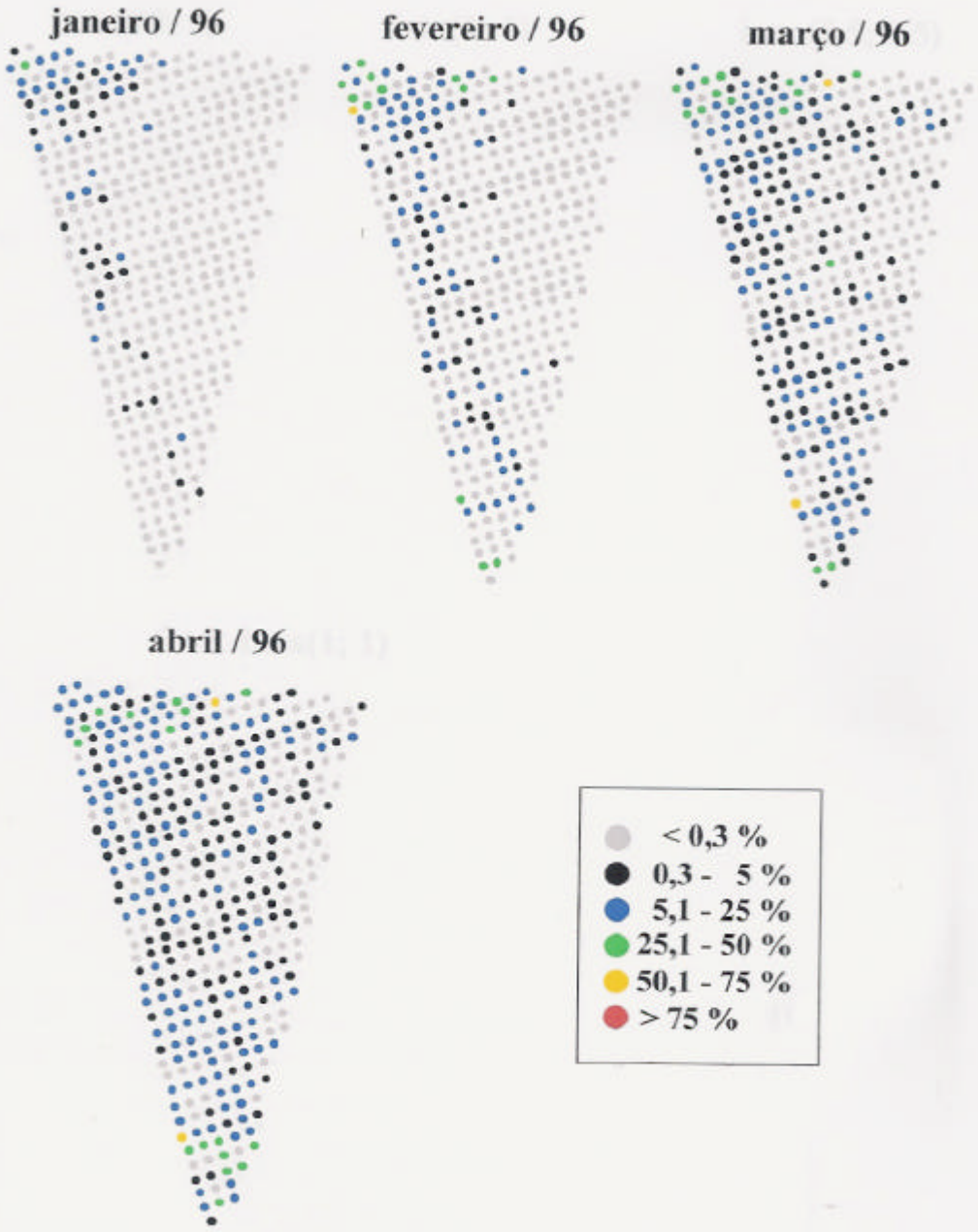

Figura 29 – Médias *a posteriori* do risco de infestação da broca para o modelo de mistura avaliado separadamente em cada tempo (com *priori* beta (1,1) para o parâmetro *p)*

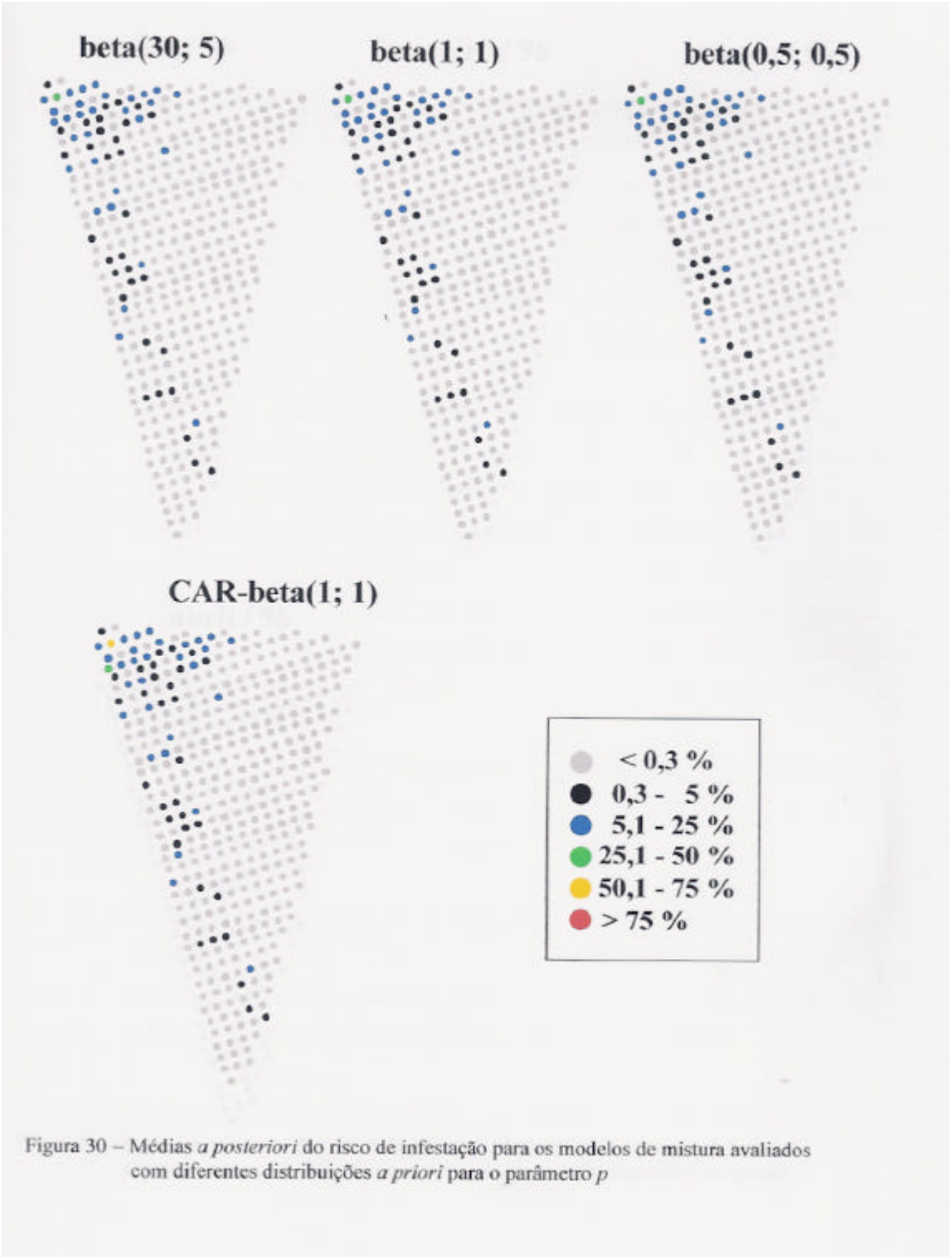

Figura 30 – Médias *a posteriori* do risco de infestação para os modelos de mistura avaliados com diferentes distribuições *a priori* para o parâmetro *p*

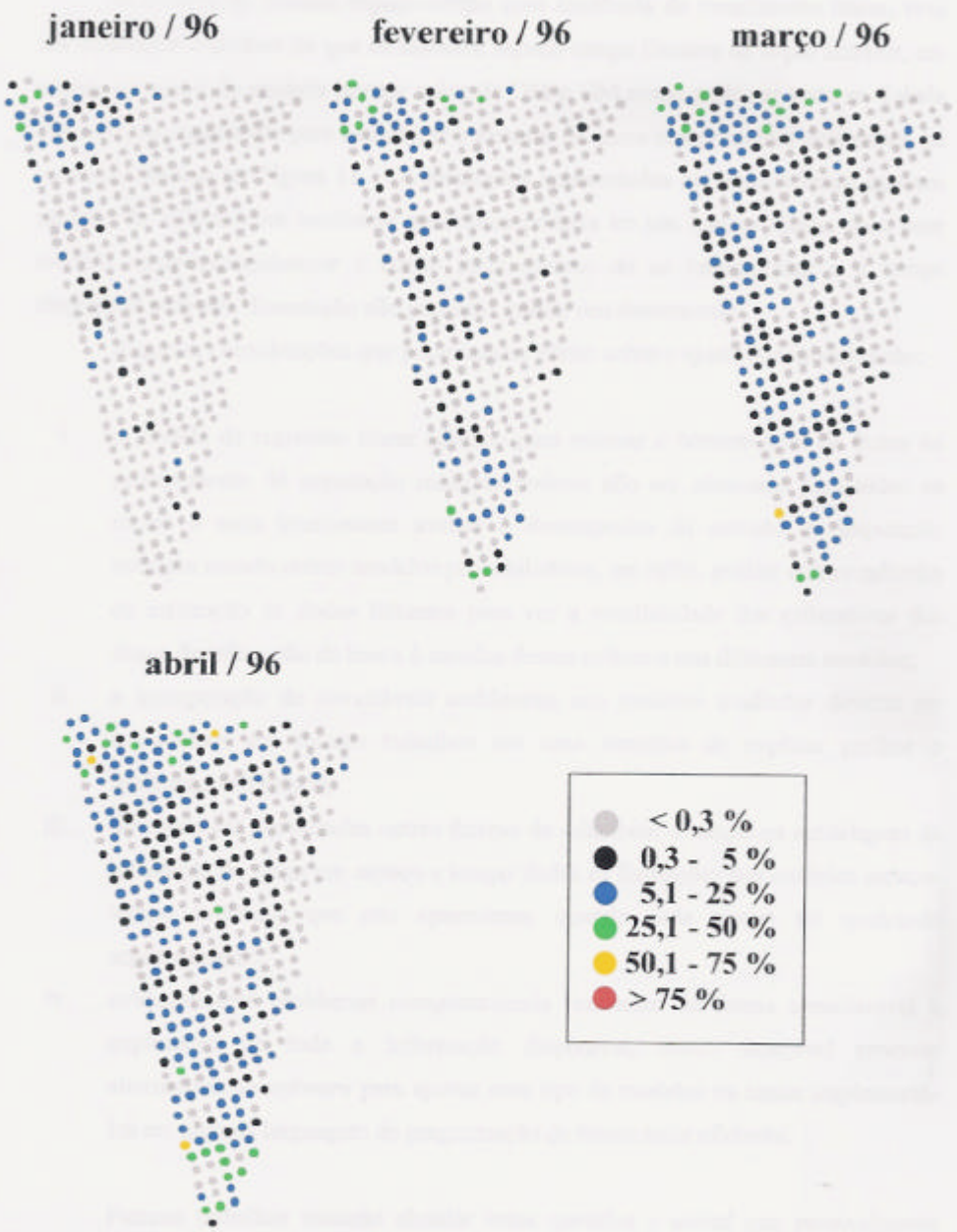

Figura 31 – Médias *a posteriori* dos riscos de infestação da broca para o modelo de mistura espaço-tempo com tendência de crescimento linear, sem dependência espacial.

O modelo de mistura espaço-tempo com tendência de crescimento linear, teve um desempenho melhor do que os modelos espaço-tempo lineares da seção anterior, em termos de ajuste do modelo (menor valor de DM e GM como pode ser visto na Tabela 10) e da sua capacidade para considerar o excesso de zeros nos primeiros dois meses tal como se observa na Figura 31. Os resultados apresentados até aqui indicam que um modelo de mistura com tendência quadrática poderia ter um melhor ajuste, mas, esse modelo computacionalmente é muito mais custoso de se implementar e o tempo disponível para esta dissertação não permitiu avaliar seu desempenho.

Algumas considerações que poderiam ter efeito sobre o ajuste dos modelos são:

i. o modelo de regressão linear adotado para estimar o número total de frutos no procedimento de imputação múltipla poderia não ser adequado para todos os meses e seria interessante avaliar o desempenho do método de imputação múltipla usando outros modelos probabilísticos, ou então, avaliar outros métodos de estimação de dados faltantes para ver a sensibilidade das estimativas dos riscos de infestação da broca à escolha desses métodos nos diferentes modelos;
ii. a incorporação de covariáveis ambientais nos modelos avaliados deveria ser considerada em futuros trabalhos em uma tentativa de explicar melhor o fenômeno sob estudo;
iii. deveriam ser pesquisadas outras formas de relacionar o tempo na modelagem da infestação da praga em espaço e tempo dadas as limitações dos modelos espaço-tempo avaliados que não apareceram quando cada tempo foi modelado separadamente;
iv. evidentemente problemas computacionais limitaram de forma considerável a exploração de toda a informação disponível, sendo desejável procurar alternativas de *software* para ajustar esse tipo de modelos ou tentar implementá-los em alguma linguagem de programação de forma mais eficiente.

Futuros trabalhos tentarão abordar estas questões e outras que possivelmente possam surgir de uma posterior análise dos resultados aqui apresentados.

5 CONCLUSÕES

Com base nos resultados apresentados foram obtidas as conclusões que se seguem.

A dispersão da infestação da broca do café no espaço e no tempo pode ser modelada adequadamente usando modelos hierárquicos Bayesianos que permitem incorporar facilmente efeitos aleatórios com e sem dependência espacial, além de covariáveis.

Em geral, tanto os modelos avaliados no espaço, como os modelos espaço-tempo foram pouco influenciados pela escolha de distribuições *a priori* para seus parâmetros e hiperparâmetros.

O uso de efeitos aleatórios espacialmente dependentes nos modelos permitiu identificar mais claramente tendências de crescimento ou decréscimo na infestação ao longo do tempo.

Modelos espaço-tempo com tendência de crescimento quadrática na infestação da broca tiveram um melhor ajuste aos dados do que modelos com tendência de crescimento linear.

Modelos de mistura, em particular os modelos inflacionados de zeros tiveram um melhor desempenho em termos de ajuste em relação aos modelos baseados em só uma distribuição padrão, principalmente no que se refere às estimativas no início da infestação.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSEN, M. Properties of some density-dependent integrodifference equation models. **Mathematical and Biological Sciences**, v.104, p.135-157, 1991.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, 2., Budapest: Akadémiai Kiadó, 1973. p.267-281.

ASSUNÇÃO, R.M. Estatística espacial com aplicações em epidemiologia, economia, sociologia. In: ESCOLA DE MODELOS DE REGRESSÃO, 7., São Carlos, 2001. **Minicurso**. São Carlos: UFSCar, Departamento de Estatística, 2001. 131p.

ASSUNÇÃO, R.M.; REIS, I.A.; OLIVEIRA, C.L. Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a bayesian space-time model. **Statistics in Medicine**, v.20, p.2319-2335, 2001.

ASSUNÇÃO, R.M.; POTTER, J.E.; CAVENAGHI, S.M. A bayesian space varying parameter model applied to estimating fertility schedules. **Statistics in Medicine**, v.21, 2002. /No prelo/

BAKER, P.S. Some aspects of the behaviour of the coffee berry borer in relation to its control in Southern Mexico. **Folia Entomológica Mexicana**, v.61, p.9-24, 1984.

BAKER, P.S. A sampling plan for a control project against the coffee berry borer (*Hypothenemus hampei*) in Mexico. **Tropical Pest Management**, v.35, p.169-172, 1989.

BAKER, P.S. **The coffee berry borer in Colombia**; final report of the DFID-Cenicafé-CABI Bioscience IPM Project (CNTR 93/1536A). Chinchiná: DFID-CENICAFÉ, 1999. 154p.

BAKER, P.S.; BARRERA, J.F. A field study of a population of coffee berry borer, *Hypothenemus hampei* (Coleoptera ; Scolytidae) in Chiapas, Mexico. **Tropical Agriculture**, v.70, p.351-355, 1993.

BAKER, P.S.; BARRERA, J.F.; VALENZUELA, J.E. The distribution of coffee berry borer (*Hypothenemus hampei*, Scolytidae) in Southern Mexico: A survey for a biocontrol project. **Tropical Pest Management**, v.35, p.163-168, 1989.

BARRERA, F.J. Dynamique des populations du scolyte des fruits du cafeier *Hypothenemus hampei* (Coleoptera: Scolytidae), et lutte biologique avec le parasitoide *Cephalonomia stephanoderis* (Hymenoptera: Bethylidae), au Chiapas, Mexique. Toulouse, 1992. 315p. Thesis (Ph.D.) - Universite Paul-Sabatier.

BERGAMIN, J. Contribuição para o conhecimento da biologia da broca do café *Hypothenemus hampei* (Ferr) (Coleoptera: Ipidae). **Arquivos do Instituto Biológico**, v.14, p.31-72, 1943.

BERGER, J.O.; PERICCHI, L.R. (1996) The intrinsic Bayes factor for linear models. In: BERNARDO, J.M.; BERGER, J.O.; DAVID, A.P.; SMITH, A.F.M. (Ed.) **Bayesian statistics *5***. London: Oxford University Press, 1996. p.25-44.

BERNARDINELLI, L.; MONTOMOLI, C. Empirical bayes versus fully bayesian analysis of geographical variation in disease risk. **Statistics in Medicine**, v.11, p.983-1007, 1992.

BERNARDINELLI, L.; CLAYTON, D.; MONTOMOLI, C. Bayesian estimates of disease maps: How important are priors?. **Statistics in Medicine**, v.14, p.2411-2431, 1995.

BERNARDINELLI, L.; CLAYTON, D.; PASCUTTO, C.; MONTOMOLI, C.; GHISLANDI, M. Bayesian analysis of space-time variation in disease risk. **Statistics in Medicine**, v.14, p.2433-2443, 1995b.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems (with discussion) **Journal of the Royal Statistical Society Series B**, v.36, p.192-236, 1974.

BESAG, J.; HIGDON, D. Bayesian analysis of agricultural field experiments. **Journal of the Royal Statistical Society Series B**, v.61, p.691-746, 1999.

BESAG, J.; KOOPERBERG, C. On conditional and intrinsic autoregressions. **Biometrika**, v.82, p.733-746, 1995.

BESAG, J.; YORK, J.C.; MOLLIÉ, A. Bayesian image restoration with two applications in spatial statistics (with discussion). **Annals of the Institute of Statistical Mathematics**, v.43, p.1-59, 1991.

BESAG, J.; GREEN, P; HIGDON, D.; MENGERSEN, K. Bayesian Computation and Stochastic Systems (with discussion). **Statistical Science**, v.10, p.3-66, 1995.

BEST, N.; COWLES, M.K.; VINES, K. **CODA: convergence diagnosis and output analisys software for Gibbs sampling output**; version 0.30. Cambridge: Cambridge University, 1996. 41p.

BEST, N.G.; ARNOLD, R.A.; THOMAS, A., WALLER, L.A.; CONLON, E.M. Bayesian models for spatially correlated disease and exposure data. In: BERNARDO, J.M.; BERGER, J.O.; DAVID, A.P.; SMITH, A.F.M. (Ed.) **Bayesian statistics 6**. London: Oxford University Press, 1999. p. 131-156.

BLISS, C.I. Statistical problems in estimating populations of japanese beetle larvae. **Journal of Economic Entomology**, v.34, p.221-232, 1941.

BORTH, P.W.; HUBERT, R.T. Modelling pink bollworm establishment and dispersion in cotton with the kriging technique. In: BELTWIDE COTTON PRODUCTION RESEARCH CONFERENCE, Dallas, 1987. **Proceedings**. Memphis: National Cotton Council of America, 1987. p.267-274.

BRESLOW, N.E.; CLAYTON, D.G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v.88, p.9-25, 1993.

BREWSTER, C.C.; ALLEN, J.C. Spatiotemporal model for studing insect dynamics in large-scale cropping systems. **Environmental Entomology**, v.26, p.473-482, 1997.

BUCKLAND, S.T.; ELSTON, D.A. Empirical models for the spatial distribution of wildlife. **Journal of Applied Ecology**, v.30, p.478-495, 1993.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs sampler. **The American Statistician**, v.46, p.167-174, 1992.

CHRISTENSEN, O.F., DIGGLE, P.J.; RIBEIRO Jr, P.J. Analysing positive-valued spatial data: the transformed Gaussian model. In: MONESTIEZ, P; ALLARD, D.; FROIDEVAUX, C. (Ed.) **GeoENV III - geostatistics for environmental applications**. Dordrecht: Kluwer Academic (Kluwer Series, 11), 2001. p.287-298.

CHRISTENSEN, O.F.; MOLLER, J.; WAAGEPETERSEN, R. **Analysis of spatial data using linear mixed models and Langevin-type Markov chain Monte Carlo**: technical report. Aalborg: Aalborg University, Department of Mathematical Sciences, 2000. 25p.

CLAYTON, D. Generalized linear mixed models. In: GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.) **Markov chain Monte Carlo in practice**. New York: Chapman and Hall, 1996. cap.16, p.275-301.

CLAYTON, D.; KALDOR, J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. **Biometrics**, v.43, p.671-681, 1987.

CLIFF, A.D.; ORD, J.K. **Spatial processes:** models & applications. London: Pion Press, 1981. 266p.

COHEN, A.C. A note on certain discrete mixed distributions. **Biometrics**, v.22, p.566-572, 1966.

CONGDON, P. **Bayesian Statistical Modelling**. New York: John Wiley, 2001. 531p.

CONLON, E.M.; WALLER, L.A. **Flexible Neighborhood Structures in Hierarchical Models for Disease Mapping**: research report 018. Minneapolis: University of Minnessota, School of Public Health, 1998. 25p.

COSTA, T.C.; VILLACORTA, A. Modelo acumulativo para *Hypothenemus hampei* (Ferrari, 1867) (Coleoptera : Scolytidae) com base em suas exigências térmicas. **Anais da Sociedade Entomológica do Brasil**, v.18, p.90-97, 1989.

COSTA, E.B.; SILVA, A.E.S. da; ANDRADE NETO, A.P.M. de; DAHER, F.A. **Manual técnico para a cultura do café no estado de Espírito Santo**. Vitória: SEAG, 1995. 63p.

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

CRESSIE, N.; CHAN, N.H. Spatial modeling of regional variables. **Journal of the American Statistical Association**, v.84, p.393-401, 1989.

CRESSIE, N.; HUANG, H.C. Classes of nonseparable, spatio-temporal stationary covariance functions. **Journal of the American Statistical Association**, v.94, p.1330-1340, 1999.

CRESSIE, N.; MUGGLIN, A.S. Spatio-temporal hierarchical modeling of an infectious disease from (simulated) count data. In: BETHLEHEM, J.G.; VANDER HEIJDEN, P. (Ed.) **Computational statistics**. Heilderberg: Physica-Verlag, 2000. p.41-52.

CURE, J.R.; SANTOS, R.H.S.; MORAES, J.C.; VILELA, E.F.; GUTIERREZ, A.P. (1998) Fenologia e Dinâmica Populacional de Broca do Café *Hypothenemus hampei* (Ferr.) Relacionadas às Fases de desenvolvimento do Fruto. **Anais da Sociedade Entomológica do Brasil**, v.27, p.325-335, 1998.

DALE, M.R. Concepts of spatial pattern analysis in plant ecology. In: DALE, M.R. (Ed.) **Pattern analysis in plant ecology**. Cambridge: Cambridge University Press, 1999. p.1-30.

DEAN, C.B.; MACNAB, Y.C. Modeling of rates over a hierarchical health administrative structure. **Canadian Journal of Statistics**, v.29, p.405-419, 2001.

DARNELL, J.S.; MEINKE, L.J.; YOUNG, L.J.; GOTWAY, C. Geostatistical investigation of the small-scale spatial variation of western corn rootworm (Coleoptera : Chrysomelidae) adults. **Environmental Entomology**, v.28, p.266-274, 1999.

DAVID, F.N.; MOORE, P.G. Notes on contagious distributions in plant populations. **Annals of Botany**, v.18, p.47-53, 1954.

DECAZY, B., OCHOA, H.; LOTODE, R. Indices de distribution spatiale et methode d'echantillonnage des populations du scolyte des drupes du cafeier *Hypothenemus hampei*. **Café Cacao Thé**, v.33, p.27-41, 1989.

DIGGLE, P.J.; RIBEIRO Jr., P.J. Model based geostatistics. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 14., Caxambú, 2000. **Minicurso**. São Paulo: Associação Brasileira de Estatística, 2000. 102p.

DIGGLE, P.J.; TAWN, J.A.; MOYEED, R.A. Model-based geostatistics. **Applied Statistics**, v.47, p.299-350, 1998.

FAHRMEIR, L.; LANG, S. Bayesian inference for generalized additive mixed models based on Markov random fields priors. **Applied Statistics**, v.50, p.201-220, 2001.

FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA. **Anexo estadístico del informe del LX congreso nacional de cafeteros**. http://www/cafedecolombia.com/LXcongreso/espanhol/principal_07122001.html (20 Mar. 2002)

GELFAND, A.E. Model determination using sampling-based methods. In: GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.) **Markov chain Monte Carlo in practice**. New York: Chapman and Hall, 1996. cap.9, p.145-161.

GELFAND, A.E.; GHOSH, S.K. Model choice: A minimum posterior preditive loss approach. **Biometrika**, v.85, p.1-11, 1998.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. **I.E.E.E. transactions of Pattern Analysis and Machine Intelligence**, v.6, p.721-741, 1984.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J.M.; BERGER, J.O.; DAVID, A.P.; SMITH, A.F.M. (Ed.) **Bayesian statistics 4**. Oxford: Clarendon Press, 1992. p.145-155.

GHOSH, S.K.; MUKHOPADHYAY, P.; LU, J.; CHEN, D. **Bayesian analysis of zero-inflated regression models**: technical report. Raleigh: North Caroline State University, Department of Statistics, 2001. 30p.

GIBSON, G.J. Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. **Applied Statistics**, v.46, p.215-233, 1997.

GIBSON, G.J.; AUSTIN, E.J. Fitting and testing spatio-temporal stochastic models with application in plant epidemiology. **Plant Pathology**, v.45, p.172-184, 1996.

GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.) **Markov chain Monte Carlo in practice**. New York: Chapman and Hall, 1996. 486p.

GREIG-SMITH, P. The use of random and contiguous quadrats in the study of the structure of plant communities. **Annals of Botany**, v.16, p.293-316, 1952.

GUTIERREZ, A.P.; VILLACORTA, A.; CURE, J.R.; ELLIS, C.K. Tritrophic Analysis of the Coffee (*Coffea arabica*) - Coffee Berry Borer [*Hypothenemus hampei* (Ferrari)] - Parasitoid System. **Anais da Sociedade Entomológica do Brasil**, v.27, p.357-385, 1998.

HALL, D.B. Zero-inflated Poisson and regression with randon effects: a case study. **Biometrics**, v.56, p.1030-1039, 2000.

HARGREAVES, H. Notes of the coffee berry borer (*Stephanoderis hampei*, Ferr.) in Uganda. **Bulletin of Entomological Research**, v.16, p.347-354, 1926.

HASTINGS, A. Spatial heterogeneity and ecological models. **Ecology**, v.71, p.426-428, 1990.

HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operations Research**, v.31, p.1109-1144, 1983.

HOLLAND, J.M.; PERRY, J.N.;WINDER,L. The within-field spatial and temporal distribution of arthropods in winter wheat. **Bulletin of Entomological Research**, v.89, p. 499-513, 1999.

IWAO, S. A new regression model for analysing the aggregation pattern of animal populations. **Researches on Population Ecology**, v.10, p.1-20, 1968.

KAISER, M.S.; CRESSIE, N.; LEE, J. Spatial mixture based on exponential family conditional distributions. **Statistica Sinica**, v.12, 2002. /No prelo/

KELSALL, J.E.; WAKEFIELD, J.C. Discussion of "Bayesian models for spatially correlated disease and exposure data", by Best et al. In: BERNARDO, J.M.;

BERGER, J.O.; DAVID, A.P.; SMITH, A.F.M. (Ed.) **Bayesian statistics 6**. London: Oxford University Press, 1999. p. 151.

KNORR-HELD, L. Bayesian Modelling of inseparable Space-time variation in disease risk. **Statistics in Medicine**, v.19, p.2555-2567, 2000.

KNORR-HELD, L.; BESAG, J. Modelling risk from a disease in time and space. **Statistics in Medicine**, v.17, p.2045-2060, 1998.

LACAYO, M.; GUHARAY, F. Tamaño óptimo de muestra para determinar el nivel de infestación de broca en los cafetales de la VI región de Nicaragua. In: AVANCES TÉCNICOS, Turrialba, 1992. Turrialba: MIP, 1992. p. 15-16.

LANDRUM, M.B.; BECKER, M.P. A multiple imputation strategy for incomplete longitudinal data. **Statistics in Medicine**, v.20, p.2741-276, 2001.

LE PELLEY, R.H. **The pests of coffee**. London: Longmans Green, 1968. 590p.

LIEBHOLD, A.M.; ZHANGg, X.; HOHN, M.E.; ELKINTON, J.S.; TICEHURST, M.; BENZON, G.L.; CAMPBELL, R.W. Geostatistical analysis of gypsy moth (Lepidoptera:Lymantriidae) eggs mass populations. **Environmental Entomology**, v.20, p.1407-1417, 1991.

LITTLE, R.J.A.; RUBIN, D.B. **Statistical Analysis with Missing Data**. New York: John Wiley, 1986. 278p.

LLOYD, M. Mean crowding. **Journal of Animal Ecology**, v.36, p.1-30, 1967.

MACNAB, Y.C.; DEAN, C.B. Parametric bootstrap and penalized quase-likelihood inference in conditional autoregressive models. **Statistics in Medicine**, v.19, p.2421-2435, 2000.

MACNAB, Y.C.; DEAN, C.B. Autoregressive Spatial Smoothing and Temporal Spline Smoothing for Mapping Rates. **Biometrics**, v.57, p. 949-956, 2001.

MATÉRN, B. **Spatial variation.** 2.ed. Berlin: Springer-Verlag, 1986. 151p.

MATHIEU, F.; BRUN, O.; FRÉROT, B. Factors related to native host abandonment by the coffee berry borer *Hypothenemus hampei* (Ferr.) (Col., Scolytidae). **Journal of Applied Entomology**, v.121, p.175-180, 1997.

McCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2.ed. London: Chapman and Hall, 1989. 511p.

MENDOZA, J.R.; GÓMEZ, J.O.; FERREIRA, E. Respuesta de la broca del café, *Hypothenemus hampei* a estimulos visuales y olfativos del fruto de café. **Sanidad Vegetal**, v.5, p.31-39, 1993.

MINISTERIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO. **Estatísticas agrícolas**. http://www.agricultura.gov.br/spa/pagespa/index.htm (20 Mar 2002).

MOLLIÉ, A. Bayesian mapping of disease. In: GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.) **Markov chain Monte Carlo in practice**. New York: Chapman and Hall, 1996. cap.20, p.359-379.

MOLLIÉ, A. Bayesian mapping of Hodgkin's disease in France. In: ELLIOTT, P.; WAKEFIELD, J.C.; BEST, N.G.; BRIGGS, D.J. (Ed.), **Spatial epidemiology methods and applications**, London: Oxford University Press, 2000. cap. 11, p.267-285

MORAN, P.A.P. The interpretation of statistical maps. **Journal of the Royal Statistical Society**, Series B, v.10, p.243-251, 1948.

MORISITA, M. I_δ-index, a measure of dispersion of individuals. **Researches on Population Ecology**, v.4, p.1-7, 1962.

MURPHY, S.T.; MOORE, D. Biological control of coffee berry borer, *Hypothenemus hampei* (Ferrari) (Coleoptera : Scolytidae): previous programmes and possibilities for the future. **Biocontrol News and Information**, v.11, p.107-117, 1990.

NESTEL, D.; KLEIN, M. Geostatistical analysis of leaf hopper (Homoptera : Cicadellidae) colonization and spread in deciduous orchads. **Environmental Entomology**, v.24, p.1032-1039, 1995.

O'HAGAN, A. Fractional Bayes factor for model comparison (with discussion). **Journal of the Royal Statistical Society Series B**, v.57, p.99-138, 1995.

OKUBO, A. **Diffusion and ecological problems: mathematical models.** Berlin: Springer, 1980. 475p.

PASCUTTO, C.; WAKEFIELD, J.C.; BEST, N.G.; RICHARDSON, S.; BERNARDINELLI, L.; STAINES, A.; ELLIOTT, P. Statistical issues in the analysis of disease mapping data. **Statistics in Medicine**, v.19, p.2493-2519, 2000.

PERRY, J.N. Spatial analysis by distance indices. **Journal of Animal Ecology**, v.64, p.303-314, 1995.

PERRY, J.N. Measures of spatial pattern for counts. **Ecology**, v.79, p.1008-1017, 1998.

PERRY, J.N.; DIXON, P. A new method for measuring spatial association in ecological count data. **Ecoscience**, v.9, 2002. /No prelo/

PERRY, J.N.; HEWITT, M. A new index of aggregation for animal counts. **Biometrics**, v.47, p.1505-1518, 1991.

PERRY, J.N.; GONZÁLEZ-ANDUJAR, J.L. (1993) A metapopulation neighbourhood model of an annual plant with a seedbank. **Journal of Ecology**, v.81, p.453-463, 1993.

PERRY, J.N.; WINDER, L.; HOLLAND, J.M.; ALSTON, R.D. Red blue plots for detecting clusters in count data. **Ecology Letters**, v.2, p.106-113, 1999.

PICKLE, L.W. Exploring spatio-temporal patterns of mortality using mixed effects models. **Statistics in Medicine**, v.19, p.2251-2263, 2000.

RAFTERY, A.L.; LEWIS, S. Comment: one long run with diagnostics: implemantation strategies for Markov chain Monte Carlo. **Statistical Science**, v.7, p.493-497, 1992.

REMOND, F., CILAS, C., VEGA-ROSALES, M.I. & GONZÁLEZ, M.O. Methodologie d'echantillonnage pour estimer les attaques des baies du cafeier par les scolytes (*Hypothenemus hampei* Ferr). **Café Cacao Thé**, v.37, p.35-51, 1993.

RICHARDSON, S.; MONFORT, C. Ecological correlation studies. In: ELLIOTT, P.; WAKEFIELD, J.C.; BEST, N.G.; BRIGGS, D.J. (Ed.), **Spatial epidemiology methods and applications**, London: Oxford University Press, 2000, cap. 11, p.205-220.

RICHARDSON, S.; MONFORT, C.; GREEN, M.; DRAPER, G.; MUIRHEAD, C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. **Statistics in Medicine**, v.14, p.2487-2501, 1995.

RIDOUT, M.S., DEMÉTRIO, C.G.B.; HINDE, J. Models for count data with many zeros. In: INTERNATIONAL BIOMETRIC CONFERENCE, 19. Cape Town: 1998. **Invited papers**, Cape Town, 1998, p.179-192.

ROBERT, C.P. Mixtures of distributions: inference and estimation. In: GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.), **Markov chain Monte Carlo in practice**, London: Chapman and Hall, 1996, cap. 24, p.441-464.

RUBIN, D.B. **Multiple imputation for nonresponse in surveys**. New York: John Wiley, 1987, 258p.

RUDD, W.G.; GANDOUR, R.W. Difussion model for insect dispersal. **Journal of Economic Entomology**, v.78, p.295-301, 1985.

RUIZ, C.R.; BAKER, P.S.; BUSTILLO, A.E. Efecto de la fenología del fruto de café sobre los parámetros de la tabla de vida de la broca del café *Hypothenemus hampei* Ferrari. In: CONGRESO DE LA SOCIEDAD COLOMBIANA DE ENTOMOLOGÍA, 23, Cartagena, 1996. p.56.

SÁNCHEZ Y RAMÍREZ, V. **Combate económicamente oportuno de la broca del grano de café**. Chiapas: Instituto Mexicano del café, Gerencia de Investigaciones Agrícolas, Dirección Adjunta de Producción y Mejoramiento de la Caficultura, 1984. 56p.

SCHOTZKO, D.J.; O'KEEFFE, L.E. Geostatistical description of the spatial distribution of *Lygus hesperus* (Heteroptera:Miridae) in lentils. **Journal of Economic Entomology**, v.82, p.1277-1288, 1989.

SCHOTZKO, D.J.; QUISENBERRY, S.S. Pea leaf weevil (Coleoptera:Curculionidae) spatial distribution in peas. **Environmental Entomology**, v.28, p.477-484, 1999.

SMYTH, G.K.; CHAKRABORTY, S.; CLARK, R.G.; PETIT, A.N. A stochastic model for anthracnose development in *Stylosanthes scabra*. **Phytopathology**, v.82, p.1267-1272, 1992.

SOKAL, R.R.; ODEN, N.L. Spatial autocorrelation in biology 1. Methodology. **Biological Journal of the Linnean Society**, v.10, p.199-228, 1978.

SPIEGELHALTER, D.J.; BEST, N.G.; CARLIN, B.P. **Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models**: technical report. Cambridge: MRC Biostatistics Unit, 1998. 31p.

SPIEGELHALTER, D.; THOMAS, A.; BEST, N. **WinBUGS**: version 1.3; user manual. Cambridge: Cambridge University. 2000. 35p.

SPIEGELHALTER, D.J.; BEST, N.G.; CARLIN, B.P.; VAN DER LINDE, A. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society Series B**, v.64, p.1-34, 2002.

SREEDHARAN, K.; BALAKRISHNAN, M.M.; PRAKASAM, C.B.; KRISHNAMOORTHY, B.P.; NAIDU, R. Bio-ecology and management of coffee berry borer. **Indian Coffee**, v.58, p.5-13, 1994.

SUN, D.; TSUTAKAWA, R.K.; KIM, H.; HE, Z. Spatio-temporal interaction with disease mapping. **Statistics in Medicine**, v.19, p.2015-2035, 2000.

TAYLOR, L.R. Aggregation, variance and mean. **Nature**, v.189, p.732-735, 1961.

THOMAS, C.F.G.; PARKINSON, L.; GRIFFITHS, G.J.K.; FERNANDEZ GARCIA, A.; MARSHALL, E.J.P. Aggregation and temporal stability of carabid beetle distributions in field and hedgerow habitats. **Journal of Applied Ecology**, v. 38, p.100-116, 2001.

TURECHEK, W.W.; MADDEN, L.V. Spatial pattern analysis of strawberry leaf blight in perennial production systems. **Phytopathology**, v. 89, p.421-433, 1999.

VAN DER LINDE, A.; WITZKO, K.H.; JÖCKEL, K.H. Spatial-temporal analysis of mortality using splines. **Biometrics**, v.51, p.1352-1360, 1995.

VIEIRA, A.M.C.; HINDE, J.; DEMÉTRIO, C.G.B. Zero-inflated proportion data models applied to a biological control assay. **Journal of Applied Statistics**, v.27, p.373-389, 2000.

WAKEFIELD, J.C.; BEST, N.G.; WALLER, L. Bayesian approaches to disease mapping. In: ELLIOTT, P.; WAKEFIELD, J.C.; BEST, N.G.; BRIGGS, D.J. (Ed.) **Spatial epidemiology methods and applications**. London: Oxford University Press, 2000. cap. 7, p.104-127.

WALLER, L.A.; CARLIN, B.P.; XIA, H. Structuring correlation within hierarchical spatio-temporal models for disesase rates. In: GREGORIE, T.G.; BRILLINGER, D.R.; DIGGLE, P.J.; RUSSEK-COHEN, E.; WARREN, W.G.; WOLFINGER, R.D. (Ed.) **Modelling bngitudinal and spatially correlated data**. Berlin: Springer, 1998. p.309-319.

WALLER, L.A.; CARLIN, B.P.; XIA, H.; GELFAND, A.E. Hierarchical spatio-temporal mapping of disease rates. **Journal of the American Statistical Association**, v.92, p.607-617, 1997.

WILLIAMS, L.D.; SCHOTZKO, D.J.; MCCAFFREY, J.P. Geostatistical description of the spatial distribution of *Limonius californicus* (Coleoptera:Elateridae) wire-worms in the Northwestern United States, with comments on sampling. **Environmental Entomology**, v.21, p.983-995, 1992.

WU, H.; WU, L. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. **Statistics in Medicine**, v.20, p.1755-1769, 2001.

XIA, H.; CARLIN, B.P. Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. **Statistics in Medicine**, v.17, p.2025-2043, 1998.

ZHOU, X.; ECKERT, G.J.; TIERNEY, W.M. Multiple imputation in public health research. **Statistics in Medicine**, v.20, p.1541-1549, 2001.

ZHU, L.; CARLIN, B.P. Comparing hierarchical models for spatio-temporallly misaligned data using the deviance information criterion. **Statistics in Medicine**, v.19, p.2265-2278, 2000.

APÊNDICE

Códigos em WinBUGS e S-Plus dos programas utilizados

***Programa* 1**. Código WinBUGS para gerar os valores imputados da variável número total de frutos usados para estimar os dados ausentes na seção 3.2.1

```
model
{
  for (i in 1:regions){
     for (t in 1:tempos){
      pop[i,t]~dpois(mu[i,t])
      log(mu[i,t])<- alpha[i] + beta[i]*t

      }
        alpha[i]~dnorm(4.6, 1.6)
        beta[i]~dnorm(0.1, 83)
    }

}
```

```
list(regions = 392, tempos = 10,

pop=structure(.Data=c(107, 105, 107, 103, 93, 101, 92,
94, 128, 183, NA, 10, 13, ................. , 38, 41,
NA, NA, NA, NA, 104, 217, 210, 210, ),.Dim=c(392,10)))
```

```
list(alpha=c(0, 0, ........................., 0, 0),

beta=c(0, 0, ................................., 0, 0),

pop=structure(.Data=c(NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, 104, NA, NA, ............................... , NA,
NA, 131, 153, 169, 175, NA, NA, NA, NA),.Dim=c(392,10)))
```

***Programa* 2**. Código WinBUGS para ajustar o modelo (1) da Tabela 1 (seção 3.2.2)

```
model       #  modelagem somente do espaco abr/96  -  imputacao 1

{
   b[1:regions] ~ car.normal(adj[], weights[], num[], taub)

for (i in 1:regions){
         O[i] ~dbin(pi[i],pop[i])
         logit(pi[i])<-  psi + g[i] + b[i]
         Orep[i] ~dbin(pi[i],pop[i])
         dif[i] <- Orep[i] - O[i]
         g[i] ~dnorm(0, taug)

         }


 for(k in 1:sumNumNeigh) {
                           weights[k] <-1
                          }


          psi  ~dflat()
          taub  ~dgamma(0.001, 0.001)
          sigmab <- 1 / sqrt(taub)
          taug  ~dgamma(0.001, 0.001)
          sigmag <- 1 / sqrt(taug)
          Gm <- inprod(dif[], dif[])
          media.b<-mean(b[])
          media.g<-mean(g[])
          desv.b<-sd(b[])
          desv.g<-sd(g[])

}



list(regions = 392, sumNumNeigh=2848,

pop=(.............................),
O=(.............................),
adj=c(.............................),
num=c(.............................))



list(taub = 1, psi = 0,  taug = 1,

g=c(0, 0, ....................., 0, 0),
b=c(0, 0, ....................., 0, 0),
Orep=(0, 0, ..................., 0, 0))
```

***Programa* 3**. Código S-Plus para gerar os vetores do esquema de vizinhança usado no modelo (11) da Tabela 1 (vizinhança baseada em distância com raio igual a 5 metros)

```
xcoor <- scan("c:\\temp\\xx.txt")
ycoor <- scan("c:\\temp\\yy.txt")
nareas <- length(xcoor)

# Cálculo da matriz de distâncias:
Dist <- matrix(0, ncol=nareas, nrow=nareas)
for (i in 1:nareas){
    Dist[i,] <- sqrt((xcoor[i]-xcoor)^2+(ycoor[i]-ycoor)^2)
}

#Cálculo da matriz de vizinhança com distância d=5 metros:
Vizd5 <- matrix(0, ncol=nareas, nrow=nareas)
d <- 5
for (i in 1:nareas){
for (k in 1:nareas){
if (Dist[i,k] < d) { Vizd5[i,k] <- 1 }
Vizd5[i,i] <- 0
}
}


#Cálculo do vetor com número de vizinhos de cada planta:
nvizd5 <- apply(Vizd5, 1, sum)

#Cálculo do vetor com o índice dos vizinhos de cada planta:

dimnames(Vizd5) <- list(c(1:nareas), c(1:nareas))
ivizd5 <- vector(mode="numeric", length=0)
for (i in 1:nareas) {
auxi <- as.numeric(dimnames(Vizd5)[[2]][Vizd5[i,]==1])
ivizd5 <- c(ivizd5,auxi)
}

#O mesmo procedimento foi usado para obter as vizinhanças
#para d=3, d=7 e d=10 metros.
```

***Programa* 4**. Código WinBUGS para ajustar o modelo (7) da Tabela 3 (seção 3.2.3)

```
model   #modelo 11 (quadrático)dependência espacial em delta e theta
        #Imputacao1

{
delta[1:regions] ~car.normal(vizinhos[], weights[], num[], tau.delta)
theta[1:regions] ~car.normal(vizinhos[], weights[], num[], tau.theta)

 for (i in 1:regions){
   for (t in 1:tempos){
      O[i,t] ~dbin(pi[i,t],pop[i,t])
      logit(pi[i,t])<- psi + eta*pow(t,2) + gama[i] + delta[i]*t +
      theta[i]*pow(t,2)
      Orep[i,t] ~dbin(pi[i,t],pop[i,t])

      }
       gama[i]~dnorm(0, tau.gama)
 }


 for(k in 1:sumNumNeigh)  {
                           weights[k] <-1
                         }


          psi ~dflat( )
          eta ~dflat( )
          tau.gama ~dgamma(0.001, 0.001)
          sigma.gama <- 1 / sqrt(tau.gama)
          tau.delta  ~dgamma(0.001, 0.001)
          sigma.delta <- 1 / sqrt(tau.delta)
          tau.theta  ~dgamma(0.001, 0.001)
          sigma.theta <- 1 / sqrt(tau.theta)
}


list(regions = 392,  tempos = 4,  sumNumNeigh = 2848,

pop=structure(.Data=c(.....................),.Dim=c(392,4)),
O=structure(.Data=c(.....................),.Dim=c(392,4)),
vizinhos=c(0, 0, ....................................., 0, 0),
num=c(0, 0, ....................................., 0, 0))



list(tau.gama=1, tau.delta=1, tau.theta=1, psi=0, eta=0,

gama=c(0, 0, ....................................., 0, 0),
delta=c(0, 0, ....................................., 0, 0),
theta=c(0, 0, ....................................., 0, 0),
Orep=structure(.Data=c(0, 0, ........., 0, 0),.Dim=c(392,4)))
```

***Programa* 5**. Código WinBUGS para ajustar o primeiro modelo de mistura descrito na seção 3.2.4

```
model       #  modelo de mistura com priori beta(30, 5)
            #  janeiro/96  *** Imputacao1 ***
{

  for (i in 1:N){
        O[i] ~dbin(pi[i],pop[i])
        state[i]~dbern(theta)
        state1[i] <- state[i] + 1
        pi[i]<-prop[i, state1[i]]
        logit(prop[i,1])<- delta[i]
        prop[i,2]<- 0
        delta[i]~dnorm(mu,tau)
        Orep[i] ~dbin(pi[i],pop[i])
        dif[i] <- Orep[i] - O[i]
        }

        Gm <- inprod(dif[], dif[])
        media.delta<- mean(delta[])
        desv.delta<- sd(delta[])
        theta ~dbeta(a, b)
        mu ~dnorm(0, 1.0E-6)
        tau ~dgamma(0.001, 0.001)
        sigma.tau <- 1 / sqrt(tau)

}


list(N = 392, a=30, b=5,

pop=c(.............................),

O=c(.............................))


list(tau=1, mu=0, theta=0.5,

delta=c(0, 0, ................., 0, 0),

state=c(0, 0, ................., 0, 0),

Orep=c(0, 0, ................., 0, 0))
```

***Programa* 6**. Código WinBUGS para ajustar o segundo modelo de mistura descrito na seção 3.2.4 (mistura de duas distribuições normais)

```
model             #modelo de mistura 2  (janeiro/96)  - Imputacao7
{

  for (i in 1:N){
        O[i] ~dbin(pi[i],pop[i])
        logit(pi[i]) <- delta[i]
        delta[i] ~dnorm(mu[i], tau[i])
         mu[i] <- lambda[T[i]]
        tau[i] <- prec[T[i]]
        T[i] ~ dcat(P[])

        }

        P[1] ~dbeta(a, b)
        [2] <- 1 - P[1]
        lambda[1] ~dnorm(0, 1.0E-6)
        lambda[2] <- lambda[1] + theta
        theta ~dnorm(0, 1.0E-6) I(0,  )
        prec[1] ~dgamma(0.001, 0.001)
        prec[2] ~dgamma(0.001, 0.001)

}



list(N = 392, a = 1, b = 1,

pop=c(.....................................................),
O=c(.....................................................),
T=c(1, NA, NA, NA, NA, ....................., NA, NA, NA, NA, 2))




list(lambda=c(0.00001, NA),  prec=c(1, 1), theta=0.5, P=c(0.5, NA),

delta=c(0, 0, ........................................., 0, 0),
T=c(NA, 1, 1, 1, 1, ........................., 1, 1, 1, 1, NA))
```

***Programa* 7.** Código Win BUGS para ajustar o modelo de mistura espaço-tempo descrito na seção 3.2.4

```
model   #  modelo de mistura espaco-tempo  (jan/96 - abr/96)
        #  ***Imputacao5  ***

{

  for (i in 1:N){
        for (t in 1:tempos){
                O[i,t] ~dbin(pi[i,t],pop[i,t])
                state[i,t]~dbern(theta[t])
                state1[i,t] <- state[i,t] + 1
                pi[i,t]<-prop[i,t,state1[i,t]]
                logit(prop[i,t,1])<- gama[i] + delta[i]*t
                prop[i,t,2]<-0
                Orep[i,t] ~dbin(pi[i,t],pop[i,t])
                }

                gama[i]~dnorm(mu.gama,tau.gama)
                delta[i]~dnorm(mu.delta,tau.delta)

                }

  for (t in 1:tempos){
                    theta[t] ~dbeta(a[t], b[t])
                    }


  mu.gama ~dnorm(0.0, 1.0E-6)
  mu.delta ~dnorm(0.0, 1.0E-6)
  tau.gama ~dgamma(0.001, 0.001)
  tau.delta ~dgamma(0.001, 0.001)
  sigma.gama <- 1 / sqrt(tau.gama)
  sigma.delta <- 1 / sqrt(tau.delta)

}


list(N = 392, tempos=4,  a=c(1,1,1,1), b=c(1,1,1,1),

pop=structure(.Data=c(.....),.Dim=c(392,4)),
O=structure(.Data=c(.....),.Dim=c(392,4)))



list(tau.gama=1,  tau.delta=1,  mu.gama=0,  mu.delta=0,  theta=c(0.5, 0.5, 0.5, 0.5),

gama=c(0, 0,  ..................................   0, 0),
delta=c(0, 0,  ..................................   0, 0),
state=structure(.Data=c(0, 0, ..... 0, 0),.Dim=c(392,4)),
Orep=structure(.Data=c(0, 0, ..... 0, 0),.Dim=c(392,4)))
```

***Programa* 7.** Código Win BUGS para ajustar o modelo de mistura espaço-tempo descrito na seção 3.2.4

```
model   #  modelo de mistura espaco-tempo  (jan/96 - abr/96)
        #  ***Imputacao5  ***

{

  for (i in 1:N){
        for (t in 1:tempos){
                O[i,t] ~dbin(pi[i,t],pop[i,t])
                state[i,t]~dbern(theta[t])
                state1[i,t] <- state[i,t] + 1
                pi[i,t]<-prop[i,t,state1[i,t]]
                logit(prop[i,t,1])<- gama[i] + delta[i]*t
                prop[i,t,2]<-0
                Orep[i,t] ~dbin(pi[i,t],pop[i,t])
                }

                gama[i]~dnorm(mu.gama,tau.gama)
                delta[i]~dnorm(mu.delta,tau.delta)

                }

  for (t in 1:tempos){
                    theta[t] ~dbeta(a[t], b[t])
                    }


  mu.gama ~dnorm(0.0, 1.0E-6)
  mu.delta ~dnorm(0.0, 1.0E-6)
  tau.gama ~dgamma(0.001, 0.001)
  tau.delta ~dgamma(0.001, 0.001)
  sigma.gama <- 1 / sqrt(tau.gama)
  sigma.delta <- 1 / sqrt(tau.delta)

}


list(N = 392, tempos=4,  a=c(1,1,1,1), b=c(1,1,1,1),

pop=structure(.Data=c(.............),.Dim=c(392,4)),
O=structure(.Data=c(.............),.Dim=c(392,4)))



list(tau.gama=1,  tau.delta=1,  mu.gama=0,  mu.delta=0,  theta=c(0.5, 0.5, 0.5, 0.5),

gama=c(0, 0,  ....................................   0, 0),
delta=c(0, 0,  ....................................   0, 0),
state=structure(.Data=c(0, 0, ..... 0, 0),.Dim=c(392,4)),
Orep=structure(.Data=c(0, 0, ..... 0, 0),.Dim=c(392,4)))
```