MODELO AUTOLOGÍSTICO PARA DADOS DE CITRUS

Luziane FRANCISCON [1]
Elias Teixeira KRAINSKI [2]
Paulo Justiniano RIBEIRO JUNIOR [3]

- RESUMO: O modelo autologístico é flexível para modelar a incidência de doenças em plantas. A estimação do modelo é feita por pseudo-verossimilhança. A inferência é feita usando bootstrap via amostrador de Gibbs. Este procedimento é computacionalmente muito caro e nós propomos o método de Monte Carlo para o teste dos parâmetros de dependência espacial do modelo. Também é proposta uma extensão do modelo para considerar a dimensão temporal em dados com várias avaliações no tempo. A metodologia foi aplicada a dados de Morte Súbita dos Citrus e os resultados obtidos pelos métodos bootstrap e Monte Carlo foram comparados. Nos três modelos analisados, verificou-se significância das covariáveis de vizinhança na linha, tanto pelo procedimento bootstrap quanto pelo método de Monte Carlo. As funções em R que implementam a metodologia encontra-se disponível para download no pacote Rcitrus.

- PALAVRAS-CHAVE: modelo autologístico, amostrador de Gibbs, morte súbita dos citrus.

1 Introdução

A produção citrícola brasileira tem notável importância na economia nacional. Com a maior parte da produção destinada para a indústria de suco, a qual, responde por 53% do suco de laranja produzido no mundo e por 80% do suco concentrado que

[1]Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná-UFPR. Caixa Postal 19081, CEP 81531-990, Curitiba, Paraná, Brasil. E-mail: luziane@ufpr.br

[2]Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná-UFPR. Caixa Postal 19081, CEP 81531-990, Curitiba, Paraná, Brasil. E-mail: elias@est.ufpr.br

[3]Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná-UFPR. Caixa Postal 19081, CEP 81531-990, Curitiba, Paraná, Brasil. E-mail: paulojus@est.ufpr.br

transita pelo mercado internacional. Citricultores, indústrias e cientistas brasileiros criaram um setor de ponta na agroindústria nacional. Estes trabalham em busca do aumento de produtividade e também controle e manutenção da capacidade produtiva agrícola, ameaçada por doenças que geram grandes danos em pomares produtores.

As doenças em pomares produtores comprometem a quantidade e qualidade das frutas cítricas. Algumas doenças, como o cancro cítrico e a *Morte Súbita dos Citrus* - MSC, podem causar a erradicação de talhões inteiros de plantas.

A MSC representa uma ameaça potencial para a citricultura paulista e nacional, uma vez que, afeta todas as variedades comerciais enxertadas em limoeiro *Cravo*. Esta doença provoca diminuição de tamanho, peso e quantidade de frutos, além de rápido definhamento e a morte das plantas. Suspeita-se que seja causada por um vírus transmitido de forma bastante eficiente por um vetor aéreo. A gravidade desta doença é devida a grande quantidade de plantas produtoras enxertadas sobre limoeiro *Cravo*.

O modelo autologístico é flexível para modelar a incidência de doenças em plantas. Trata-se de um modelo interessante e de simples interpretabilidade que modela a dependência espacial, porém não conhecido na literatura de citrus. Na agricultura este modelo foi inicialmente estudado para a análise da incidência de *Phytophthora* em pimentas de sino (GUMPERTZ M. L. ; GRAHAM; RISTAINO, 1997).

Neste trabalho, o modelo autologístico é aplicado a dados de MSC para verificar a dependência espacial da doença. É proposto uma extensão do modelo na dimensão temporal para os dados. A metodologia de estimação do modelo foi implementada em R . Estas funções foram adicionadas ao pacote **Rcitrus**, disponível para *download* em `http://www.est.ufpr.br/Rcitrus`.

A Seção 2 traz a metodologia do modelo autologístico para modelagem de dependência espacial em dados binários. É descrito a estrutura de vizinhança utilizada, os métodos computacionalmente intensivos utilizados e a expansão do modelo para a dimensão temporal. Na Seção 3 aplica-se o modelo autologístico em dados de incidência de MSC.

2 Metodologia

A modelagem de variáveis binárias pode ser feita via regressão logística. A regressão logística é um caso particular dos modelos de regressão da família exponencial, os chamados modelos lineares generalizados (NELDER; WEDDERBURN, 1972). Em dados de incidência de doenças em plantas, o modelo de regressão logística usual não é adequado. Na maioria dos casos não há indepêndencia das observações. Espera-se que plantas próximas tenham características similares, particularmente, que a localização das plantas doentes apresentem algum padrão espacial. Quando as plantas são dispostas em linhas e colunas em uma área, talhão ou *lattice*, pode-se utilizar um modelo autoregressivo espacial para dados binários, onde as observações vizinhas são utilizadas como covariadas.

2.1 Especificação do modelo autologístico

No modelo autologístico, a probabilidade de sucesso é modelada como combinação linear de presença ou ausência de sucesso em locais vizinhos e de covariadas que captam informações adicionais nesses locais:

$$P(Y = y_i|x_{ji}, y_{(i)}) = logit(y_i) = \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{k=1}^{q} \gamma_k y_{(i)} \ , \tag{1}$$

onde $y_{(i)}$ é o vetor de observações sem a i-ésima observação, β_j mede a influência da covariada x_j e γ_k mede a influência de vizinhança de ordem k. No caso de *lattices* essa expressão pode ser reescrita como

$$P(Y = y_{kl}|x_{kl}, y_{(kl)}) = logit(y_{kl}) = \sum_{j=1}^{p} \beta_j x_{kl} + \sum_{k=1}^{q} \gamma_k y_{(kl)} \ , \tag{2}$$

onde o índice kl indica a posição da observação na linha k e na coluna l.

Um *lattice* D é uma região com n posições, cada uma descrita pelas coordenadas (k, l) especificando a linha e coluna do *lattice* que a observação é alocada. Em cada posição é observada uma resposta binária $y_{k,l}$, onde $y_{k,l}$ tem valor 1 na presença da variável de interesse e 0 caso contrário, e um vetor $p \times 1$ de covariadas $x_{k,l}$. As respostas binárias de n posições correspondem a $Y = (y_{k,l}, (k, l) \in D)$ que constituem um mapa da ocupação da variável de interesse HE F. ; ZHOU e ZHU (2003).

A expressão do modelo autologístico é dada por

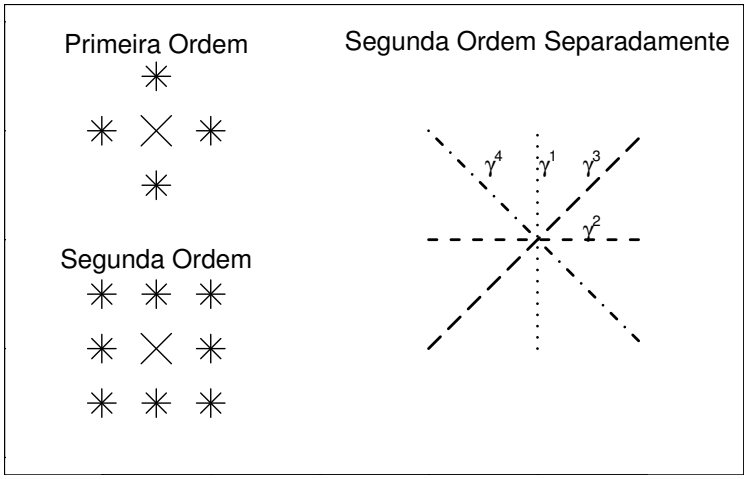
$$Pr(Y_{k,l} = y_{k,l}|x_{k,l}, y_{(k,l)}, (k, l) \in D) = \frac{exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y_{(k,l)}\}}{1 + exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y_{(k,l)}\}} \ , \tag{3}$$

onde os $\beta' s$ são parâmetros de regressão e $\gamma' s$ são os parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariadas e $y_{(k,l)}$ representam as covariadas de vizinhança.

Estrutura de Vizinhança: A estrutura de vizinhança pode ser construída de diferentes formas, considerando simplesmente a ordem ou considerando linhas e colunas separadamente. Na Figura 1 estão representadas algumas estruturas de vizinhança.

Na estrutura de vizinhança de segunda ordem separadamente, a primeira covariada inclui informações dos dois vizinhos na mesma linha em colunas adjacentes, posições $(k, l+1)$ e $(k, l-1)$. A segunda covariada inclui informações dos dois vizinhos na mesma coluna em linhas adjacentes, posições $(k-1, l)$ e $(k+1, l)$, a terceira covariada, que inclui dois vizinhos da diagonal $(-1, 1)$, posições $(k-1, l+1)$ e $(k+1, l-1)$. E a quarta covariada inclui dois vizinhos da diagonal $(1, 1)$, posições $(k-1, l-1)$ e $(k+1, l+1)$. Esta estrutura mais flexível é particularmente útil no caso com doenças de plantas que possuem espaçamento diferente entre linhas e colunas.

Figura - 1: Estruturas de vizinhança: primeira ordem, segunda ordem e segunda ordem separadamente



2.2 Inferência para o modelo autologístico

No modelo autologístico não é possível obter a expressão da função de verossimilhança para a estimação dos parâmetros do modelo. A probabilidade de sucesso de uma observação $s_{k,l}$, neste modelo, é condicional ao *status* das outras observações $v_{k,l}$.

2.2.1 Estimação por pseudo-verossimilhança

(BESAG, 1972) sugere um método de estimação que maximize uma função de pseudo-verossimilhança para o modelo ajustado, dada por:

$$\begin{aligned} l_{ps} &= \sum_{(k,l)\in D} ln[P(Y_{k,l} = y_{k,l}|todos\ os\ outros\ valores)] \ , \\ l_{ps} &= \sum_{(k,l)\in D} \{y_{k,l} f_{k,l}(\theta) - ln[1 + \exp(f_{k,l}(\theta))]\} \ , \end{aligned}$$

onde $f_{k,l}(\theta) = \beta_0 + \beta_1 x_{k,l}^T + \gamma_1 y_{(k,l)} + \gamma_2 y_{(k,l)} + \gamma_3 y_{(k,l)} + \gamma_4 y_{(k,l)}$, e $\theta = (\beta_0, \beta_1^T, \gamma_1, \gamma_2, \gamma_3, \gamma_4)^T$

Neste método não há grandes problemas ao estimar os efeitos, porém as estimativas de variância dos efeitos são subestimadas. É preciso um procedimento alternativo para reestimar os erros-padrão das estimativas. O procedimento proposto por (BESAG, 1972) é o método de Bootstrap paramétrico.

2.2.2 Estimação pelo método de Bootstrap

A metodologia bootstrap, consiste em gerar amostras dos dados originais para estimar a quantidade de interesse com estas amostras, ou seja, simular N *lattices* $(\tilde{y}^1, ..., \tilde{y}^N)$. Devido ao fato de que cada y_i^t estar condicionada a posição original dentro do *lattice*, é preciso preservar esta condição em cada amostra simulada. Para considerar isso, o procedimento bootstrap deve ser paramétrico. Neste caso, deve-se ter informação suficiente sobre a forma da distribuição dos dados. O modelo autologístico é ajustado aos dados observados, pelo método da pseudo-verossimilhança, γ_t^0. As estimativas dos parâmetros são obtidas a partir dos N γ^t, obtidos do modelo ajustado ao N $\tilde{y}^t$ simulados. Devido a configuração dos dados, esse procedimento de reamostragem só é possível utilizando o algoritmo amostrador de Gibbs.

Amostrador de Gibbs: O amostrador de Gibbs é um procedimento de amostragem baseada em distribuições condicionais. A idéia desse amostrador é de retirar amostras sucessivas da distribuição de cada dado em particular, condicionado a todos os outros.

Construímos um algoritmo de Gibbs para obter N amostras $\hat{\gamma}$ da seguinte forma:

1. Iniciar com os dados observados, e fazer um ciclo pelas plantas atualizando cada planta em ordem aleatória.

2. Para atualizar cada planta deve-se simular o status (doente/sadia) de acordo com o modelo ajustado dos dados observados e o estado atual dos vizinhos.

3. Uma varredura completa é realizada após atualizar o status de todas as plantas uma vez.

4. Ajustar o modelo com os dados atuais e guardar as estimativas dos parâmetros.

5. Repetir os passos anteriores N vezes.

Para obter a estimativa de γ é preciso simular y_t de $P(y_i^t|y_{i-1}^{t-1})$, onde $y_{i-1}^{t-1} = (y_0^t, ..., y_{i-1}^t, y_{i+1}^{t-1}, ..., y_n^{t-1})$ são as observações atuais da cadeia, reconstruir as covariadas de vizinhança e obter as estimativas de $\hat{\gamma}$ ajustando um GLM usual. Os N vetores de $\hat{\gamma}$ são utilizados para obter o erro-padrão para $\hat{\gamma}$ estimado com os dados observados.

2.2.3 Inferência via método de Monte Carlo

O método de Monte Carlo é o procedimento mais simples para se fazer inferência sobre parâmetros, a respeito dos quais desconhecemos a distribuição e, consequentemente, uma estatística de teste. Esse método consiste em obter amostras do parâmetro de interesse sob hipótese a nula. Isso é feito ajustando-se o modelo proposto à dados simulados sob a hipótese nula. A partir das amostras dos

parâmetros obtida, obtem-se a distribuição empírica dos parâmetros sob a hipótese nula. O valor observado na amostra de dados reais é avaliado na distribuição empírica obtida e o valor-p é obtido.

No caso do ajuste do modelo autologístico à dados de incidência de doenças em plantas, a hipótese nula é que não há dependência espacial ou o padrão espacial é aleatório. Então, ao se ajustar o modelo autologístico, os parâmetros das covariáveis de vizinhança não serão significativamente diferentes de zero. A amostra desses parâmetros sob a hipótese de padrão aleatório é obtida a partir de uma distribuição aleatória das plantas, doentes e sadias.

O algoritmo de Monte Carlo nesse caso é dado pelos passos:

1. Ajustar o modelo autologístico aos dados reais;
2. Aleatorizar as plantas no talhão;
3. Ajustar o modelo autologístico aos dados aleatorizados;
4. Repetir os passos 2 e 3 N vezes,
5. Comparar as estimativas do passo 1 com as estimativas do passo 3.

O método de Monte Carlo é mais simples que o método de bootstrap e sua aplicação extremamente mais rápida. A vantagem do procedimento de bootstrap é que são obtidas estimativas por intervalo para os parâmetros. Porém, o método de Monte Carlo traz resultados muito semelhantes sobre a hipótese de dependência espacial, tornando-se uma opção aplicável a grandes conjuntos de dados.

2.3 Covariadas de vizinhança no período de tempo anterior

Existem situações onde as observações são medidas em diferentes períodos de tempo. Neste caso podería-mos ajustar um modelo autologístico para as observações em cada um dos períodos. Uma proposta é modelar os dados de um período de tempo t, utilizando as covariadas de vizinhança no período de tempo anterior $t-1$.

Esta abordagem permite explorar o potencial preditivo do modelo. Pois o modelo ajustado em cada avaliação, tendo como base as covariadas do tempo anterior, pode ser utilizado para predizer a incidência no tempo futuro, baseado nas covariadas de vizinhança do período atual.

A expressão do modelo autologístico, neste caso, é dada por

$$Pr(Y_{k,l} = y^t_{k,l} | x_{k,l}, y^{t-1}_{(k,l)}, (k,l) \in D) = \frac{exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y^{t-1}_{(k,l)}\}}{1 + exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y^{t-1}_{(k,l)}\}} \ , \quad (4)$$

onde os $\beta' s$ são parâmetros de regressão e $\gamma' s$ são os parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariadas e $y^{t-1}_{(k,l)}$ representam as covariadas de vizinhança.

Nessa situação, não há indução de autocorrelação pelo uso repetitivo das observações. Portanto a estimação dos parâmetros de variância, podem ser feitas por pseudo-verossimilhança, ou seja, as estimativas fornecidas pelo modelo de regressão logística usual.

2.4 Covariadas de vizinhança no tempo atual e anterior

A partir do modelo ajustado com covariadas no tempo anterior, podemos testar se ainda sobrou algum efeito possível de ser explicado pelas covariadas no tempo atual. Ou seja, construir o modelo autologístico considerando covariadas no período de tempo atual e período de tempo anterior em um mesmo modelo. A expressão do modelo é dada por

$$Pr(Y_{k,l} = y^t_{k,l} | x_{k,l}, y^t_{k,l}, y^{t-1}_{(k,l)}, (k,l) \in D) = \frac{exp\{\eta^*\}}{1 + exp\{\eta^*\}} \ , \tag{5}$$

onde η^* é dado por

$$\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y^t_{(k,l)} + \gamma_t y^{t-1}_{(k,l)} \ .$$

Aqui covariadas de vizinhança estão tanto no período de tempo anterior, $t-1$ quanto no período de tempo t. Neste caso os erros-padrão das estimativas dos parâmetros no tempo t precisam ser estimados pelo procedimento de reamostragem, pois há a indução de autocorrelação e os erros padrões usuais não são apropriados.

3 Resultados

Os dados analisados são dados de incidência de MSC em um talhão da Fazenda Vale Verde, localizada no município de Comendador Gomes, estado de Minas Gerais. Esse talhão possui 20 linhas de plantas com 48 plantas em cada linha. O espaçamento entre linhas é de $7,5$ metros e entre as plantas na linha é de 4 metros. Foram analisadas 11 avaliações feitas entre os dias 05/11/2001 e 07/10/2002.

A incidência da doença variou de $14,9\%$ na primeira avaliação até $45,73\%$ na 11^o avaliação. O mapa com a incidência da doença em cada avaliação está no apêndice.

A variável resposta de interesse é a presença ou ausência de MSC. Na modelagem estatística foi considerada a estrutura de vizinhança de segunda ordem separadamente. Devido ao espaçamento entre linhas ser diferente do espaçamento de plantas na linha, essa estrutura de vizinhança permite verificar se a correlação espacial depende do espaçamento. Se o contágio tiver um alcance pequeno, espera-se que apenas as covariadas de vizinhança na linha sejam signicativas.

Os modelos propostos são

$$\begin{array}{lll}
1 & logit(p^t_{kl}) = & \beta_0 + \gamma_1 L^t_{kl} + \gamma_2 C^t_{kl} + \gamma_3 Da^t_{kl} + \gamma_4 Db^t_{kl} \\
2 & logit(p^t_{kl}) = & \beta_0 + \gamma_1 L^{t-1}_{kl} + \gamma_2 C^{t-1}_{kl} + \gamma_3 Da^{t-1}_{kl} + \gamma_4 Db^{t-1}_{kl} \\
3 & logit(p^t_{kl}) = & \beta_0 + \gamma_1 L^{t-1}_{kl} + \gamma_2 C^{t-1}_{kl} + \gamma_3 Da^{t-1}_{kl} + \gamma_4 Db^{t-1}_{kl} + \\
 & & \gamma_5 L^t_{kl} + \gamma_6 C^t_{kl} + \gamma_7 Da^t_{kl} + \gamma_8 Db^t_{kl}
\end{array}$$

O teste da dependência espacial, consiste em testar a significância dos coeficientes associados as covariadas de vizinhança. Para isso utilizamos o seguinte

resultado assintótico

$$\frac{\hat{\gamma}}{ep_{\hat{\gamma}}} \sim N(0,1) \; . \tag{6}$$

No modelo 1, o teste da significância dos coeficientes, estará testando a existência da dependência espacial e permitirá ver se a agregação ocorre apenas nas linhas (curto alcance), entre as linhas (médio alcance) ou também nas diagonais (longo alcance). Além disso, a análise do efeito das covariadas de vizinhança nas diagonais permitirá avaliar a existência de um efeito direcional da agregação.

A análise geral da significância dos coeficientes do modelo 2, estará testando a capacidade preditiva deste modelo. Será interessante saber se é possível prever o *status* das plantas em um momento futuro. A análise da significância de cada coeficiente separadamente, permitirá estudar a forma de propagação da doença. Por exemplo, se o coefiente da covariada de vizinhança na linha for significativo, podemos dizer que a doença tendeu a se propagar na linha.

No modelo 3, a significância dos coeficientes das covariadas no tempo anterior, poderá ser interpretado da mesma forma que no modelo 2. Nesse modelo, o teste da significância das covariadas no tempo atual, estará testando se houve a formação ou incremento de agregação entre a data da avaliação anterior e a data atual.

Na Tabela 1 estão apresentadas as datas das avaliações, as incidências e as estimativas dos parâmetros dos modelos para cada uma das avaliações.

Tabela - 1: Datas, incidências e estimativas dos parâmetros do modelo 1, ajustado para as 11 avaliações

Data	Incidência	$\hat{\beta}_0$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$
05/11/01	0.14895	-2.02052	0.32079	-0.02221	0.00699	0.20609
05/12/01	0.17293	-1.97306	0.34912	0.22879	0.13854	0.16093
04/01/02	0.21875	-1.84436	0.62823	-0.02529	0.16258	0.23295
13/02/02	0.23840	-1.78096	0.70992	-0.09897	0.21175	0.20843
14/03/02	0.26354	-1.68169	0.58892	-0.02604	0.30987	0.16706
05/04/02	0.27812	-1.63307	0.63199	-0.00680	0.18708	0.23912
08/05/02	0.32292	-1.45117	0.60624	0.06947	0.09455	0.19067
03/06/02	0.33125	-1.39161	0.62401	0.13288	0.02720	0.13081
06/07/02	0.34167	-1.28953	0.60778	0.07711	-0.05904	0.18728
06/09/02	0.37500	-0.90676	0.47809	0.01132	-0.11679	0.07101
07/10/02	0.45729	-0.90008	0.52397	0.12469	-0.07815	0.16272

A Tabela 2 traz os valores-p obtidos para os coeficientes estimados do modelo 1. Os erros-padrão utilizado para o cálculo desses valores-p foram estimados pelo método bootstrap. No procedimento bootstrap foram realizadas 1100 simulações e o cálculo dos erros-padrão foi feito com as últimas 1000 simulações.

Analisando a Tabela 2 vemos que existe dependência espacial somente dentro da linha a partir da terceira avaliação, valores-p menores que $0,01$. Esse resultado

Tabela - 2: Valores-p da hipótese de nulidade dos parâmetros do modelo 1 ajustado para as 11 avaliações

Avaliação	$\hat{\beta}_0$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$
01	0.00000	0.27365	0.94062	0.98163	0.46181
02	0.00000	0.16735	0.37606	0.58584	0.52980
03	0.00000	0.00221	0.90968	0.44806	0.26035
04	0.00000	0.00036	0.63991	0.28060	0.29245
05	0.00000	0.00126	0.89330	0.08113	0.37826
06	0.00000	0.00025	0.97196	0.31854	0.17815
07	0.00000	0.00018	0.69008	0.56112	0.24626
08	0.00000	0.00014	0.44090	0.87295	0.42052
09	0.00000	0.00005	0.65054	0.70722	0.24503
10	0.00002	0.00190	0.94235	0.43899	0.64234
11	0.00006	0.00028	0.40094	0.59804	0.25456

permite duas conclusões: a dependência espacial é de curto alcance, pois o espaçamento dentro da linha é de 4 metros e entre linhas é de 7,5 metros; e, quando há baixa incidência não existe dependência espacial.

Através do modelo ajustado podemos calcular a chance de determinada planta estar doente condicionada ao status da doença das plantas vizinhas dentro da linha. Por exemplo, analisando a 9^o avaliação vemos que se as plantas vizinhas dentro da linha estão doentes, a chance da planta estar doente é 3.60 vezes maior do que se as plantas vizinhas dentro da linha estiverem sadias.

Na Tabela 3 observa-se os resultados obtidos pelo método de Monte Carlo. Devido a esse método ser muito mais rápido que o procedimento de bootstrap, foram feitas 9999 simulações. Observa-se que os resultados são extremamente semelhantes, exceto para $\hat{\beta}$ pois neste caso o teste não é para a nulidade de β, mas para um valor sob hipótese de aleatoriedade.

Na Tabela 4 estão as estimativas dos coeficientes das covariadas de vizinhança para o modelo 2, calculadas no tempo anterior. Foram ajustados 10 modelos, um para cada avaliação a partir da 2^a.

A interpretação desse modelo pode ser feita a partir dos valores-p, Tabela 5. Esses valores-p referem-se ao teste da hipótese de nulidade dos coeficientes da Tabela 4. No cálculo dos valores-p, foram utilizadas a estimativas dos erros-padrão obtidas pelo método da pseudo-verossimilhança.

Analisando a Tabela 5 e a Tabela 6, observa-se por ambos os métodos os p-valores são semelhantes, exeto para β. Dessas tabelas, pode-se concluir que o potencial preditivo passa a ter significância no modelo ajustado para a 3^a avaliação, com covariadas de vizinhança da 2^a avaliação, ou seja, o modelo passa a ter algum potencial preditivo a partir da segunda avaliação. Porém, observa-se que nenhuma covariada de vizinhança da 11^a avaliação é significativa para prever o *status* da última avaliação. Por exemplo, ao considerar o *status* das duas plantas vizinhas na

Tabela - 3: Valores-p obtidos pelo método de Monte Carlo para testar a hipótese de aleatoriedade espacial no modelo 1.

Avaliação	$\hat{\beta}_0$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$
01	0.04930	0.25073	0.93529	0.97930	0.45665
02	0.00670	0.16252	0.35724	0.56846	0.51515
03	0.00020	0.00220	0.90129	0.41524	0.25073
04	0.00020	0.00080	0.60566	0.27103	0.27703
05	0.00010	0.00180	0.88189	0.08991	0.34453
06	0.00020	0.00060	0.96760	0.27503	0.17012
07	0.00000	0.00030	0.66397	0.54535	0.23412
08	0.00020	0.00010	0.39854	0.85859	0.39604
09	0.00050	0.00020	0.61826	0.69807	0.22492
10	0.03100	0.00180	0.94069	0.42604	0.62716
11	0.00130	0.00010	0.36834	0.57786	0.24802

Tabela - 4: Incidência e estimativas dos parâmetros do modelo 2, ajustado para as 11 avaliações

Incidência	$\hat{\beta}_0$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$
0.14896	-1.85856	0.35758	0.09636	0.06987	0.13624
0.17292	-1.72653	0.44081	0.11601	0.16528	0.24510
0.21875	-1.72087	0.63954	-0.04746	0.20502	0.22401
0.23854	-1.62404	0.61800	-0.07895	0.29245	0.20671
0.26354	-1.57709	0.59488	-0.02847	0.24230	0.20916
0.27812	-1.36210	0.58248	0.02760	0.13193	0.21061
0.32292	-1.39174	0.61067	0.10177	0.07507	0.15503
0.33125	-1.29333	0.60794	0.10203	-0.01651	0.15265
0.34167	-0.95363	0.50256	0.01584	-0.08967	0.12701
0.37500	-0.64497	0.43635	0.02653	-0.04305	0.12646
0.45729	1.66580	0.16143	0.19122	0.17514	0.16592

Tabela - 5: Valores-p a hipótese de nulidade dos parâmetros do modelo 2 obtidos pelo procedimento de booststrap em cada uma das 11 avaliações

Avaliação	$\hat{\beta}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$
01	0.0000	0.0542	0.6177	0.7170	0.4588
02	0.0000	0.0054	0.4801	0.3110	0.1156
03	0.0000	0.0000	0.7506	0.1570	0.1130
04	0.0000	0.0000	0.5690	0.0296	0.1231
05	0.0000	0.0000	0.8260	0.0594	0.1007
06	0.0000	0.0000	0.8215	0.2794	0.0787
07	0.0000	0.0000	0.3840	0.5199	0.1842
08	0.0000	0.0000	0.3770	0.8869	0.1816
09	0.0000	0.0000	0.8874	0.4235	0.2493
10	0.0000	0.0000	0.8003	0.6861	0.2260
11	0.0000	0.3378	0.2612	0.3147	0.3410

Tabela - 6: Valores-p para hipótese de aleatoriedade espacial obtidos pelo método de Monte Calo para o modelo 2 ajustado às 11 avaliações

Avaliação	$\hat{\beta}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$
01	0.00530	0.07131	0.62386	0.71987	0.45415
02	0.00000	0.00930	0.48685	0.30723	0.12351
03	0.00000	0.00000	0.74197	0.15432	0.10931
04	0.00000	0.00000	0.55496	0.03040	0.12861
05	0.00000	0.00000	0.82568	0.06011	0.09631
06	0.00000	0.00000	0.82248	0.27573	0.07421
07	0.00000	0.00000	0.37474	0.50555	0.17662
08	0.00000	0.00000	0.36524	0.88149	0.17672
09	0.00040	0.00000	0.88499	0.42254	0.23962
10	0.00020	0.00000	0.80048	0.68797	0.21732

linha de uma planta sadia na avaliação 3. A probabilidade dessa planta ficar doente se nenhuma das suas vizinhas na linha estiver doente é de 15.18%, enquanto que se uma dessas vizinhas estiver doente essa probabilidade passa a ser 25.33% e se as duas vizinhas na linha estiverem doente esta probabilidade passa a ser 39.13%.

As estimativas dos coeficientes estimados para o modelo 3 estão na Tabela 7.

Tabela - 7: Estimativas dos parâmetros do modelo 3, ajustado para as avaliações 2 a 11

Av.	$\hat{\beta}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$	$\hat{\gamma_5}$	$\hat{\gamma_6}$	$\hat{\gamma_7}$	$\hat{\gamma_8}$
02	-1.9735	.0456	-.6440	-.4037	-.1461	.2855	.7686	.4831	.2889
03	-1.8624	-.5592	.3194	.0036	.1073	1.0454	-.2879	.1877	.1804
04	-1.7900	-.2739	.2634	-.0466	.2453	.9420	-.3296	.2717	-.0074
05	-1.6925	.2898	-.4051	-.0741	.2026	.3297	.3337	.3833	-.0013
06	-1.6379	-.2783	-.3310	.8787	-.2150	.8914	.3175	-.6469	.4345
07	-1.4447	.0766	-.1204	.2176	.1786	.5423	.1665	-.0924	.0279
08	-1.4061	.0795	-.4918	1.3170	.6446	.5613	.6233	-1.2719	-.4966
09	-1.2594	.3813	.7266	1.7413	-.3848	.2281	-.6408	-1.7751	.5425
10	-.8544	.3846	.1663	.2043	.5757	.1105	-.1620	-.3192	-.4759
11	-.8728	-.1086	-.3673	.1053	-.1053	.6024	.4173	-.1708	.2444

Na Tabela 8, estão os valores-p calculados para o teste da hipótese de nulidade dos parâmetros do modelo. Nesse cálculo, foram utilizadas as estimativas dos erros-padrão dos coeficientes obtida pelo procedimento bootstrap, com 1100 simulações, onde descartou-se as 100 primeiras simulações.

Tabela - 8: Valores-p da hipótese de nulidade dos parâmetros do modelo 3, ajustado para as avaliações 2 a 11

Av.	$\hat{\beta}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$	$\hat{\gamma_5}$	$\hat{\gamma_6}$	$\hat{\gamma_7}$	$\hat{\gamma_8}$
02	0.0000	0.4111	0.0026	0.0331	0.2320	0.0964	0.0001	0.0148	0.0830
03	0.0000	0.0017	0.0313	0.4916	0.2666	0.0000	0.1170	0.1941	0.2061
04	0.0000	0.0349	0.0426	0.3790	0.0468	0.0000	0.0724	0.0889	0.4850
05	0.0000	0.0175	0.0050	0.3024	0.0779	0.0396	0.0392	0.0168	0.4973
06	0.0000	0.0169	0.0148	0.0000	0.0580	0.0000	0.0483	0.0002	0.0070
07	0.0000	0.2654	0.1690	0.0453	0.0760	0.0005	0.1679	0.2927	0.4354
08	0.0000	0.3000	0.0009	0.0000	0.0000	0.0038	0.0011	0.0000	0.0024
09	0.0001	0.0114	0.0000	0.0000	0.0087	0.1523	0.0017	0.0000	0.0015
11	0.0003	0.0004	0.0844	0.0427	0.0000	0.2461	0.1523	0.0212	0.0014
11	0.0007	0.1644	0.0005	0.1751	0.1728	0.0000	0.0042	0.1249	0.0509

A análise da Tabela 8 não traz conclusões consistentes, pois não há um padrão definido na disposição das covariadas significativas. Em muitas avaliações

há covariadas de vizinhança significativas na linha e nas diagonais, na avaliação anterior ou na atual.

Na Tabela 9, estão os valores-p para a hipótese de aleatoriedade espacial, calculados utilizando-se 9999 simulações de Monte Carlo. Aqui também observa-se a semelhança nos resultados, ressaltando que os valores-p geralmente são próximos ao dobro do valor dos valores-p obtidos pelo procedimento bootstrap, exceto para β cuja hipótese testada é diferente.

Tabela - 9: Valores-p da hipótese de aleatoriedade espacial obtidos pelo método de Monte Carlo, para o model 3 nas avaliações 2 a 11

Av.	$\hat{\beta}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$
02	0.0164	0.8125	0.0016	0.0405	0.4360	0.2554	0.0037	0.0489	0.2431
03	0.0007	0.0013	0.0583	0.9833	0.4958	0.0000	0.1650	0.3563	0.3744
04	0.0011	0.0580	0.0803	0.7513	0.0849	0.0000	0.0934	0.1561	0.9697
05	0.0004	0.0278	0.0047	0.5812	0.1312	0.0748	0.0732	0.0345	0.9940
06	0.0003	0.0279	0.0113	0.0000	0.0977	0.0000	0.0682	0.0001	0.0136
08	0.0003	0.5180	0.3287	0.0705	0.1392	0.0010	0.2942	0.5545	0.8599
08	0.0008	0.4766	0.0000	0.0000	0.0000	0.0006	0.0002	0.0000	0.0019
09	0.0045	0.0006	0.0000	0.0000	0.0008	0.1423	0.0000	0.0000	0.0004
10	0.0767	0.0001	0.1415	0.0651	0.0000	0.4542	0.2802	0.0302	0.0016
11	0.0045	0.3003	0.0003	0.3151	0.3160	0.0000	0.0037	0.2210	0.0825

No apêndice pode-se visualizar os valores obtidos nas 1100 simulações para cada uma das 11 avaliações da MSC analisadas e as densidades.

4 Conclusão

O modelo autologístico foi eficiente para captar a dependência espacial em dados de incidência de doenças em plantas. A estrutura de vizinhança adotada possibilitou investigar vários aspectos da dependência espacial. Além disso, a incorporação da dimensão temporal permite explorar o potencial preditivo do modelo.

A estratégia de análise revelou-se importante para avaliar o alcance da dependência espacial. No talhão analisado, verificou-se que a dependência espacial é de curto alcance, ou seja, apenas a covariada de vizinhança dentro da linha foi significativa na maioria das avaliações. A capacidade preditiva também foi verificada, pois no modelo somente com covariadas no tempo anterior, a covariada de vizinhança dentro da linha foi significativa na maioria das avaliações. Isso indica que o *status* das plantas vizinhas no tempo anterior $t-1$ está influenciando a incidência da doença na planta do tempo t.

Uma vantagem adicional do modelo autologístico frente a alguns métodos tradicionais de análise de dados de doenças de plantas, é a possibilidade de modelar

os dados originais sem discretização de informação, como é feita na análise por *quadrats*.

Na comparação dos métodos utilizados, conclui-se que o método de Monte Carlo fornece resultados semelhantes ao procedimento bootstrap em um tempo extremamente mais curto. Sendo assim recomenda-se o seu uso em caso de grandes quantidades de dados.

- *ABSTRACT: The autologistic model is flexible for modeling the incidence diseases in plants. The single estimatiom method is using pseudo-likelihood. The inference is made using bootstrap saw Gibbs sampler. This procedure is computational very expensive and we consider Monte Carlo method for test the spatial dependence parameters. Also we consider an extencion of the model, for consider the temporal dimension in data with some evaluations in time. The methodologies was applied in Citrus Sudden Death and the results are compared. In three models analyzed, the significance of neighborhood in lines covariate was verified in both, bootstrap procedure and Monte Carlo method. The functions implemented in R are available for download in Rcitrus package.*

- *KEYWORDS: Autologistic model, Gibbs sampling, Citrus Sudden Death*

Referências

BESAG, J. Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistics Society, Series B*, 1972.

GUMPERTZ M. L. ; GRAHAM, J. M.; RISTAINO, J. B. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics*, 1997.

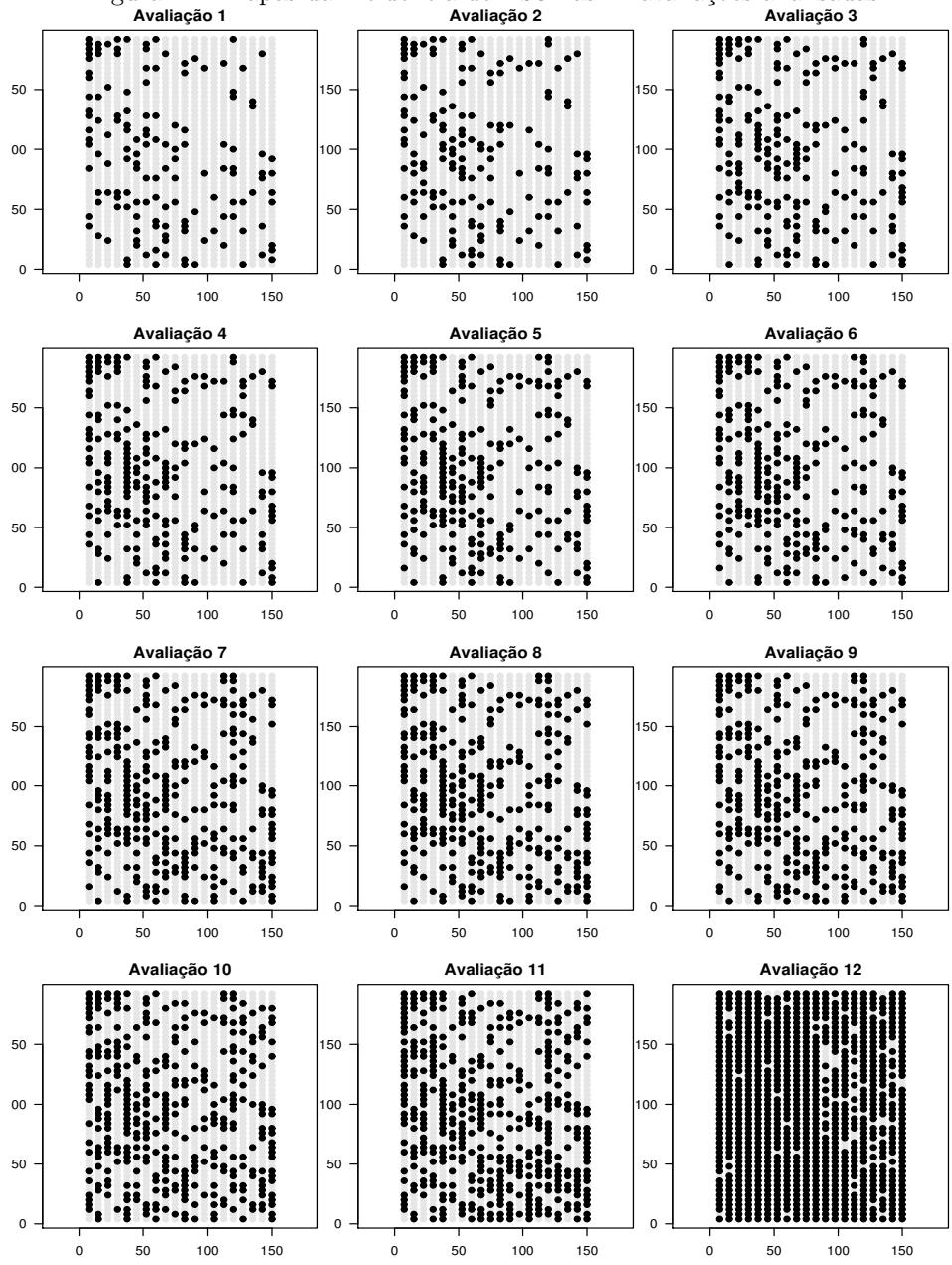
HE F. ; ZHOU, J.; ZHU, H. Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological, and Environmental Statistics*, 2003.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistics Society, Series A*, 1972.

R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2005. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

Apêndice - Figuras

Figura - 2: Mapas da incidência de MSC nas 11 avaliações analisadas



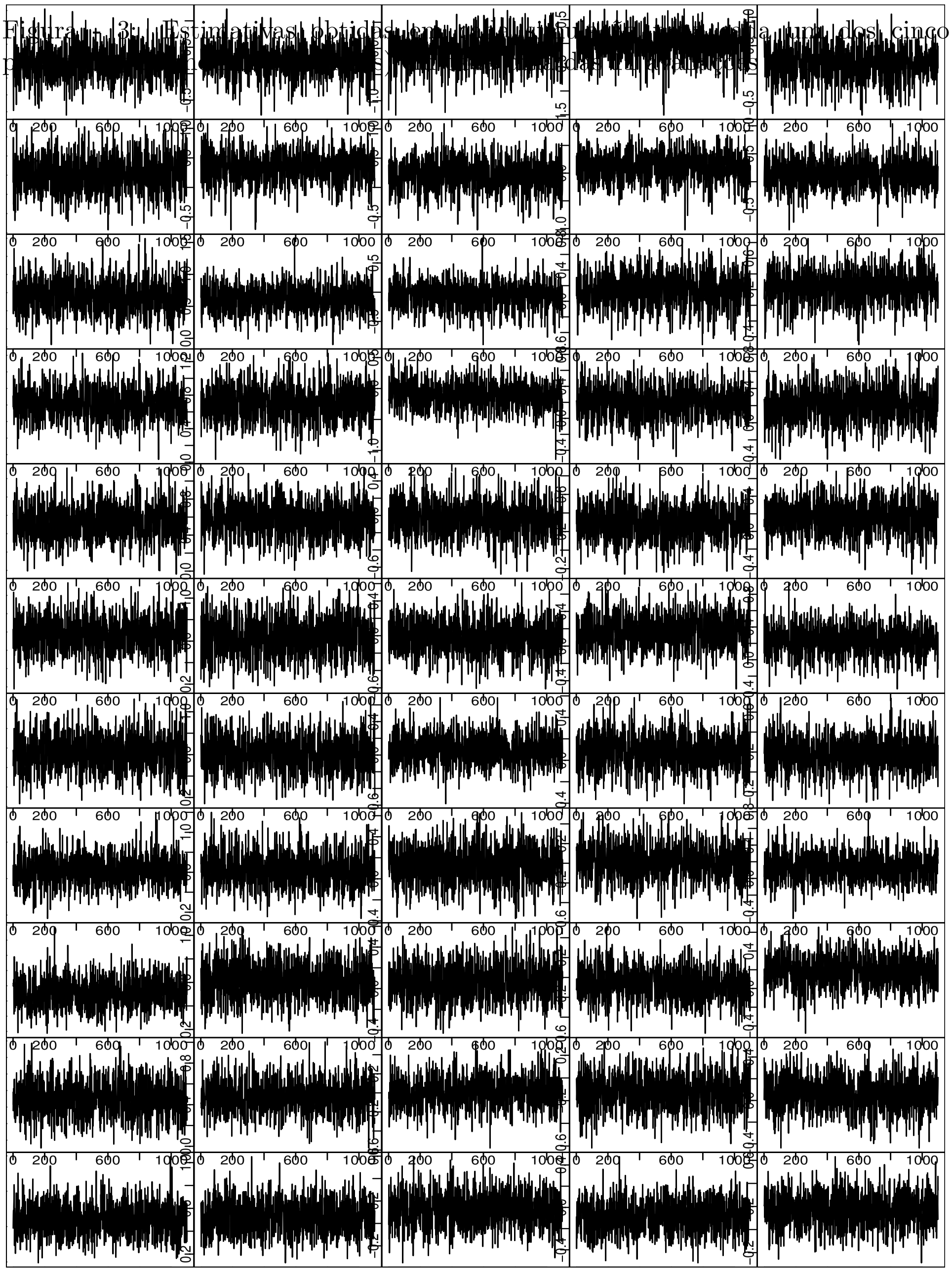


Figura 3: Estimativas obtidas em cada [illegible] cada um dos cinco [illegible]

Figura - 4: Densidades para cada um dos cinco parâmetros do modelo 1 (colunas) em cada uma das 11 avaliações (linhas).

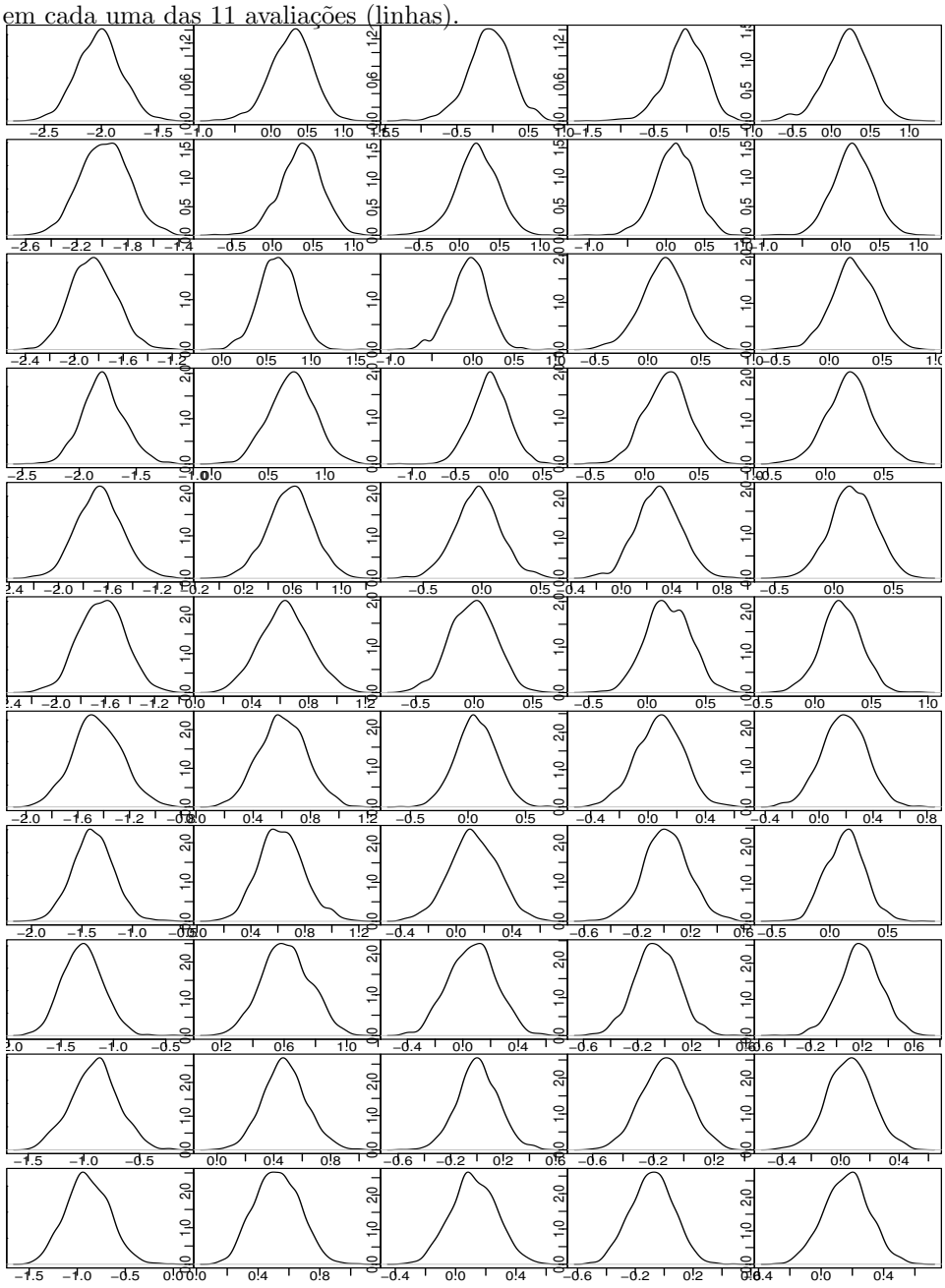


Figura - 5: Série das 9999 amostras de Monte Carlo para cada um dos cinco parâmetros do modelo 1 (colunas) em cada uma das 11 avaliações (linhas).

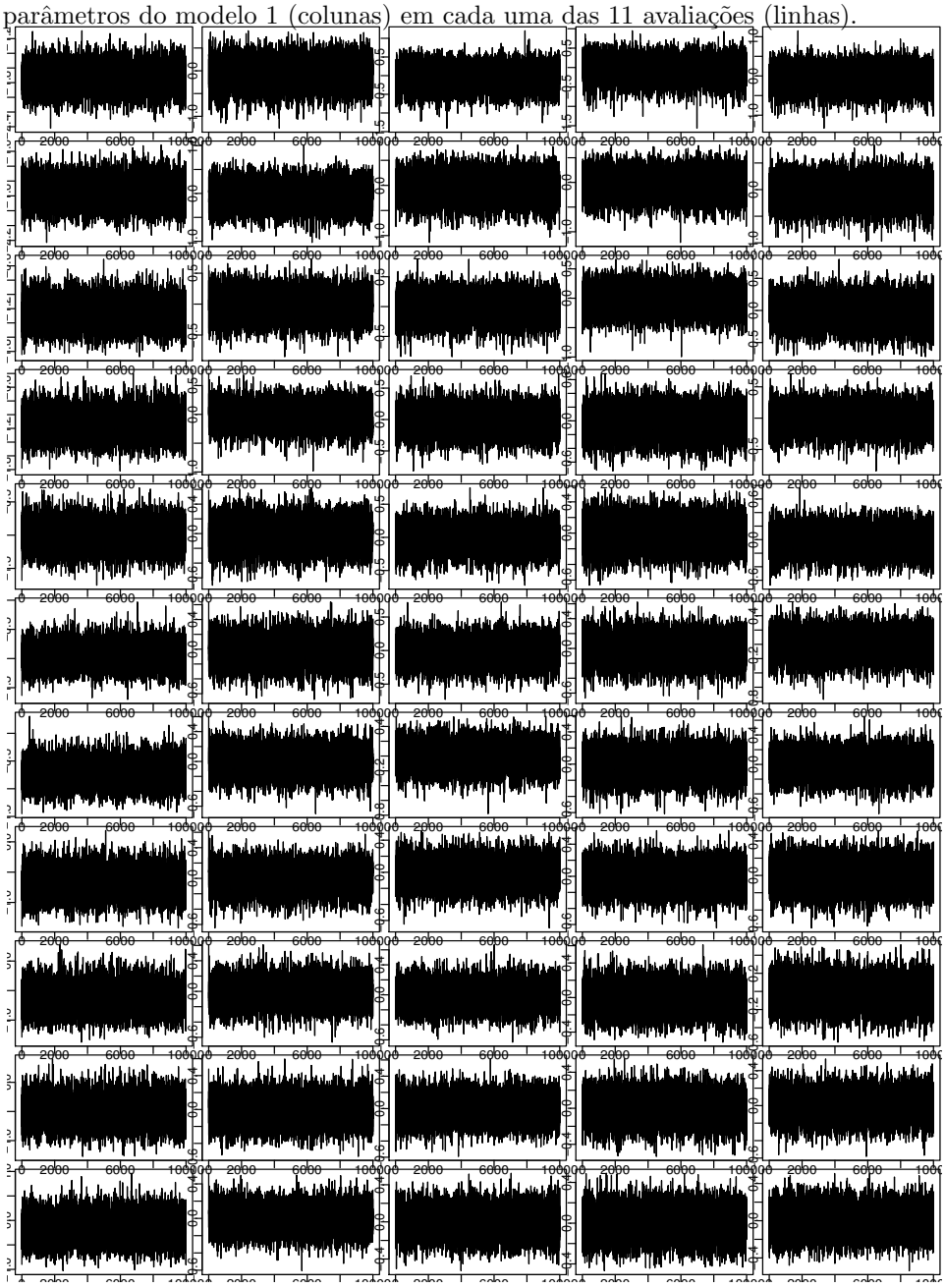


Figura - 6: Histogramas das 9999 amostras de Monte Carlo para cada um dos cinco parâmetros do modelo 1 (colunas) em cada uma das 11 avaliações (linhas) e respectivas estimativas (linha vermelha)