

Modelos hierárquicos bayesianos para estudar a distribuição espacial da infestação da broca do café em nível local

RAMIRO RUIZ C.*
CLARICE G. B. DEMÉTRIO †
RENATO M. ASSUNÇÃO ‡
ROSELI A. LEANDRO §

Resumo

Estudar a distribuição espacial de pragas em sistemas agrícolas pode fornecer informação importante sobre os mecanismos de dispersão das espécies e sua interação com fatores ambientais, sendo útil também no desenvolvimento de planos de amostragem, na otimização de programas de manejo integrado de pragas e no planejamento de experimentos. Neste trabalho foram comparados vários modelos para estudar a variação espacial da infestação da broca do café visando descrever adequadamente a dispersão da infestação da praga e identificar áreas de crescimento ou decréscimo na infestação. Foram usadas diferentes combinações de efeitos aleatórios representando variabilidade não estruturada e estruturada espacialmente. Foram também avaliados diferentes esquemas de vizinhança para representar a correlação espacial dos dados. Adicionalmente foram testados modelos de mistura para levar em consideração o excesso de zeros no início da infestação. O ajuste dos modelos foi feito usando métodos MCMC. Os resultados são apresentados como uma seqüência de mapas de risco de infestação.

*Departamento de Ciências Exatas, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo. Piracicaba - Brasil.

†Departamento de Ciências Exatas, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo. Piracicaba - Brasil. email: clarice@carpa.ciagri.usp.br

‡Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte - Brasil.

§Departamento de Ciências Exatas, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo. Piracicaba - Brasil.

Palavras chave: métodos MCMC, mapas de risco, modelos de mistura, modelos inflacionados de zeros.

Abstract

Studying the spatial distribution of agricultural pests can provide 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. This work compared several models for studying the spatial variation of the coffee berry borer infestation in order to produce risk maps and identify areas of low/high levels of infestation. Firstly spatial analysis was carried out using different combinations of random effects representing spatially structured and unstructured variability. Also different neighborhood schemes were used to represent the spatial correlation of the data. Mixture models allowing for the excess of zeros in the first months were also considered. The model fitting was done using MCMC methods. The results are presented as a sequence of risk maps.

Keywords: Markov chain Monte Carlo, risk maps, mixture models, zero inflated models.

1 Introdução

Descrições detalhadas da distribuição espacial de populações de pragas em culturas comerciais são de grande importância na otimização do uso de agentes de controle em programas de manejo integrado, no desenvolvimento de planos de amostragem e no planejamento de experimentos no campo dentre outras aplicações. Porém, o padrão espacial de populações de insetos tem sido pouco pesquisado, em parte pelo esforço intensivo de amostragem que é 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 em relações variância/média (David & Moore, 1954; Taylor, 1961; Morisita, 1962; Lloyd, 1967; Iwao, 1968, dentre outros). Estes índices, porém, ignoram a localização espacial das amostras, 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, esses métodos são 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 geográficas das unidades de amostragem e requerem que a amostragem seja feita a intervalos de espaço regulares. Além disso, há uma perda de informação espacial ao passar de dados pontuais para contagens por quadrats. Tem sido comum também o uso de métodos convencionais de Geoestatística para caracterizar padrões espaciais de insetos (Schotzko & O’Keeffe, 1989; Liebhold et al., 1991; Williams et al., 1992; Nestel & Klein, 1995; Darnell et al., 1999; Schotzko & Quisenberry, 1999). Entretanto, Perry (1998) sustenta que as contagens de indivíduos de uma espécie animal ou vegetal em particular, por serem discretas, distribuídas freqüentemente em aglomerados e com uma grande quantidade de valores iguais a zero, podem não ter a estrutura de covariância espacial estável assumida pelos métodos geoestatísticos, desenvolvidos originalmente para variáveis medidas em uma escala contínua.

A broca do café *Hypothenemus hampei* Ferrari (Coleoptera: Scolytidae) tem sido descrita como a praga mais importante da cafeicultura no mundo. 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). Alguns aspectos do padrão de dispersão e estratégias de amostragem foram estudados para este inseto em vários países, mostrando que a praga tem um padrão de distribuição agregado no campo. 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. O presente trabalho pretende estudar aspectos da variação espacial em pequena escala da infestação da broca do café em condições de campo na Colômbia, usando modelos estatísticos que descrevam adequadamente a dispersão da praga em um lote de café em estágios iniciais de infestação, e construir mapas de risco de infestação da praga que permitam identificar áreas de crescimento ou decréscimo da infestação nos diferentes meses. Esses resultados para a fase inicial da infestação fazem parte de uma pesquisa mais abrangente, ainda em andamento, para modelar o processo de infestação no espaço e no tempo simultaneamente.

A seção 2 descreve em detalhe a metodologia utilizada e os diferentes modelos considerados. A seção 3 apresenta e discute os resultados obtidos à luz dos diferentes modelos avaliados e sugere algumas considerações que poderiam melhorar o ajuste dos modelos propostos. Na seção 4 são apresentadas as conclusões e recomendações para futuros trabalhos.

2 Metodologia

Este trabalho analisou os dez primeiros meses de avaliação de um levantamento da infestação da broca, nos quais a praga se dispersou a partir de focos iniciais até colonizar quase a totalidade de uma área experimental de 2214 plantas de café (*Coffea arabica* var. Colômbia), distribuídas em uma 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. Esses dados fazem parte de um experimento sobre a dinâmica populacional e desenvolvimento de técnicas de amostragem para a broca do café, realizado pelo “Centro Nacional de Investigaciones de Café”, CENICAFE, Colômbia. 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. Nenhum método de controle da praga foi realizado durante este período além da colheita permanente de frutos maduros, sobremaduros e secos. A informação considerada para análise começou a ser obtida mensalmente a partir de julho de 1995 (três meses depois do registro da primeira floração importante na cultura) até abril de 1996. Iniciou-se com uma inspeção 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. Isso gera um problema estatístico de dados faltantes para a estimação do modelo e é abordado na próxima sub-seção. A localização (coordenadas $X - Y$) das 2214 plantas do lote foi referenciada num plano cartesiano a partir de uma origem arbitrária. Devido à extensão e à complexidade do tema em estudo e a sua alta demanda computacional, este trabalho considerou unicamente a análise de uma subárea de 392 plantas dentre as 2214 disponíveis. Essa subárea, porém, é representativa do que aconteceu na parcela completa nesse período.

Do ponto de vista da Estatística Espacial, o problema foi abordado como um problema de dados de área, fazendo uma analogia entre os métodos usados na área de epidemiologia médica para modelar a variação geográfica de taxas de doenças e a distribuição espaço-temporal de pragas e doenças de frutos em sistemas agrícolas perenes. Assim, cada planta pode ser considerada como equivalente a uma pequena área ou distrito, com o número total de frutos

dessa planta sendo equivalente à população sob risco nessa área, enquanto que o número de frutos afetados pela praga é equivalente ao número de casos da doença. As vantagens e desvantagens da implementação dessa classe de modelos na situação anteriormente descrita são consideradas ao final do artigo. Uma abordagem hierárquica Bayesiana foi adotada para a estimação dos parâmetros de todos os modelos considerados neste trabalho.

2.1 Estimação dos dados faltantes

Como foi mencionado anteriormente, não se contava com a informação sobre o número total de frutos das plantas não infestadas, e este teve que ser estimado. Para isto, foi usado o método de imputação múltipla (Rubin, 1987). Este método está baseado na substituição de cada valor ausente ou deficiente por $m \geq 2$ valores aceitáveis, representando uma distribuição de possibilidades que descreve a incerteza sobre o verdadeiro valor que está faltando. 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. A escolha deste método foi devido a sua simplicidade de implementação e ao fato de permitir que a estimação dos dados ausentes seja feita separadamente da modelagem do risco de infestação, sendo possível, depois, o uso de métodos padrões 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.

A 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 (ver Figura 1). Além disso, constatou-se uma ausência de dependência espacial significativa entre os totais de frutos de cada planta, determinada mediante o cálculo de um índice de autocorrelação de Moran (Moran, 1948) nos meses de março e abril de 1996, já que nesses meses existia informação do total de frutos para 58% e 70% das plantas respectivamente.

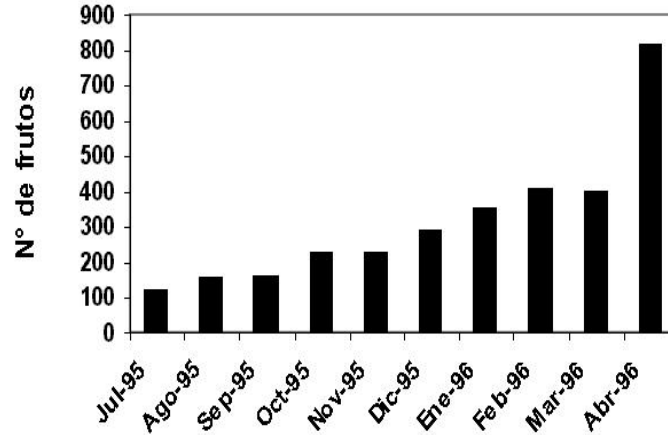
Portanto, para a modelagem das contagens N_{it} de frutos da árvore i no mês t , foi adotado um modelo de crescimento da forma

$$N_{it} \sim \text{Poisson}(\mu_{it})$$

com

$$\text{Log}(\mu_{it}) = \alpha_i + \beta_i t \quad i = 1, \dots, 392; \quad t = 1, \dots, 10. \quad (1)$$

Figura 1: 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.



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. Condiçoadas nos valores de μ_{it} , as contagens de frutos foram supostas independentes. Além disso, foi admitido que, *a priori*,

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

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 valores para os parâmetros destas prioris representam valores plausíveis que forneceriam 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. 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 ($3,2 = 4\sigma$), 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 .

Usando o teorema de Bayes, é possível usar os dados observados para atualizar o conhecimento sobre o vetor de parâmetros composto pelos α 's e β 's, bem como pelos dados faltantes. Esta atualização é expressa pela distribuição de probabilidade *a posteriori* conjunta dada por:

$$P(\alpha, \beta, n_{ausentes} | n_{observados}) \propto \prod_{i=1}^{392} \prod_{t=1}^{10} \left\{ \frac{\mu_{it}^{n_{it}} \exp(-\mu_{it})}{n_{it}!} \right\} \exp \left\{ - \sum_{i=1}^{392} (\alpha_i - 4.6)^2 \right\} \\ \exp \left\{ - \sum_{i=1}^{392} (\beta_i - 0.1)^2 \right\}.$$

Esta atualização é feita numericamente por meio de métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC). Esses métodos geram uma amostra da distribuição *a posteriori* dos parâmetros de interesse e foram 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 cinco) 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).

O modelo (1) 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 a incerteza sobre os valores ausentes a serem estimados.

2.2 Análise espacial

O modelo básico para os dados assume que n_i e y_i , são respectivamente, o número total observado de frutos e o número observado de frutos infestados pela broca na planta i , $i = 1, \dots, 392$, e que o número de frutos infestados Y_i segue uma distribuição binomial com parâmetros n_i e π_i . A probabilidade de risco de infestação desconhecida, π_i , foi modelada em um contexto de modelos

lineares generalizados mistos com uma função de ligação logística e um preditor linear η_i que se decompõe aditivamente em efeitos fixos e aleatórios segundo o modelo específico que esteja sendo usado.

Para a análise espacial foi escolhido inicialmente o mês de março/96 sobre o qual foi feita uma análise para verificar o quanto diferem as estimativas *a posteriori* do risco de infestação em relação à escolha de diferentes esquemas de vizinhança e diferentes distribuições *a priori* para os parâmetros e os hiperparâmetros do modelo básico

$$Y_i | \pi_i \sim \text{Binomial}(n_i, \pi_i)$$

com

$$\text{logit}(\pi_i) = \eta_i = \xi + \theta_i + \phi_i, \quad i = 1, \dots, 392 \quad (2)$$

sendo que o intercepto ξ tem uma distribuição *a priori* uniforme $(-\infty, +\infty)$ e representa o logaritmo do nível geral do risco relativo de infestação na área experimental. O modelo considera efeitos aleatórios para cada área, como definidos em Besag et al (1991), constituídos pela soma de um componente de heterogeneidade não estruturada, $\phi_i \sim \text{normal}(0, \tau_\phi)$ e um componente espacialmente estruturado, θ_i com uma *priori* condicional autoregressiva Gaussiana (CAR) dada pela expressão $\theta_i | \theta_{-i} \sim N(\bar{\theta}_{(i)}, \sigma_\theta^2 / r_i)$, sendo que θ_{-i} representa o grupo de todos os θ 's excluindo θ_i ; $\bar{\theta}_{(i)} = \sum_{j \in \partial_i} \theta_j / r_i$ é a média dos θ 's na vizinhança de cada planta i e r_i é o número desses vizinhos. Estes dois tipos de efeitos representam características não especificadas da planta i que possuem (θ) e não possuem (ϕ) estrutura espacial e que podem ser interpretados como substitutos de covariáveis não medidas. Foram avaliadas, também, diferentes combinações desse modelo, incluindo somente efeitos aleatórios não estruturados, estruturados espacialmente e a combinação de ambos os efeitos. Os modelos avaliados estão resumidos na Tabela 1.

Para estudar a influência da escolha das *prioris* para os hiperparâmetros τ_θ e τ_ϕ 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 τ_θ e τ_ϕ . As distribuições escolhidas, porém, não são muito informativas, refletindo a incerteza sobre esses parâmetros.

Dado que as plantas encontravam-se regularmente espaçadas na área experimental, foram inicialmente considerados esquemas de vizinhança definidos para reticulados 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). No primeiro caso, a vizinhança

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) $\eta_i = \alpha + \theta_i + \phi_i$	2 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(2) $\eta_i = \alpha + \theta_i + \phi_i$	2 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,5; 0,0005)$	$\Gamma(0,5; 0,0005)$
(3) $\eta_i = \alpha + \theta_i$	2 ^a ordem	-	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	-
(4) $\eta_i = \phi_i$	2 ^a ordem	$N(0, \tau_\phi)$	-	-	$\Gamma(0,001; 0,001)$
(5) $\eta_i = \alpha + \theta_i + \phi_i$	2 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,25; 0,005)$	$\Gamma(0,25; 0,005)$
(6) $\eta_i = \alpha + \theta_i + \phi_i$	2 ^a ordem	$t(0, \tau_\phi, 3)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(7) $\eta_i = \alpha + \theta_i + \phi_i$	2 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,01; 0,001)$	$\Gamma(0,25; 0,005)$
(8) $\eta_i = \alpha + \theta_i + \phi_i$	1 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(9) $\eta_i = \alpha + \theta_i + \phi_i$	4 ^a ordem	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(10) $\eta_i = \alpha + \theta_i + \phi_i$	3 metros	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(11) $\eta_i = \alpha + \theta_i + \phi_i$	5 metros	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(12) $\eta_i = \alpha + \theta_i + \phi_i$	7 metros	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$
(13) $\eta_i = \alpha + \theta_i + \phi_i$	10 metros	$N(0, \tau_\phi)$	$CAR(\tau_\theta)$	$\Gamma(0,001; 0,001)$	$\Gamma(0,001; 0,001)$

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 circunferências com raios de 3, 5, 7 e 10 metros. Estes sistemas de vizinhança foram usados na especificação das auto-regressões condicionais Gaussianas que constituem as distribuições *a priori* para os efeitos aleatórios espacialmente estruturados θ do modelo (2). Assim, a distribuição *a posteriori* conjunta para o modelo (2) fica dada por

$$P(\theta, \phi, \tau_\theta, \tau_\phi | y) \propto \prod_{i=1}^{392} \left\{ \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \right\} \tau_\phi^{-\frac{392}{2}} \exp \left\{ -\frac{\tau_\phi}{2} \sum_{i=1}^{392} (\phi_i - \bar{\phi})^2 \right\} \\ \tau_\theta^{-\frac{392}{2}} \exp \left\{ -\frac{\tau_\theta}{2} \sum_{i \sim j} (\theta_i - \theta_j)^2 \right\} p(\tau_\theta) p(\tau_\phi).$$

sendo que $i \sim j$ 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 cinco) 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 (1). Os diferentes modelos foram comparados com base na estatística *DM* de Gelfand & Ghosh (1998).

Esta estatística é a soma de dois componentes, um representando o grau de ajuste aos dados (GM) e outro representando a complexidade do modelo e que atua como um termo de penalidade (PM). Quanto menor o valor de $DM = PM + GM$, melhor é o ajuste do modelo aos dados, de acordo com este critério.

Como a análise apresentada para os dados de março de 1996 mostrou que o modelo (3) foi o mais apropriado, ele foi ajustado separadamente para cada um dos outros nove meses. Isto permite estudar a variação dos dois parâmetros através do tempo e sugere possíveis estratégias de modelagem espaço-tempo.

2.3 Modelo 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. Seja Y a variável aleatória *número de frutos com broca em n frutos*, com observações (y_i, n_i) , $i = 1, 2, \dots, 392$, 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 está infestada pela broca ($y_i = 0$) ou $Z_i = 0$ se a planta i tem algum grau de infestação ($y_i = 1, 2, \dots, n_i$), tal que $Z_i \sim \text{Bernoulli}(p)$. Assim, o modelo de mistura considera que no início da infestação 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 π_i . Desse modo, dizemos que Y tem uma distribuição binomial inflacionada de zeros, dada por:

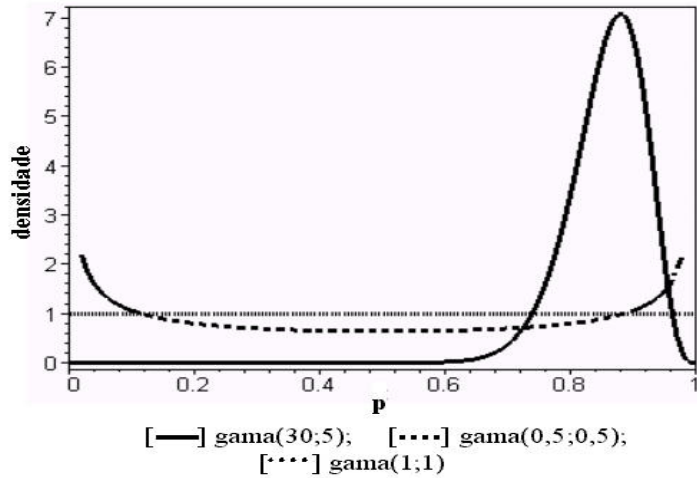
$$Pr(Y_i = y_i) = \begin{cases} p + (1 - p)(1 - \pi_i)^{n_i}, & y_i = 0 \\ (1 - p) \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} & y_i = 1, 2, \dots, n_i. \end{cases}$$

com $0 \leq p < 1$. De forma semelhante aos modelos anteriores, foi usada uma função de ligação logística para a binomial, com $\text{logit}(\pi_i) = \delta_i$. O parâmetro do preditor linear δ_i foi modelado, inicialmente, sem considerar dependência espacial *a priori*. Assim, foi assumido que $\delta_i \sim \text{normal}(\mu, \tau)$ com *hiperprioris* de parâmetros conhecidos $\mu \sim \text{normal}(0; 1, 0E - 6)$ e $\tau \sim \text{gama}(0, 001; 0, 001)$, 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, $\text{logit}(\pi_i) = \xi + \gamma_i$ em que $\xi \sim$

$uniforme(-\infty, +\infty)$ e $\gamma_i \sim CAR(\tau)$, 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 (Figura 2).

Figura 2: Distribuições a priori para o parâmetro p nos modelos de mistura.



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, 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 $\mu = a/(a + b)$ e $\sigma^2 = \mu(1 - \mu)/(a + b + 1)$. Assim, os parâmetros a e b podem ser calculados a partir das expressões: $a = \mu[\mu(1 - \mu)/\sigma^2 - 1]$ e $b = a(1 - \mu)/\mu$, dando os valores da *priori* informativa $a = 30$ e $b = 5$.

Para o ajuste de cada um desses modelos de mistura, foi gerada uma cadeia de 45000 iterações das quais foram descartadas as primeiras 5000 e guardada 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.

3 Resultados e discussão

Todos os modelos propostos na metodologia, foram satisfatoriamente ajustados usando o software WinBUGS 1.3 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.

3.1 Análise espacial

A partir do ajuste dos modelos, foi possível obter estimativas *a posteriori* das variâncias marginais para os efeitos aleatórios não estruturados (ϕ) e estruturados espacialmente (θ), as quais são apresentadas na Tabela 2 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* das variâncias marginais empíricas

Tabela 2: Estimativas pontual e por intervalo da variância marginal empírica para os efeitos aleatórios θ e ϕ e razão entre as estimativas das variâncias.

Modelo	S_{θ}^2 (estruturada espacialmente)			S_{ϕ}^2 (não estruturada)			Razão S_{θ}^2/S_{ϕ}^2
	Média	IC 95%		Média	IC 95%		
		LI	LS		LI	LS	
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	1,85	5,399	1,35
6	5,794	3,667	8,91	5,148	2,271	10,963	1,125
7	5,077	2,999	8,497	3,682	1,853	5,37	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	5,27	1,482
11	6,48	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	0,38	3,845	4,935

para os efeitos aleatórios não estruturados, ϕ , e estruturados espacialmente, θ , foi maior do que um para todos os modelos avaliados em março/96, indicando um domínio da variabilidade espacialmente estruturada nos modelos. Em particular, esta dominância foi mais forte à medida que a ordem ou o comprimento do raio do esquema de vizinhança aumentava. 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 (Figura 3). 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. 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 (resultados não mostrados). Foi

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

Em todos os modelos avaliados houve algumas plantas não infestadas pela broca que, foram classificadas como tendo algum grau de infestação (Tabela 3). 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 com distribuição *a priori* t sobre ϕ que teve 28 plantas passando da categoria 1 para a categoria 2. Por outro lado, o modelo somente com efeito aleatório espacial teve 25 plantas que passaram da categoria 1 para a 2, enquanto que no modelo sem dependência espacial nos riscos de infestação, somente 5 plantas mudaram de categoria 1 para a 2. Isto sugere que o modelo binomial padrão não é apropriado para modelar o início da infestação, quando se tem um excesso de zeros, sendo necessário considerar modelos que levem isso em consideração.

As estatísticas do critério de Gelfand & Ghosh (1998) para seleção dos modelos avaliados em março/96, mostraram similaridade entre todos os modelos espacialmente estruturados (Tabela 4), tendo somente uma diferença igual a 25 entre o maior e o menor valores das médias *a posteriori* da *deviance* preditiva esperada (DM), indicando que, com base nesse critério, nenhum desses modelos teve um melhor ou pior ajuste. Por outro lado, o modelo sem dependência espacial, teve uma maior média *a posteriori* para a estatística DM , com uma diferença de 47 unidades acima do maior valor de DM para os outros 12 modelos avaliados. 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 na Figura 4. A Figura 5 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

Tabela 3: 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%
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
MM*	168	120	87	14	3
Obs**	175	113	87	14	3

* Modelo de mistura

** dados observados

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 aumentam rapidamente. Isso está refletido nos mapas das médias *a posteriori* dos riscos de infestação da Figura 4.

Entretanto, parece pouco provável que esse rápido incremento na infestação obedeça só à dinâmica da praga dentro da área experimental. Aparentemente fatores externos influenciaram também o aumento da infestação nos últimos dois meses avaliados (Fev-Mar/96). Esse período coincide com a época em que culturas mais velhas e de baixa produção, que ficavam por perto, tinham sido submetidas a uma poda drástica ou “decepa” (esta atividade, geralmente, é feita nos dois primeiros meses de cada ano), obrigando as populações de broca

Tabela 4: 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
MM	8771	8658	113

DM = Deviance preditiva esperada ($DM = PM + GM$)

PM = termo de penalização

GM = Medida de qualidade de ajuste

MM = modelo de mistura.

destas culturas a procurar refúgio em frutos de culturas mais novas incluindo a área experimental sob estudo. Segundo Sreedharan et al. (1994), a dispersão da broca é também grandemente ajudada pelo vento. Isso sugere que seja considerada em futuros trabalhos a incorporação de covariáveis relacionando fatores ambientais e práticas de manejo da lavoura na modelagem do risco de infestação desta praga, em uma tentativa de explicar melhor o fenômeno sob estudo.

3.2 Modelos de mistura

As médias *a posteriori* dos riscos de infestação para os modelos de mistura avaliados separadamente em cada tempo 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 va-

lores 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 6. De forma semelhante, 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 (Figuras 6a e 6b). A robustez dessas estimativas *a posteriori* à escolha de diferentes *prioris*, indica que são os dados, e não a informação *a priori*, que estão direcionando o ajuste do modelo nesse caso.

4 Considerações finais

Com base nos resultados apresentados pode-se concluir que a dispersão da infestação da broca do café no espaço pode ser modelada adequadamente usando modelos hierárquicos Bayesianos dada a sua facilidade para incorporar facilmente efeitos aleatórios com e sem dependência espacial, além de covariáveis. O modelo igualmente pode ser estendido para incluir um componente temporal. Entretanto, dados os problemas computacionais que surgiram durante a realização deste trabalho e que limitaram de forma considerável a exploração de toda a informação disponível, é desejável procurar alternativas de *software* para ajustar esse tipo de modelos ou tentar implementá-los em alguma linguagem de programação de forma mais eficiente.

Em geral, os modelos avaliados foram pouco influenciados pela escolha de distribuições *a priori* para seus parâmetros e hiperparâmetros, sugerindo que são os dados, e não a informação *a priori*, que estão direcionando o ajuste do modelo nesses casos. Entretanto, isso deve ser confirmado com uma análise envolvendo um intervalo de valores mais amplo para as *prioris* utilizadas.

O uso de efeitos aleatórios espacialmente dependentes nos modelos permitiu identificar mais claramente áreas de maior ou menor intensidade da infestação nos diferentes meses avaliados. Por outro lado, os 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. Isto deverá ser levado em consideração na modelagem espaço-temporal do risco da infestação da broca.

Embora os modelos aqui apresentados tenham sido avaliados em pequena escala, eles poderiam ser adaptados para modelar a infestação da praga em nível regional, considerando a infestação das diferentes unidades produtivas (fazendas) de uma determinada região.

O modelo permite manter a economia na tomada de informação ao permitir modelar as taxas de infestação da praga mesmo quando não é feita a contagem do total de frutos das árvores não infestadas. O uso do método de imputação múltipla permitiu estimar razoavelmente bem esses dados faltantes. Porém, seria interessante avaliar o desempenho do método usando outros modelos probabilísticos diferentes do modelo de regressão linear, 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.

Faz-se necessário incorporar o efeito de covariáveis relacionando fatores ambientais e práticas de manejo da lavoura na modelagem do risco de infestação da praga, visando explicar melhor o processo de infestação.

A inclusão de covariáveis ambientais nos modelos para estimar o total de frutos por árvore também é de interesse, dado que a formação de flores (que posteriormente darão origem aos frutos) está diretamente influenciada pela intensidade e distribuição dos períodos de chuva ao longo do ano, dentre outras variáveis climáticas, influenciando assim diretamente a distribuição de frutos nas árvores.

5 Agradecimentos

Os autores agradecem ao Centro Nacional de Investigaciones de Café, (CENI-CAFE, Chinchiná, Colombia) pelo fornecimento dos dados sobre infestação da broca do café.

Bibliografia

- Besag, J. (1974), ‘Spatial interaction and the statistical analysis of lattice systems (with discussion)’, *Journal of the Royal Statistical Society Series B* **36**, 192–236.
- Besag, J., York, J. C. & Mollié, A. (1991), ‘Bayesian image restoration with two applications in spatial statistics (with discussion)’, *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

- Best, N., Cowles, M. K. & Vines, K. (1996), *CODA: convergence diagnosis and output analysis software for Gibbs sampling output and version 0.30*, Cambridge: Cambridge University.
- Bliss, I. (1941), 'Statistical problems in estimating populations of japanese beetle larvae', *Journal of Economic Entomology* **34**, 221–232.
- Darnell, J., Meinke, L., Young, L. & Gotway, C. (1999), 'Geostatistical investigation of the small-scale spatial variation of western corn rootworm (coleoptera : Chrysomelidae) adults', *Environmental Entomology* **28**, 266–274.
- David, F. & Moore, P. (1954), 'Notes on contagious distributions in plant populations', *Annals of Botany* **18**, 47–53.
- Gelfand, A. & Ghosh, S. (1998), 'Model choice: A minimum posterior predictive loss approach', *Biometrika* **85**, 1–11.
- Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to calculating posterior moments, in 'Bayesian statistics 4', Oxford: Clarendon Press, pp. 145–155. Bernardo, J. M. and Berger, J.O. and David, A.P. and Smith, A.F.M. (Ed.).
- Greig-Smith, P. (1952), 'The use of random and contiguous quadrats in the study of the structure of plant communities', *Annals of Botany* **16**, 293–316.
- Heidelberger, P. & Welch, P. (1983), 'Simulation run length control in the presence of an initial transient', *Operations Research* **31**, 1109–1144.
- Iwao, S. (1968), 'A new regression model for analysing the aggregation pattern of animal populations', *Researches on Population Ecology* **10**, 1–20.
- Le Pelley, R. (1968), *The pests of coffee*, Longmans Green, London.
- Liebhold, A., Zhang, X., Hohn, M., Elkinton, J., Ticehurst, M., Benzon, G. & Campbell, R. (1991), 'Geostatistical analysis of gypsy moth (lepidoptera:lymantriidae) eggs mass populations', *Environmental Entomology* **20**, 1407–1417.
- Lloyd, M. (1967), 'Mean crowding', *Journal of Animal Ecology* **36**, 1–30.
- Moran, P. (1948), 'The interpretation of statistical maps', *Journal of the Royal Statistical Society, Series B* **10**, 243–251.

- Morisita, M. (1962), 'I -index, a measure of dispersion of individuals', *Researches on Population Ecology* **4**, 1–7.
- Nestel, D. & Klein, M. (1995), 'Geostatistical analysis of leaf hopper (homoptera : Cicadellidae) colonization and spread in deciduous orchards', *Environmental Entomology* **24**, 1032–1039.
- Perry, J. (1998), 'Measures of spatial pattern for counts', *Ecology* **79**, 1008–1017.
- Raftery, A. & Lewis, S. (1992), 'Comment: one long run with diagnostics: implementation strategies for markov chain monte carlo', *Statistical Science* **7**, 493–497.
- Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, John Wiley, New York.
- Schotzko, D. & O'keeffe, L. (1989), 'Geostatistical description of the spatial distribution of lygus hesperus (heteroptera:miridae) in lentils', *Journal of Economic Entomology* **82**, 1277–1288.
- Schotzko, D. & Quisenberry, S. (1999), 'Pea leaf weevil (coleoptera:curculionidae) spatial distribution in peas', *Environmental Entomology* **28**, 477–484.
- Spiegelhalter, D., Thomas, A. & Best, N. (2000), *WinBUGS: version 1.3 and user manual*. Cambridge, Cambridge University.
- Sreedharan, K., Balakrishnan, M., Prakasam, C., Krishnamoorthy, B. & Naidu, R. (1994), 'Bio-ecology and management of coffee berry borer', *Indian Coffee* **58**, 5–13.
- Taylor, L. (1961), 'Aggregation, variance and mean', *Nature* **189**, 732–735.
- Williams, L., Schotzko, D. & McCaffrey, J. (1992), 'Geostatistical description of the spatial distribution of limonius californicus (coleoptera:elateridae) wire-worms in the northwestern united states, with comments on sampling', *Environmental Entomology* **21**, 983–995.

Figura 3: Médias *a posteriori* do risco da infestação da broca do café para diferentes esquemas de vizinhança (a – d), e para modelos sem estrutura espacial (f) ou só com efeitos espacialmente estruturados (e).

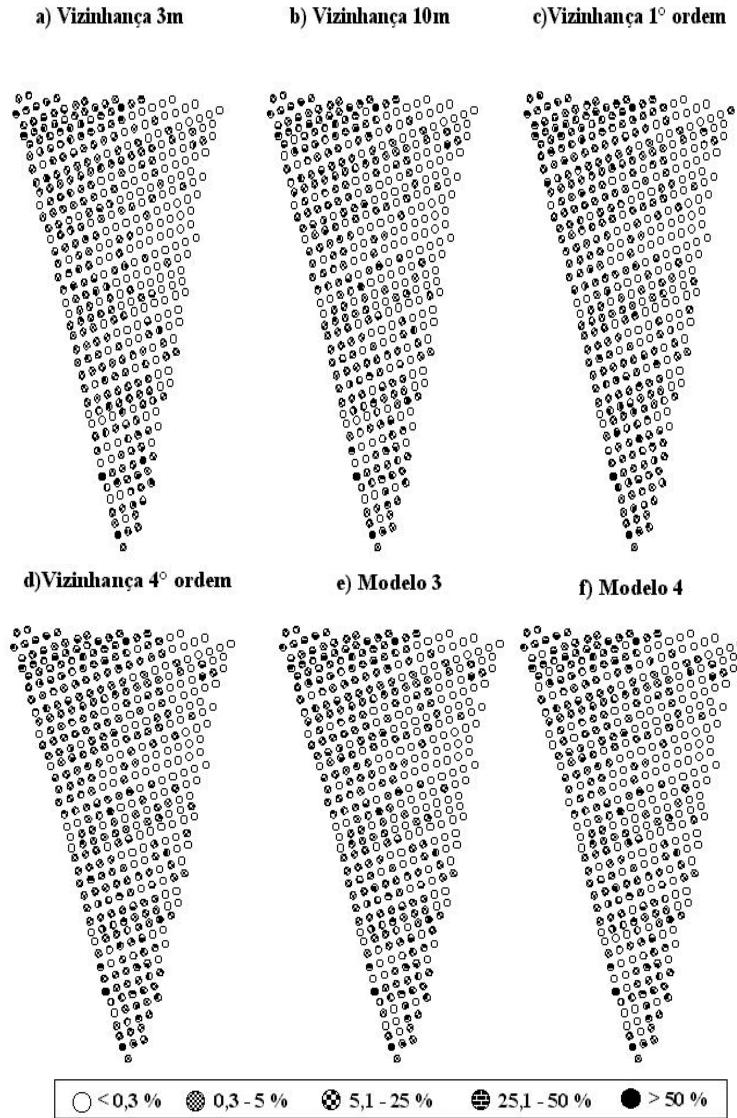


Figura 4: Médias a posteriori do risco da infestação da broca do café obtidas com o modelo 3 (só com efeitos espacialmente estruturados).

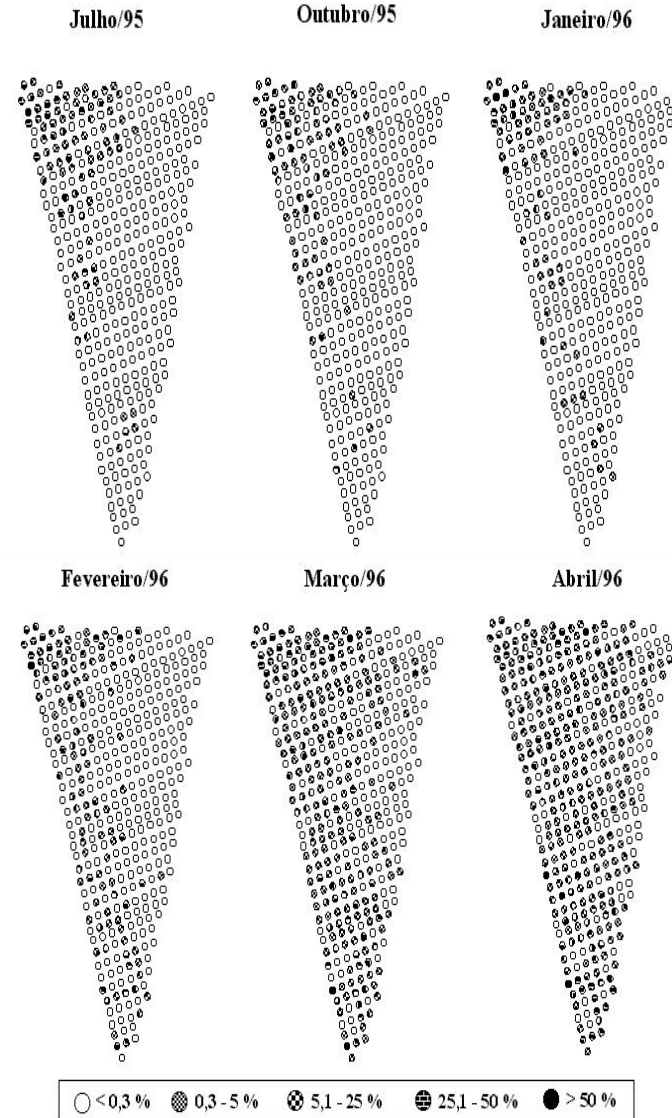


Figura 5: Médias a posteriori e intervalos de credibilidade Bayesianos para os parâmetros do modelo 3 nos diferentes meses avaliados.

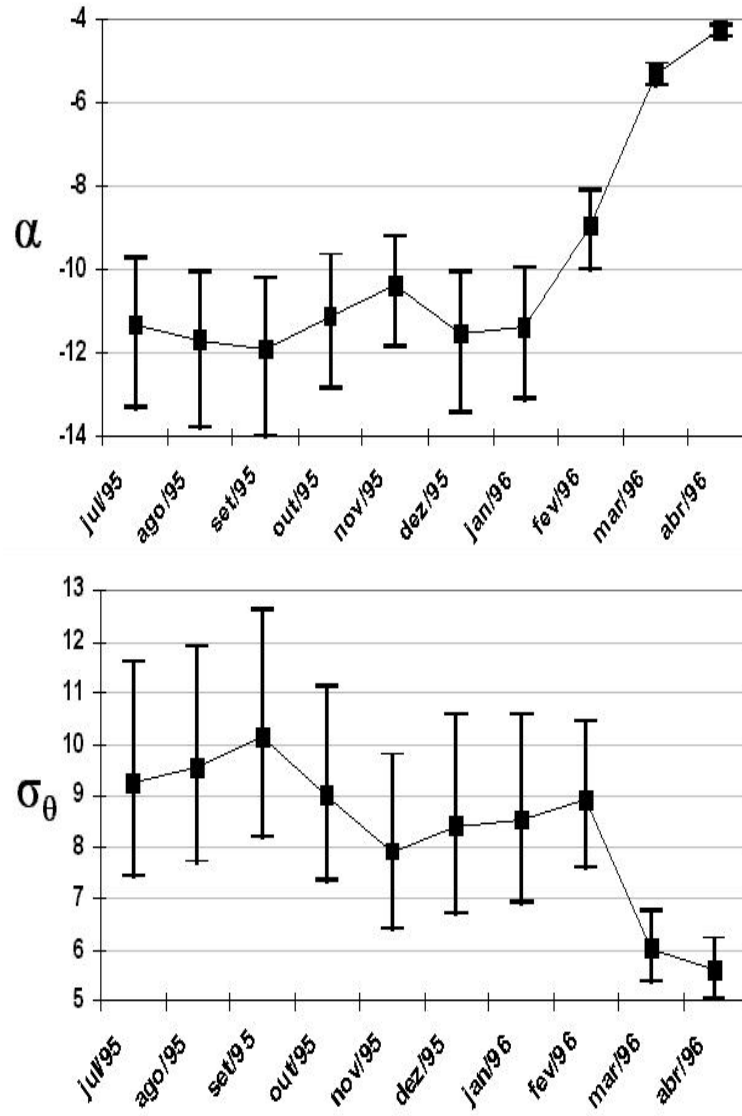


Figura 6: Médias a posteriori do risco da infestação da broca usando diferentes priors para o parametro p no modelo de mistura.

