LUZIANE FRANCISCON

Modelo autologístico aplicado a dados de citrus

CURITIBA

JUNHO 2005

LUZIANE FRANCISCON

Modelo autologístico aplicado a dados de citrus

Trabalho de Conclusão de Curso apresentado na disciplina de Laboratório de Estatística II do Curso de Estatística do Departamento de Estatística, Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientador: Paulo Justiniano Ribeiro Junior

CURITIBA

JUNHO 2005

Termo de Aprovação

LUZIANE FRANCISCON

Modelo autologístico aplicado a dados de citrus

Trabalho de Conclusão de Curso apresentado na disciplina de Laboratório de Estatística II do Curso de Estatística do Departamento de Estatística, Setor de Ciências Exatas, Universidade Federal do Paraná, aprovado pela seguinte banca examinadora:

Prof. PhD. Paulo Justiniano Ribeiro Junior
(Orientador)
Universidade Federal do Paraná

Prof. PhD. Ricardo Sandes Ehlers
Universidade Federal do Paraná

Curitiba, 30 de junho de 2005

Agradecimentos

A Deus, pela vida.

Aos meus pais, pelo carinho e incentivo.

Ao prof. Paulo Justiniano Ribeiro Junior pela orientação e motivações.

Ao prof. Ricardo Sandes Ehlers pelas discussões e sugestões e aos demais professores do departamento pelos conhecimentos passados.

Aos colegas de turma pela amizade e apoio durante o curso.

Ao Renato Beozzo Bassanezzi, pesquisador do Fundo de Defesa da Citricultura (FUNDECITRUS) (`http://www.fundecitrus.com.br`) por ter disponibilizado os dados de *Morte Súbita dos Citrus*.

Ao Laboratório de Estatística e Geoinformação (LEG) (`http://www.est.ufpr.br/LEG`), pelos recursos computacionais ao colega Elias Teixeira Krainski pelo auxílio com a implementação computacional.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Neste trabalho foi estudado o modelo autologístico para dados binários de lattices. É feita uma revisão de literatura e as principais abordagens para a estimação e inferência. É descrita a metodologia de estimação do modelo autologístico baseada em pseudo-verossimilhança e o método de reamostragem bootstrap para estimação dos erros-padrão das estimativas. O modelo autologístico é aplicado em um conjunto de dados da literatura e em um conjunto de dados de Morte Súbita dos Citrus.

Palavras-chave: Estatística Espacial; Modelo Autologístico; Morte Súbita dos Citrus.

Abstract

We study the autologistic model for binary lattice data. The literature revision and main approaches for estimation and inference is made. We describe the methodology of estimation of the autologistic model based on pseudo-likelihood estimation and parametric bootstrap standard errors. The autologistic model is applied to a data set of literature and in a data set of Sudden Death of the Citrus.
Key-words: Spatial Statistics; Autologistic Model; Citrus Sudden Disease.

INTRODUÇÃO

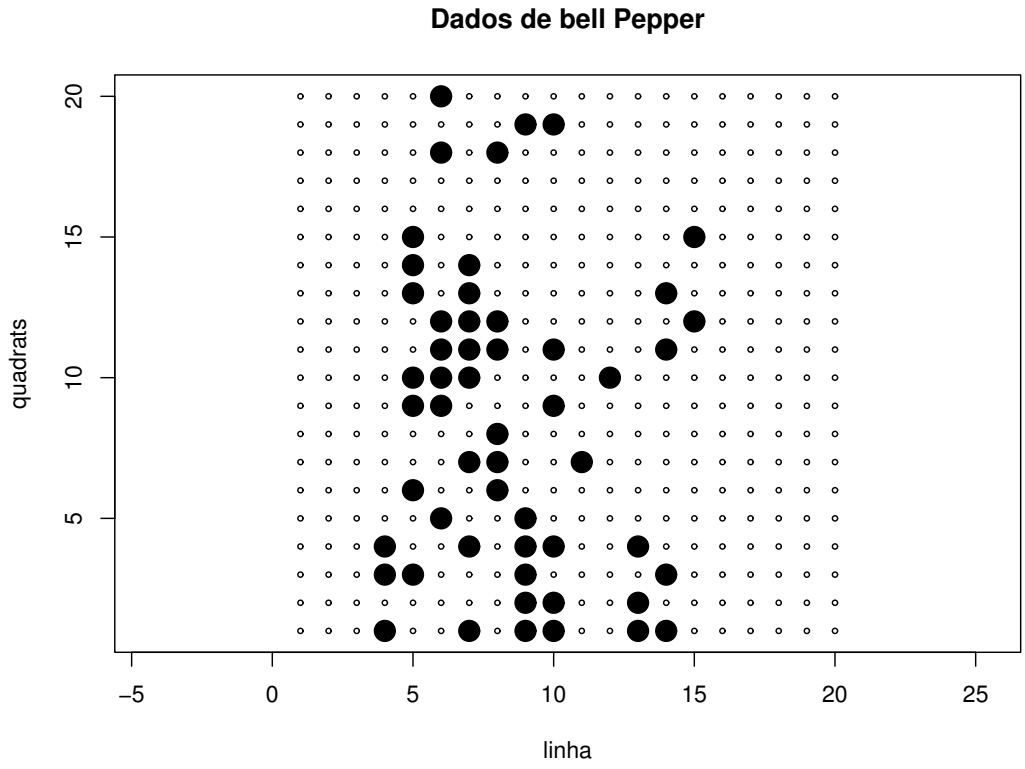
É razoável assumir que de modo geral os fenômenos observados em locais geograficamente mais próximos tendem a ser mais semelhantes do que se forem observados em locais mais distantes. Nesse caso, diz-se que existe dependência espacial no fenômeno em estudo. Em geral a correlação entre as observações diminui a medida que se aumenta a distância entre elas.

Em alguns estudos, a dependência espacial pode ser considerada de diferentes formas. Um exemplo são os experimentos onde quer se estudar a incidência de doenças em plantas. Nesse caso, a dependência espacial pode ser avaliada considerando: a localização da ocorrência das plantas doentes como um processo pontual; a probabilidade de uma planta ficar doente como uma medida contínua na região e usar um modelo de campo aleatório; ou, modelar a ocorrência da doença em uma planta ou *quadrat* como combinação linear da presença ou não da doença em observações vizinhas.

Nesse contexto, a análise da incidência de doenças pode ser feita por um modelo autoregressivo. Considerando-se o *status* da planta como resposta binária, o modelo será denominado autologístico, onde os coeficientes de regressão dão a estimativa do acréscimo na probabilidade da presença ou não da doença. Na estrutura de vizinhança consideram-se observações adjacentes em uma mesma fileira ou em fileiras próximas. Considerando vizinhos da mesma fileira e fileiras vizinhas separadamente, obtêm-se o grau da dispersão da doença em diferentes direções. A autocorrelação é evidentemente induzida, pois a mesma informação é utilizada como resposta e covariável, (GUMPERTZ M. L. ; GRAHAM; RISTAINO, 1997).

No Capítulo 1 será apresentada a revisão de literatura sobre modelos lineares generalizados, análises de diagnóstico e o modelo de regressão logística. Na Capítulo 2 será apresentado o modelo autologístico para modelar a dependência espacial em dados binários, a estrutura de vizinhança utilizada e os métodos computacionalmente intensivos utilizados para a estimação dos erros das estimativas dos parâmetros. Na Capítulo 3 aplica-se o modelo autologístico em um conjunto

Figura 1: Exemplo de dados de láttice. Dados de incidência de doenças em pimentas de sino.



de dados de incidência de determinada doença em pimentas de sino (GUMPERTZ M. L. ; GRAHAM; RISTAINO, 1997). Na Capítulo 4 é feita a aplicação do modelo autologístico em dados de *Morte Súbita dos Citrus* (NADA,).

Nos Apêndices são listadas as funções implementadas em R (R Development Core Team, 2005) e a forma de uso das mesmas, os dados e os códigos utilizados para as análises. Estas funções foram adicionadas ao pacote **Rcitrus**, disponível para *download* em `http://www.est.ufpr.br/Rcitrus`.

1 MODELOS LINEARES GENERALIZADOS

Os Modelos Lineares Generalizados (GLM'S), também denominados modelos exponenciais lineares, foram introduzidos por NELDER e WEDDERBURN (1972). A idéia básica consiste em diferentes opções para a distribuição da variável resposta, assumindo que a mesma pertença a família exponencial de distribuições e tenha maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear.

1.1 Especificação do Modelo

A estrutura de um GLM é dada por três partes:

- uma componente aleatória composta de uma variável aleatória Y com n observações independentes de uma distribuição pertencente à família exponencial e um vetor de médias μ;
- uma componente sistemática composta por variáveis explicativas $x_1, ..., x_p$ tais que, produzem um preditor linear;
- e uma função monótona diferenciável, conhecida como função de ligação, que relaciona as duas componentes.

Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias independentes, cada uma com função de densidade ou de probabilidade na família exponencial escrita como:

$$f(y_i; \theta_i, \phi) = \exp[\phi\{y_i\theta_i - b(\theta_i) + c(y_i)\} + a(y_i, \phi)] \ , \tag{1.1}$$

onde $\mathrm{E}\{Y_i\} = db(\theta_i)/d\theta_i$, que denotaremos por μ_i, $\mathrm{var}\{Y_i\} = \phi^{-1}V_i$, $V_i = d\mu_i/d\theta_i$ é a função de variância, sendo $i = 1, 2, \cdots, n$, $\theta = \theta(\beta)$ é o parâmetro canônico e ϕ é o parâmetro de dispersão.

Os modelos lineares generalizados são definidos por (1.1) e pela componente sistemática,

$$g(\mu_i) = \eta_i \ , \tag{1.2}$$

onde $\eta_i = x_i\beta$ é o preditor linear, $\beta = (\beta_1, \cdots, \beta_p)^\top$, $p < n$, é o vetor dos parâmetros da regressão a serem estimados, $x_i = (x_{i1}, \cdots, x_{ip})$ representa os valores de p variáveis explicativas e $g(\cdot)$ uma função monótona e diferenciável, denominada função de ligação.

As funções de ligação mais utilizadas são obtidas quando o parâmetro canônico coincide com o preditor linear, isto é, quando $\theta = \eta$ e a função de ligação nestas situações é chamada de ligação canônica. Nas ligações canônicas o preditor linear modela diretamente o parâmetro canônico, isto geralmente resulta em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão.

Como exemplos ou casos particulares de distribuições que pertencem à família exponencial (1.1) podemos mencionar a distribuição normal, normal inversa, Poisson e binomial, dentre outras. A partir dessas distribuições pode-se obter os os modelos de regressão normal inversa, regressão Poisson, regressão logística entre outros.

1.2 Modelo de Regressão Logística

O modelo de regressão logística é um modelo linear generalizado adequado para variáveis respostas binárias. Os dados assumem distribuição de probabilidade binomial, sob a forma de Y_i sucessos em m_i ensaios de Bernoulli, $i = 1, \cdots, n$. O modelo logístico é atrativo, devido a facilidade de interpretação dos parâmetros do modelo. A partir dos parâmetros, podemos medir o aumento na chance de sucesso com o acrescimo de uma unidade da covariável associada enquanto as demais permanecem constantes.

Uma particularidade do modelo de regressão logística diz respeito à natureza da relação entre a variável resposta e as covariáveis. A quantidade de interesse chamada média condicional é expressa por $E(Y|x)$, em que Y denota a variável resposta e x denota o vetor de valores das variáveis independentes. Neste caso é modelado o valor médio da variável resposta condicionado aos valores das covariáveis,

dentro do intervalos de valores $0 \leq E(Y|x) \leq 1$.

1.2.1 Especificação do Modelo

O modelo de regressão logística para $E(Y|x) = \theta(x)$ pode ser expresso por:

$$\theta(x) = P(Y = 1|x) = \frac{\exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}}{1 + \exp\{\beta_0 + \sum_{k=1}^{p} \beta_k x_k\}} \ , \tag{1.3}$$

em que $Y_i = 1$ representa a presença da resposta e zero caso contrário, x representa as covariáveis, $x = (x_1, \cdots, x_p)$, o parâmetro β_0 é o intercepto, e $(\beta_1, ..., \beta_p)$ são os p parâmetros de regressão.

Observe que como:

$$log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \ , \tag{1.4}$$

tem-se um modelo linear para o logito, isto é, o logaritmo Neperiano da razão entre $\theta(x)$ e $1 - \theta(x)$. No contexto de (GLM'S) $\eta = \log\left(\frac{\theta(x)}{1-\theta(x)}\right)$ é a função de ligação canônica para a regressão logística, outras ligações como probit e log-log, são usadas em alguns casos particulares.

1.2.2 Inferências do Modelo

A estimação dos parâmetros do modelo é usualmente feita pelo método de máxima verossimilhança. Assumindo que as observações são independentes tem-se a seguinte expressão para a função de verossimilhança:

$$L(\beta) = \prod_{i=1}^{n} (\theta(x_i)^{y_i})(1 - \theta(x_i))^{1-y_i} \ , \tag{1.5}$$

onde $\theta(x_i)$ é a contribuição para a função de verossimilhança dos pares (y_i, x_i) em que $y_i = 1$, e $1 - \theta(x_i)$ a contribuição dos pares em que $y_i = 0$.

1.2.2.1 Análise de Diagnóstico

Na análise de diagnóstico é feita a verificação das suposições de adequação do modelo aos dados, e a possível existência de observações extremas que podem gerar problemas no ajuste. Essa etapa compreende a análise de resíduos e a detecção de pontos de influência e de alavancagem, técnica que consiste em verificar a

dependência do modelo estatístico sobre as várias observações que foram coletadas e ajustadas.

Análise de Resíduos: Dentro da análise de diagnósticos a análise de resíduos ocupa um lugar importante. É usada para explorar a adequação do modelo ajustado com respeito à escolha da função de variância, da função de ligação e dos termos do preditor linear. Além disso, são úteis para indicar a presença de pontos aberrantes que poderão ser influentes ou não. Os resíduos medem as discrepâncias entre os valores observados $y_i's$ e os seus valores ajustados $\hat{\mu}_i's$.

Uma definição usual de resíduos em GLM'S são os *resíduos deviance* que, para a i-ésima observação é dado por:

$$r_{d,i} = d(y_i, \widehat{\mu}_i)^{1/2}\text{sign}(y_i - \widehat{\mu}_i) \ , \tag{1.6}$$

onde $d(y, \mu)$ é a função deviance, definida como:

$$d(y, \mu) = 2\{t(y, y) - t(y, \mu)\} \ , \tag{1.7}$$

sendo $t(y, y) = y\theta - b(\theta)$.

Uma forma de verificar a adequação do modelo através destes resíduos é comparar a soma de quadrados dos resíduos deviance com os graus de liberdade do resíduo. Para um bom ajuste do modelo aos dados é preciso que os valores sejam próximos. Caso o valor da soma de quadrados dos resíduos deviance seja superior ao número de graus de liberdade, há indicação de existência de superdispersão nos dados que deve ser considerada na modelagem.

Outra definição de resíduos são os chamados resíduos quantis propostos por DUNN e SMITH (1996). São baseados na idéia de inverter a função de distribuição estimada para cada observação e assim obter resíduos cuja função de densidade é exatamente a normal. São resíduos usados em modelos lineares generalizados em situações de grande dispersão quando os resíduos deviance podem não ser normais.

Seja $F(y; \mu, \phi)$ a função de distribuição acumulada de Y, com função de densidade $f(y; \mu, \phi)$. Dado que F é contínua, então $F(y; \mu, \phi)$ está uniformemente distribuída no intervalo (0,1). Neste caso os quantis residuais são definidos como:

$$r_{q,i} = \Phi^{-1}\{F(y_i; \widehat{\mu}_i, \widehat{\phi}_i)\} \cdot \ . \tag{1.8}$$

Nesta definição, Φ^{-1} é a função de distribuição acumulada inversa da normal padrão. Observemos que a distribuição de probabilidade de $r_{q,i}$ converge para uma normal padrão se $\widehat{\beta}$ e $\widehat{\phi}$ são estimadores consistentes de β e ϕ, respectivamente.

Se F não for contínua, uma definição mais geral dos resíduos quantis é necessária. Seja $a_i = lim_{y \to y_i} F(y; \widehat{\mu_i}, \widehat{\phi_i})$ e $b_i = F(y_i; \widehat{\mu_i}, \widehat{\phi_i})$. A definição de resíduo quantil para y_i é

$$r_{q,i} = \Phi^{-1}(u_i) \ , \tag{1.9}$$

onde u_i é uma variável uniforme no intervalo (a_i, b_i). Outras formas de resíduos são os resíduos de Pearson e os resíduos estudentizados.

2 O MODELO AUTOLOGÍSTICO

O modelo autologístico é um modelo autorregressivo espacial para dados binários. O modelo autorregressivo espacial considera a observação em determinado local dependente das observações em locais vizinhos em forma de covariável. Essa estratégia visa modelar a possível dependência espacial existente nos dados.

Neste capítulo será apresentado o modelo autologístico para dados de *lattices*, introduzido por BESAG (1972). Em um *lattice* os dados são observados em posições dispostas regularmente em linhas e colunas na região, ou seja, uma malha regular de pontos, um exemplo típico são *pixels* em uma imagem de satélite. Em dados de plantas o espaçamento entre linhas, em geral, é maior que o espaçamento das plantas dentro da linha.

Ao ajustar um modelo logístico a dados binários de *lattices*, os resíduos desse modelo apresentarão um padrão espacial se houver dependência espacial. Dessa forma, somente as covariáveis existentes no modelo não conseguem explicar a variabilidade dos dados. O modelo autologístico é uma forma de considerar a dependência espacial, modelando a probabilidade de sucesso, condicional às covariáveis de vizinhança. É necessário criar covariáveis de vizinhança que capturem esta dependência espacial, considerando alguma estrutura de vizinhança. Porém devido à autocorrelação os erros-padrão usuais do GLM não são válidos, o que gera problemas para a inferência sobre os parâmetros do modelo. Diversos métodos para a inferência foram propostos na literatura.

O modelo autologístico é flexível para fazer predição espacial de doenças em plantas. GUMPERTZ M. L. ; GRAHAM e RISTAINO (1997) usaram o modelo autologístico para estudar a probabilidade de um particular *quadrat* ter doença, a partir das informações de vizinhança. No artigo é descrito a aplicação do modelo em dados de pimentas de sino. Na estimação dos parâmetros do modelo foi usado o método de pseudo-verossimilhança. Os erros-padrão das estimativas dos parâmetros

das covariáveis de vizinhança foram calculados utilizando o método de reamostragem Bootstrap paramétrico. No procedimento de reamostragem foi utilizado o amostrador de Gibbs, para gerar campos com dependência espacial.

HE F. ; ZHOU e ZHU (2003) investigaram e compararam três métodos de estimação dos parâmetros: o método de pseudo-verossimilhança, proposto por BESAG (1977); o método de verossimilhança Monte Carlo introduzido por GEYER e THOMPSON (1992); e aproximação estocástica Monte Carlo cadeias de Markov proposto por GU e ZHU (2001), o qual faz aproximação estocástica para calcular a estimativa de máxima verossimilhança para modelos espaciais.

HUBBELL (2001) utilizou o modelo autologístico para estudar o efeito das informações vizinhas na sobrevida de árvores tropicais. A correção dos erros foi feita a partir de método Monte Carlo via cadeias de Markov utilizando amostrador de Gibbs, semelhante ao procedimento utilizado por GUMPERTZ M. L. ; GRAHAM e RISTAINO (1997).

FRIEL e PETTITT (2004) propõem um método eficiente para calcular a constante de normalização do modelo autologístico. Em particular, o método possibilita aproximar a verossimilhança verdadeira do modelo para estimação e inferência. A verdadeira verossimilhança é aproximada pelo produto das verossimilhanças para qual a constante de normalização pode ser encontrada através de um método computacional analítico que envolve a *lattice* em um cilindro.

2.1 Especificação do Modelo

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

$$P(Y = y_i|x_{ji}, y_{(i)}) = logit(y_i) = \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{k=1}^{q} \gamma_k y_{(i)} \ , \tag{2.1}$$

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

$$P(Y = y_{kl}|x_{kl}, y_{(kl)}) = logit(y_{kl}) = \sum_{j=1}^{p} \beta_j x_{kl} + \sum_{k=1}^{q} \gamma_k y_{(kl)} \ , \tag{2.2}$$

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

Uma *lattice* D é uma região com n posições, cada uma descrita pelas coordenadas (k, l), especificando a linha e coluna da *lattice* para os quais cada 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 covariáveis $x_{k,l}$. As respostas binárias de n posições correspondem $Y = (y_{k,l}, (k, l) \in D)$, constituem um mapa da ocupação da variável de interesse (HE F. ; ZHOU; ZHU, 2003).

A expressão do modelo autologístico é dada por

$$Pr(Y_{k,l} = y_{k,l}|x_{k,l}, y_{(k,l)}, (k, l) \in D) = \frac{exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y_{(k,l)}\}}{1 + exp\{\sum_{j=0}^{p} \beta_j x_{k,l} + \sum_{t=1}^{q} \gamma_t y_{(k,l)}\}}, \quad (2.3)$$

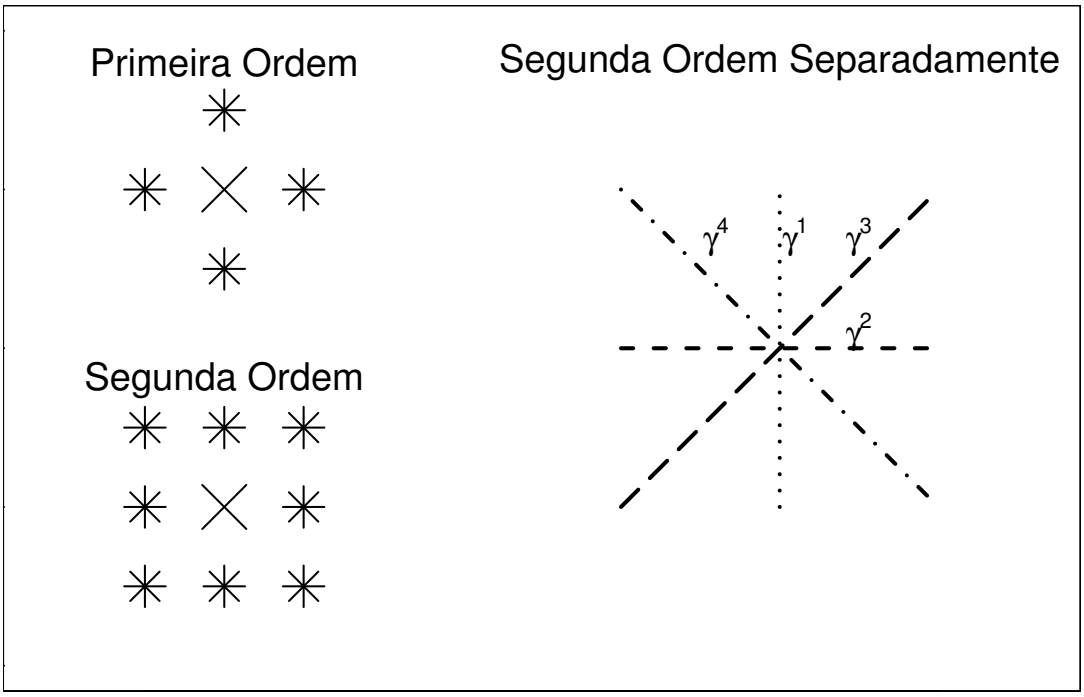
onde os $\beta' s$ são parâmetros de regressão e $\gamma' s$ são os parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariáveis e $y_{(k,l)}$ representam as covariáveis de vizinhança.

Estrutura de Vizinhança: Em dados de *lattices*, a estrutura de vizinhança pode ser construída de diferentes formas, considerando simplesmente a ordem ou considerando linhas e colunas separadamente. Na Figura 2 estão representadas algumas estruturas de vizinhança.

Considerando as ordens de vizinhança, uma possibilidade é construir uma covariável dentro de cada ordem. A covariável de primeira ordem é a soma das informações de quatro vizinhos, dois vizinhos na mesma linha em colunas adjacentes e dois vizinhos na mesma coluna em linhas adjacentes. E a covariável de segunda ordem, que é a soma das informações da covariável de primeira ordem, com as informações dos quatro vizinhos nas duas diagonais.

Outra maneira de construir covariáveis de vizinhança é considerar linhas, colunas e diagonais separadamente, dentro de determinada ordem. Na primeira ordem são construídas duas covariáveis. A primeira covariável inclui informações dos dois vizinhos na mesma linha em colunas adjacentes, posições $(k, l + 1)$ e $(k, l - 1)$. A segunda covariável inclui informações dos dois vizinhos na mesma coluna em linhas adjacentes, posições $(k - 1, l)$ e $(k + 1, l)$. Na estrutura de segunda ordem de vizinhança, somadas às covariáveis de primeira ordem uma terceira covariável, que inclui dois vizinhos da diagonal $(-1, 1)$, posições $(k - 1, l + 1)$ e $(k + 1, l - 1)$. E

Figura 2: Estruturas de vizinhança: primeira ordem, segunda ordem e segunda ordem separadamente



a quarta covariável inclui dois vizinhos da diagonal $(1,1)$, posições $(k-1, l-1)$ e $(k+1, l+1)$. Esta estrutura mais flexível pode ser particularmente útil no caso com doenças de plantas que possuem espaçamento diferente entre linhas e colunas.

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. Isto se deve ao fato de que essa função assume que a probabilidade condicional de uma observação $s_{k,l}$ é independente da probabilidade condicional de outra observação $v_{k,l}$. Isso pode ser verdade se as observações estiverem suficientemente distantes uma da outra, mas isso não é verdadeiro se as observações $s_{k,l}$ e $v_{k,l}$ forem vizinhas. Ou seja, as respostas observadas são espacialmente correlacionadas e as observações não são independentes, não é possível escrever a função de máxima verossimilhança.

Sugere-se 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 $\theta = (\beta_0, \beta_1^T, \gamma_1, \gamma_2, \gamma_3, \gamma_4)^T$

Mas os erros-padrão das estimativas dos coeficientes γ's não são apropriados para dados correlacionados, pois são subestimados. Neste caso, é preciso um procedimento para reestimar os erros-padrão das estimativas. O procedimento aqui proposto é o método de Bootstrap paramétrico.

Método de Bootstrap: O método de Bootstrap é utilizado para estimar usando simulação da distribuição amostral de uma estatística. Esse método consiste em gerar amostras dos dados originais para estimar a quantidade de interesse. O bootstrap paramétrico, é utilizado quando se tem informação suficiente sobre a forma da distribuição dos dados. A amostra bootstrap é formada realizando-se amostragem dessa distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. A distribuição da estatística de interesse aplicado aos valores da amostra bootstrap, condicional aos dados observados, é definida como distribuição bootstrap dessa estatística.

Na aplicação deste trabalho, os passos do bootstrap paramétrico consistem em simular N *lattices* $(\tilde{y}^1, ..., \tilde{y}^N)$ dado as estimativas dos parâmetros de vizinhança do modelo ajustado para os dados observados, γ_t^0. Obter a estimativa de γ^t, para cada $\tilde{y}^t$ simulado.

Mas, devido ao fato de que cada y_i^t está condicionado a posição original dentro do *lattice*, é preciso preservar esta condição em cada amostra simulada. Por isso o procedimento de amostragem utiliza o amostrador de Gibbs, para que a estrutura espacial fique preservada.

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 outros. Essas distribuições são conhecidas como "distribuições condicionais

completas". Estatísticas resumo empíricas, podem ser obtidas dessas amostras e usadas para retirada de inferências sobre seus valores verdadeiros.

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 a 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 covariáveis 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.

Os N *lattices* simuladas são correlacionadas, mas a variância converge para a verdadeira variância com o aumento do tamanho da cadeia. Isto porque o amostrador de Gibbs produz uma seqüência de *lattices* e uma correspondente seqüência de estimativas de parâmetros por pseudo-verossimilhança que contempla a escala inteira de valores possíveis (chamados cadeia de Markov). Desta maneira, a contribuição das correlações entre as *lattices* tornam-se insignificantes enquanto aumenta o tamanho da cadeia.

3 APLICAÇÃO À DADOS DE PIMENTAS DE SINO

O modelo autologístico pode ser aplicado em análises para predizer presença ou ausência de doenças de plantas com padrão espacial. Modela a probabilidade de determinada planta estar ou não doente dependendo das covariáveis e do status da doença nas plantas vizinhas.

Neste trabalho reproduzimos as análises dos dados do campo 1 apresentados por GUMPERTZ M. L. ; GRAHAM e RISTAINO (1997). No estudo foram medidas duas covariáveis do solo, umidade e densidade do patógeno, e construídas covariáveis de vizinhança. O campo 1 forma um *lattice* de 20 linhas por 20 *quadrats* com duas ou três plantas por *quadrat*. A variável resposta de interesse foi a presença ou ausência da doença no *quadrat*. O *quadrat* apresentava doença, se alguma das plantas dentro dele apresentasse lesões da doença ou estivesse morta.

A análise consistiu em ajustar três diferentes modelos de regressão logística. O modelo 1 considerou a umidade do solo e a densidade do patógeno no solo, ignorando a correlação espacial. Esse modelo foi preliminar para checar se apenas as covariáveis do solo eram suficientes para modelar a presença da doença. Se a correlação espacial ainda estivesse presente nos resíduos após ajustado o Modelo 1 concluí-se que somente as covariáveis do solo não eram suficientes para explicar a variabilidade observada nos dados.

Em alguns ajustes a correlação espacial pode ser completamente eliminada somente pela regressão nas covariáveis. Na presente aplicação porém, a doença propaga-se de uma planta para outra, e o status da doença dos *quadrats* vizinhos podem ser importante para predizer a presença da doença.

O Modelo 2 é um modelo autologístico de segunda-ordem com as covariáveis do solo. Este modelo foi construído adicionando covariáveis de vizinhança no

Modelo 1. O Modelo 3 é autologístico sem as covariáveis do solo, a predição está baseada somente na doença em *quadrats* vizinhos.

Os três modelos sugeridos pelo artigo são,

$$
\begin{aligned}
Modelo\,1 &: logit(p_{kl}) : \beta_0 + \beta_1 M_{kl} + \beta_2 L_{kl} \ , \\
Modelo\,2 &: logit(p_{kl}) : \beta_0 + \beta_1 M_{kl} + \beta_2 L_{kl} + \gamma_1 W_{kl} + \gamma_2 A_{kl} + \gamma_3 D_{kl1} + \gamma_4 D_{kl2} \ , \\
Modelo\,3 &: logit(p_{kl}) : \beta_0 + \gamma_1 W_{kl} + \gamma_2 A_{kl} + \gamma_3 D_{kl1} + \gamma_4 D_{kl2} \ .
\end{aligned}
$$

Nos três modelos, M=umidade do solo, L=densidade de patógeno no solo, o subescrito k e l indica linha e *quadrat*, respectivamente. O número de vizinhos doentes está indicado pelas covariáveis: W_{kl} são vizinhos das posições $(k, l-1)$ e $(k, l+1)$, A_{kl} são vizinhos das posições $(k-1, l)$ e $(k+1, l)$ e as diagonais D_{kl1} das posições $(k-1, l-1)$ e $(k+1, l+1)$, D_{kl2} nas posições $(k-1, l+1)$ e $(k+1, l-1)$. Observe que foram incluídas quatro covariáveis de vizinhança, que permitem examinar se a correlação dentro da linha é maior que entre as linhas e se existe alguma diagonal significativa na dispersão espacial da doença. Todos os modelos foram estimados com *lattice* 16×16 no total de 256 *quadrats*, desconsiderando duas linhas e dois *quadrats* nas bordas.

Na Tabela 1 estão apresentados os resultados dos três modelos ajustados, com os coeficientes de regressão estimados pela pseudo-verossimilhança.

Tabela 1: Estimativas dos parâmetros do modelo descrito do artigo.

Modelo	$\hat{\beta}_0$	Umidade do solo	Densidade patógeno	Dentro Linha	Entre Linhas	Diagonal (1,1)	Diagonal (-1,1)
1	-2.70	.0918	.178				
2	-3.91	.103	.0646	1.33	-0.739	.805	1.08
3	-2.81			1.36	-0.670	.730	1.09

Comparados os três modelos através do AIC (*Akaike Information Criterion)*, o modelo que melhor se ajustou aos dados do campo 1 foi o Modelo 3,

$$Modelo\,3 : logit(p_{kl}) : 2.81 + 1.36 W_{kl} - 0.670 A_{kl} + 0.730 D_{kl1} + 1.09 D_{kl2} \ .$$

Com as estimativas do modelo foi possível obter a chance de determinado *quadrat* apresentar doença, aproximadamente quatro vezes maior se um vizinho

dentro da linha estava doente, do que se os vizinhos estivessem livres da doença. Chegou-se a este resultado através da estimativa do parâmetro de γ_1 do Modelo 3 descritos na Tabela 1 $\hat{\gamma_1} = 1.36$, que corresponde a uma razão de chances $e^{\hat{\gamma_1}} = 3.9$ mantendo-se as outras covariáveis constantes.

Para este trabalho foi essa metodologia foi toda implementada programando-se funções no programa estatístico R (R Development Core Team, 2005). As funções são fornecidas no Apêndice A; a documentação é fornecida no Apêndice B. Versões mais atualizadas podem ser encontradas em `http://www.est.ufpr.br/Rcitrus`.

Na Tabela 2 mostramos os resultados dos erros-padrão das estimativas dos parâmetros obtidos na artigo. Os resultados que obtivemos considerando bordas de tamanho 2, como apresentado no artigo, e também usando bordas de tamanho 1. Nas estimativas do erro-padrão por bootstrap foram feitas obtendo-se 550 simulações e das quais as primeiras 50 simulações foram desconsideradas.

Tabela 2: Erros-padrão bootstrap e erros-padrão do modelo, usando borda de tamanho 1 e 2.

	$\hat{EP}(\beta_0)$	$\hat{EP}(\gamma_1)$	$\hat{EP}(\gamma_2)$	$\hat{EP}(\gamma_3)$	$\hat{EP}(\gamma_4)$
EP pseudo (bor=2)	0.32	0.32	0.42	0.36	0.30
EP pseudo (bor=2)	0.32	0.32	0.42	0.36	0.30
EP pseudo (bor=1)	0.29	0.27	0.37	0.33	0.27
EP bootstrap (bor=2)	0.40	0.55	0.69	0.53	0.54
EP bootstrap (bor=2)	0.36	0.51	0.69	0.56	0.51
EP bootstrap (bor=1)	0.33	0.50	0.68	0.62	0.57

Observa-se que as estimativas do erro-padrão que obtivemos por pseudo-verossimilhança com borda de tamanho 2, foram as mesmas obtidas pelo artigo. Quando usamos bordas de tamanho 1 as estimativas foram menores, pois neste caso utilizamos uma *lattice* de 18×18. Nas estimativas obtidas por bootstrap, quando utilizamos borda tamanho 2 os resultados foram semelhantes. Quando utilizamos borda de tamanho 1 obtemos estimativas mais precisas, tanto para γ quanto para o erro-padrão, pois são utilizados mais dados.

4 APLICAÇÃO À DADOS DE CITRUS

O Estado de São Paulo sozinho, responde por 80% da produção citrícola nacional, seus pomares apresentam baixa variabilidade com aproximadamente 85% das laranjeiras doces enxertadas sobre limoeiro cravo, o que gera elevada vulnerabilidade da cultura à ocorrência de novas epidemias.

A *Morte Súbita dos Citrus* (MSC), é uma nova doença dos citros, que provoca rápido definhamento e morte de plantas enxertadas em limoeiro *Cravo*. O primeiro relato oficial da doença foi realizado em fevereiro de 2001 no município de Comendador Gomes, MG, em janeiro deste mesmo ano, já eram observado 86% de plantas definhando e mortas de um pomar de 4703 plantas neste mesmo município. A MSC ocorre com maior intensidade em plantas enxertadas em limoeiro *Cravo*.

A importância desta doença é devida a representatividade desse porta-enxerto na citricultura brasileira, pela rusticidade, vigor à copa e resistência às deficiências hídrica no norte e nordeste do estado que São Paulo. Até o momento pouco se sabe sobre a etiologia da doença, apenas que se conseguiu a transmissão da doença de uma planta para outra. Também não se tem garantias que a doença ficará restrita ao norte do Estado de São Paulo ou se caminhará para a região central e mais importante do Estado.

Neste trabalho foram analisados dados de incidência de MSC em um talhão da Fazenda Vale Verde, no município de Comendador Gomes, MG. O talhão tem 20 linhas com 48 plantas em cada linha, o espaçamento dentro da linha é de 4 metros e entre as linhas é de 7,5 metros. A variável resposta de interesse é a presença ou ausência de MSC nas plantas, consideradas plantas doentes aquelas que apresentassem algum sintoma da doença ou plantas mortas pela doença. Os dados analisados são de 11 avaliações feitas de 05/11/2001 até 07/10/2002. A incidência da doença variou de 14,9% na primeira avaliação até 45,73% na 11^o avaliação.

Na modelagem estatística usada construímos um modelo autologístico considerando a estrutura de vizinhança de segunda ordem descrita na Seção 2.1, formando quatro covariáveis de vizinhança. Para saber se existe padrão espacial, é testado a hipótese no modelo ajustado.

$$H_0 = \hat{\gamma}_{kl} = 0 \ .$$

se esta hipótese for verdadeira para todos as estimativas dos parâmetros γ, então isso indica que não existe padrão espacial nos dados, se em alguma das estimativas a hipótese for rejeitada, devemos estudar a correlação espacial nos dados.

Tabela 3: Data, incidência e estimativas dos parâmetros dos modelos ajustados para as 11 avaliações.

Data	Incidência	$\hat{\beta}_0$	Dentro Linha	Entre Linhas	Diagonal (1,1)	Diagonal (-1,1)
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 estrutura de vizinhança adotada permite verificar se existe correlação espacial entre plantas na mesma linha, entre as linhas ou em alguma diagonal.

O modelo proposto para cada uma das 11 avaliações foi,

$$Modelo : logit(p_{kl}) : \beta_0 + \gamma_1 L_{kl} + \gamma_2 C_{kl} + \gamma_3 D_{kl1} + \gamma_4 D_{kl2} \ .$$

Com o uso das funções e os códigos dos Apêndices, foi ajustado o modelo autologístico para cada uma das 11 avaliações. Na Tabela 3 estão as datas, a incidência e as estimativas pontuais obtidas por pseudo-verossimilhança para cada um dos parâmetros, nas 11 avaliações.

Além das estimativas dos parâmetros precisamos precisamos obter os erros-padrão dessas estimativas para testar a hipótese da significância de cada um dos parâmetros. Na Tabela 4 estão as estimativas obtidas por pseudo-verossimilhança

dos erros-padrão das estimativas dos parâmetros.

Tabela 4: Erros-padrão obtidos por pseudo-verossimilhança

Data	$\hat{EP}(\beta_0)$	$\hat{EP}(\gamma_1)$	$\hat{EP}(\gamma_2)$	$\hat{EP}(\gamma_3)$	$\hat{EP}(\gamma_4)$
05/11/01	0.15467	0.19984	0.21252	0.21083	0.19412
05/12/01	0.15308	0.17258	0.17523	0.17753	0.17102
04/01/02	0.15104	0.14094	0.15408	0.15068	0.14586
13/02/02	0.14965	0.13035	0.14456	0.14114	0.13955
14/03/02	0.14906	0.12400	0.13201	0.13048	0.13034
05/04/02	0.15014	0.12088	0.12765	0.12703	0.12473
08/05/02	0.15552	0.11280	0.11786	0.11750	0.11756
03/06/02	0.15501	0.11228	0.11633	0.11694	0.11542
06/07/02	0.15711	0.11135	0.11466	0.11546	0.11325
06/09/02	0.16364	0.10712	0.10848	0.11085	0.10819
07/10/02	0.17197	0.10108	0.10063	0.10375	0.10363

Na Tabela 5 estão as estimativas obtidas por reamostragem bootstrap dos erros-padrão das estimativas dos parâmetros, com 1100 simulações e descartadas as 100 primeiras.

Tabela 5: Erros-padrão obtidos por bootstrap

Data	$\hat{EP}(\beta_0)$	$\hat{EP}(\gamma_1)$	$\hat{EP}(\gamma_2)$	$\hat{EP}(\gamma_3)$	$\hat{EP}(\gamma_4)$
05/11/01	0.19422	0.29304	0.29814	0.30349	0.28006
05/12/01	0.18722	0.25284	0.25847	0.25426	0.25613
04/01/02	0.18115	0.20525	0.22291	0.21430	0.20696
13/02/02	0.18349	0.19918	0.21155	0.19625	0.19798
14/03/02	0.18163	0.18264	0.19417	0.17766	0.18960
05/04/02	0.18481	0.17240	0.19332	0.18756	0.17759
08/05/02	0.18510	0.16213	0.17422	0.16269	0.16444
03/06/02	0.20016	0.16374	0.17241	0.17010	0.16239
06/07/02	0.19248	0.15058	0.17022	0.15719	0.16110
06/09/02	0.21185	0.15399	0.15652	0.15091	0.15291
07/10/02	0.22392	0.14430	0.14845	0.14823	0.14282

Observa-se que as estimativas dos erros obtidos por pseudo-verossimilhança subestimam os erros quando comparadas com as estimativas obtidas por bootstrap paramétrico, mais realísticas, são maiores que as obtidas por pseudo-verossimilhança. Portanto adotamos as estimativas obtidas por bootstrap.

Para testar a dependência espacial, testamos a significância dos coeficientes associados as covariáveis de vizinhança ao nível de significância de 5%. Para isso

utilizamos o seguinte resultado assintótico,

$$\frac{\hat{\gamma}}{ep_{\hat{\gamma}}} \sim N(0,1) \; . \tag{4.1}$$

Na Tabela 6 apresentamos os p-valores obtidos utilizando os erros-padrão bootstrap.

Tabela 6: Tabela de p-valores para testar a hipótese de que os coeficientes são iguais a zero

Data	$\hat{\beta}_0$	Dentro da Linha	Entre as Linhas	Diagonal (1,1)	Diagonal (-1,1)
05/11/01	0.00000	0.27365	0.94062	0.98163	0.46181
05/12/01	0.00000	0.16735	0.37606	0.58584	0.52980
04/01/02	0.00000	0.00221	0.90968	0.44806	0.26035
13/02/02	0.00000	0.00036	0.63991	0.28060	0.29245
14/03/02	0.00000	0.00126	0.89330	0.08113	0.37826
05/04/02	0.00000	0.00025	0.97196	0.31854	0.17815
08/05/02	0.00000	0.00018	0.69008	0.56112	0.24626
03/06/02	0.00000	0.00014	0.44090	0.87295	0.42052
06/07/02	0.00000	0.00005	0.65054	0.70722	0.24503
06/09/02	0.00002	0.00190	0.94235	0.43899	0.64234
07/10/02	0.00006	0.00028	0.40094	0.59804	0.25456

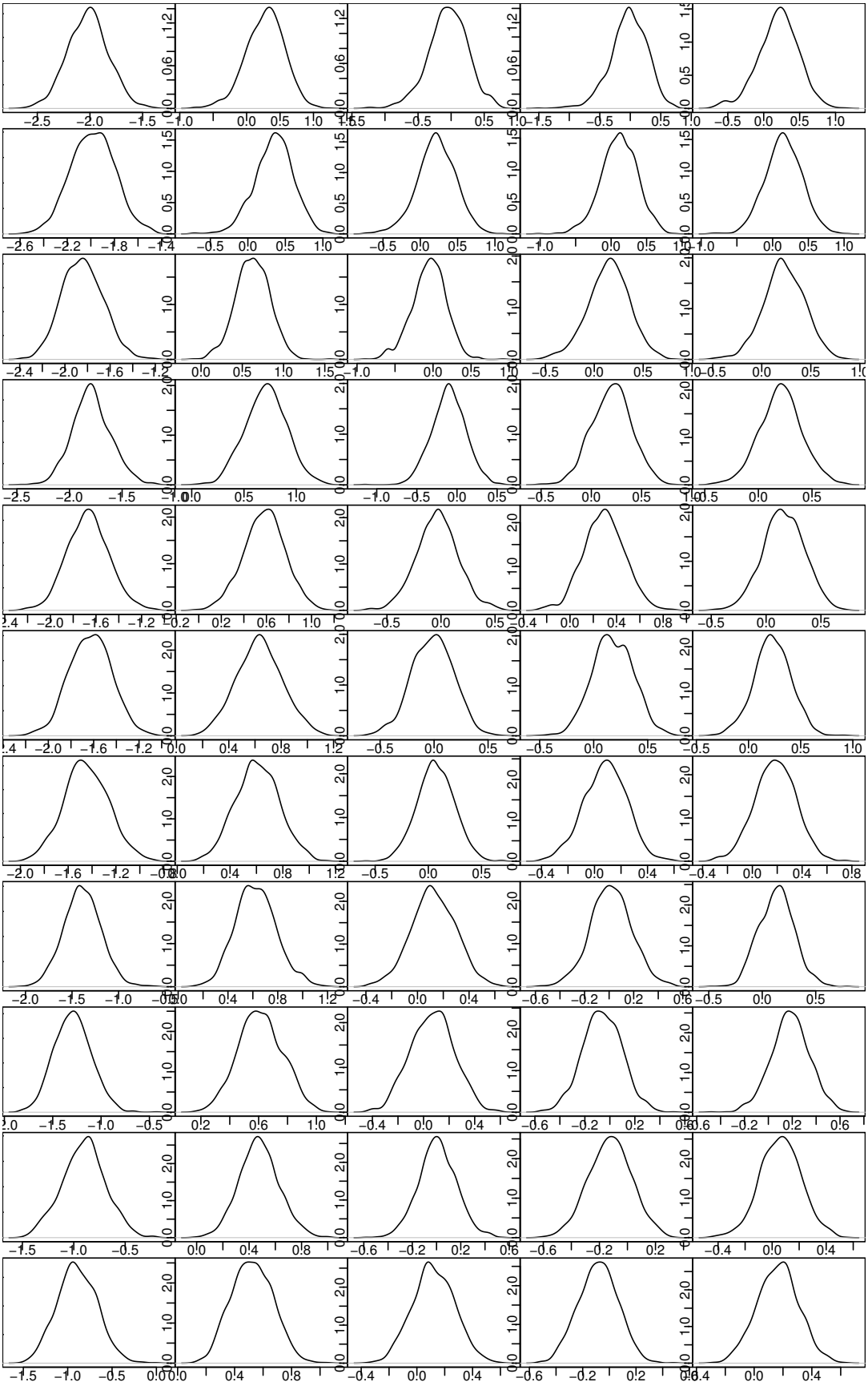
Analisando a Tabela 6 vemos que existe dependência espacial somente dentro da linha a partir da terceira avaliação. Esse resultado permite duas conclusões: a dependência espacial é de curto alcance, pois o espaçamento dentro da linha é de 4 metros e entre linhas é de 7,5 metros; e quando há baixa incidência 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 estão sadias. Utilizando o modelo para fazer predição, se considerar o status das plantas vizinhas dentro da linha, teremos maior precisão.

Na Figura 3 estão plotados os valores obtidos das 1100 simulações para cada uma das 11 avaliações da MSC analizadas e na Figura 4 estão plotadas as densidades.

Figura 3: Estimativas obtidas em cada simulação para cada um dos cinco parâmetros (colunas) em cada uma das 11 avaliações (linhas).

Figura 4: Gráficos das densidades para cada um dos cinco parâmetros (colunas) em cada uma das 11 avaliações (linhas).



5 CONCLUSÕES

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 a dependência espacial em várias direções. Na aplicação aos dados de *Morte Súbita dos Citrus*, isso permitiu modelar a dependência espacial dentro das linhas e entre linhas separadamente. Essa estratégia revelou-se importante, pois permitiu verificar que a dependência espacial ocorre dentro da linha. Esse resultado permite concluir que 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. Considerando as 11 avaliações analisadas, também pode-se concluir que, quando incidência é baixa, menor que 20% aproximadamente, não existe dependência espacial.

Outra vantagem do modelo autologístico é a possibilidade de modelar os dados originais, sem qualquer agrupamento com perda de informação, pois não é necessário agrupar os dados em *quadrats*. Nos métodos de análise baseados *quadrats*, além de haver perda de informação devida ao agrupamento por *quadrats*, há uma subjetividade na escolha do tamanho de *quadrats* a ser adotado.

Referências

FRIEL, N. and PETTITT, A. N. [S.l.].

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

BESAG, J. Efficiency of pseudo likelihood estimators for simple gaussian fields. *Biometrika*, 1977.

DUNN, P. K.; SMITH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical*, 1996.

FRIEL, N.; PETTITT, A. N. Likelihood estimation and inference for the autologistic model. *Journal of Computational and Graphical Statistics*, 2004.

GEYER, C. J.; THOMPSON, E. A. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistics Society, Series B*, 1992.

GU, M. G.; ZHU, H. T. Maximum likelihood estimation for spatial models by markov chain monte carlo stochastic approximation. *Journal of the Royal Statistics Society, Series B*, 2001.

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.

HUBBELL, S. e. a. Local neighborhood effects on long-term survival of individual trees in a neotropical forest. *Ecological Research*, 2001.

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 A – Funções

Funções em R utilizadas para a aplicação do modelo autologístico.

```
``extract.neigh'' <-
function (data, bor = 0, mat = FALSE, death = 1, healt = 0, id = NULL)
{
    if (!is.null(id))
        return(extract.n.neigh(data, mat = mat, death = death,
            healt = healt, id = id))
    bor <- as.numeric(bor)
    data <- as.matrix(data)
    dd <- dim(data)
    res <- list()
    res$disease <- data
    res$vrow <- cbind(0, data[, -dd[2]]) + cbind(data[, -1],
        0)
    res$vcol <- rbind(0, data[-dd[1], ]) + rbind(data[-1, ],
        0)
    res$vdiag1 <- rbind(0, cbind(0, data[-dd[1], -dd[2]])) +
        rbind(cbind(data[-1, -1], 0), 0)
    res$vdiag2 <- rbind(0, cbind(data[-dd[1], -1], 0)) + rbind(cbind(0,
        data[-1, -dd[2]]), 0)
    if (bor == 1)
        res <- lapply(res, function(x) x[-c(1, nrow(x)), -c(1,
            ncol(x))])
    if (bor == 2)
        res <- lapply(res, function(x) x[-c(1:2, (nrow(x) - 1):nrow(x)),
            -c(1:2, (ncol(x) - 1):ncol(x))])
```

```
    if (!mat)
        return(as.data.frame(sapply(res, function(x) as.numeric(x))))
    else return(res)
}
``autologistic.citrus'' <-
        function (data, bor = 1, bin = FALSE, death = 1, healt = 0, N = 10)
{
    logit <- function(eta) exp(eta)/(1 + exp(eta))
    fit.autologistic.citrus <- function(data, bor = 2, death = 1) {
        data <- extract.neigh(data, bor = bor, mat = FALSE, death = death)
        return(glm(disease ~ vrow + vcol + vdiag1 + vdiag2,
            data = data, family = binomial))
    }
    data <- as.matrix(check.citrus(data, death = death, healt = healt))
    dd <- dim(data)
    if (bor == 1)
        int.coords <- expand.grid(x = 2:(dd[1] - 1), y = 2:(dd[2] -
            1))
    if (bor == 2)
        int.coords <- expand.grid(x = 3:(dd[1] - 2), y = 3:(dd[2] -
            2))
    aj.ori <- fit.autologistic.citrus(data, bor = bor, bin = bin,
        death = death)
    print(coef(aj.ori))
    if (N > 0) {
        gam <- data.frame(matrix(0, N, 5))
        for (i in 1:N) {
            cat("sim", i, ": ")
            ord <- sample(1:nrow(int.coords))
            for (j in ord) {
                new <- as.data.frame(t(extract.1.neigh(data,
                  int.coords[j, ], death = death)))
                p <- logit(predict(aj.ori, newdata = new))
                if (bor == 1)
                  data[-c(1, dd[1]), -c(1, dd[2])][j] <- rbinom(1,
```

```
                    1, p)
                if (bor == 2)
                  data[-c(1:2, (dd[1] - 1):dd[1]), -c(1:2, (dd[2] -
                    1):dd[2])][j] <- rbinom(1, 1, p)
            }
            aj.i <- fit.autologistic.citrus(data, bor = bor,
                death = death)
            gam[i, ] <- coef(aj.i)
            cat("ok :", as.numeric(gam[i, ]), "\n")
        }
        row.names(gam) <- paste("sim", 1:N, sep = "")
        colnames(gam) <- names(coef(aj.ori))
        res <- list(pseudo = aj.ori, gamma.sim = gam)
        class(res) <- "autologistic"
        return(res)
    }
    else return(aj.ori)
}
``print.autologistic'' <-
function (x, ...)
{
    cat("Resultados da Pseudo-Verossimilhanca\n")
    cat("Coeficientes:\n")
    print(coef(x$pseudo))
    cat("Variancias:\n")
    print(diag(vcov(x$pseudo)))
    cat("Resultados da reamostragem bootstrap \n")
    cat("via Amostrador de Gibbs:\n")
    cat("Coeficientes:\n")
    print(apply(x$gamma.sim, 2, mean))
    cat("Variancias:\n")
    print(apply(x$gamma.sim, 2, var))
}
``plot.autologistic'' <-
function (x, y, ...)
```

```
{
    if (missing(y))
        y <- 1:nrow(x$gamma)
    par(mfrow = c(2, 3), mar = c(3, 3, 3, 1), mgp = c(2, 1, 0))
    for (i in 1:5)
      plot(x$gamma.sim[y, i], xlab = "Sim",
           ylab = colnames(x$gamma.sim)[i],
        type = "l", ...)
    return(invisible())
}
```

APÊNDICE B – Documentação das funções

Funções em R utilizadas para a aplicação do modelo autologístico.

`extract.neigh` *Extrai informacoes de vizinhos doentes*

Description

Extrai numero de plantas doentes vizinhas na linha, dentro da linha e nas diagonais separadamente.

Usage

```
extract.neigh(data, bor=0, mat=FALSE, death=1, healt=0, id=NULL)
```

Arguments

`data`	objeto da classe 'data.frame' ou 'matrix'
`bor`	inteiro que indica o numero de linhas de plantas/plantas na linha e entre linhas que serao tomadas como bordas
`mat`	logico que indica se a informacao sera retornada na forma de lista de matrizes ou um data.frame
`death`	codigo(s) que indicam planta doente
`healt`	codigo(s) que indicam planta sadia
`id`	indica a linha no talhao e planta na linha na qual sera calculado, id=NULL serao consideradas todas as plantas

Details

Se bor=1 ou bor=2, serao utilizadas as informacoes das bordas mas sera retornado apenas informacoes para plantas no interior.

Value

Se id=NULL, sera um data.frame contendo os dados na primeira coluna e as informacoes dos vizinhos na linha 'vrow', dentro da linha 'vcol', na diagonal (-1,1) 'vdiag' e na diagonal (1,1) 'vdiag2'. Se for informado a coordenada (linha, coluna) de apenas uma planta no argumento 'id', sera um vetor. Se for informada mais de uma planta no argumento 'id', sera um data.frame.

Author(s)

Elias T. Krainski and Paulo Justiniano Ribeiro Jr

See Also

`extract.1.neigh`, `extract.n.neigh` e `kdt.nneigh`

Examples

```
data(Itajobi)
dim(Itajobi)

cov <- extract.neigh(Itajobi)
dim(cov)

cov1 <- extract.neigh(Itajobi, bor=1)
dim(cov1)
```

`autologistic.citrus`

ajuste do modelo autologistico com correcao de erros

Description

Estima os parametros do modelo autologistico considerando informacao dos separadamente. A estimacao dos parametros e' feita por pseudo-verossimilhanca. A correcao dos erros e' feita utilizando-se reamostragem bootstrap parametrico via amostrador de Gibbs.

Usage

```
autologistic.citrus(data, bor = 1, bin = FALSE, death = 1, healt = 0,
                    N = 10)
```

Arguments

`data`	objeto da classe 'matrix' ou 'data.frame'
`bor`	inteiro que indica o numero de linhas de plantas/plantas na linha e entre linhas que serao tomadas como bordas
`bin`	logico que indica se a informacao sera retornada na forma binaria ou de contagem, veja detalhes
`death`	codigo(s) que indicam planta doente
`healt`	codigo(s) que indicam planta sadia
`N`	numero de amostras booststrap

Value

Se N=0, sera retornado apenas o modelo ajustado por pseudo-verossimilhanca para os dados. Se N>0, sera retornado uma lista com o ajuste para os dados originais e uma 'matrix' com 'N' linhas e 5 colunas com os resultados das estimativas obtidas em cada simulacao.

Author(s)

Elias T. Krainski, Luziane Franciscon and Paulo Justiniano Ribeiro Jr.

References

GUMPERTZ, M. L. ; GRAHAM, J. M. & RISTAINO, J. B. (1997). 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

See Also

`betabinom.citrus`, outro teste de completa aleatoriedade espacial em doencas de citrus use `disp.quadrats`, `mmdist.test` ou `Kenv.csr.citrus`.

Examples

```
data(bellPepper)
aut <- autologistic.citrus(bellPepper, N=10)
aut
plot(aut)
```

APÊNDICE C – Dados analisados

Dados analisados de Bell-pepper

```
> bellPepper
   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
2  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
3  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
4  1 0 1 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
5  0 0 1 0 0 1 0 0 1  1  0  0  1  1  1  0  0  0  0  0
6  0 0 0 0 1 0 0 0 1  1  1  1  0  0  0  0  0  1  0  1
7  1 0 0 1 0 0 1 0 0  1  1  1  1  1  0  0  0  0  0  0
8  0 0 0 0 0 1 1 1 0  0  1  1  0  0  0  0  0  1  0  0
9  1 1 1 1 1 0 0 0 0  0  0  0  0  0  0  0  0  0  1  0
10 1 1 0 1 0 0 0 0 1  0  1  0  0  0  0  0  0  0  1  0
11 0 0 0 0 0 0 1 0 0  0  0  0  0  0  0  0  0  0  0  0
12 0 0 0 0 0 0 0 0 0  1  0  0  0  0  0  0  0  0  0  0
13 1 1 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
14 1 0 1 0 0 0 0 0 0  0  1  0  1  0  0  0  0  0  0  0
15 0 0 0 0 0 0 0 0 0  0  0  1  0  0  1  0  0  0  0  0
16 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
17 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
18 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
19 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
20 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
```

Os dados analisados de *Morte Súbita dos Citrus* são muito extensos e não estão incluídos neste apêndice. Esses dados estão disponíveis no pacote **Rcitrus**, que encontra-se a disposição em `http://www.est.ufpr.br/Rcitrus`.

APÊNDICE D – Códigos para análise dos dados

Códigos em R utilizadas para a análise dos dados de Bell-Pepper com duas bordas.

```
data(bellPepper)
set.seed(123)
autBP.2bor <- autologistic.citrus(bellPepper, bor=2, N=550)
autBP.2bor
plot(autBP.2bor)
gam <- autBP.2bor$gam[-(1:50), ]
gam <- gam[!apply(!gam>-5, 1, any),]
apply(gam, 2, sd)
```

Códigos em R utilizadas para a análise dos dados de Bell-Pepper com uma borda.

```
data(bellPepper)
set.seed(123)
autBP.1bor <- autologistic.citrus(bellPepper, bor=1, N=550)
autBP.1bor
plot(autBP.1bor)
gam <- autBP.1bor$gam[-(1:50), ]
gam <- gam[!apply(!gam>-5, 1, any),]
apply(gam, 2, sd)
```

Códigos em R utilizadas para a análise dos dados de *Morte Súbita dos Citrus*.

```
data(vv202)
dados <- vv202[,,1:11]
dados[dados>1] <- 1
attributes(dados) <- list(dim=c(20,48,11),
                          dimnames=list(1:20, 1:48, 1:11))
dim(dados)
table(dados)
apply(dados, 3, table)
autol.11av <- apply(dados, 3, autologistic.citrus, bor=1, N=1100)
aut.reais <- autol.11av$pseudo
lapply(aut.reais, summary)
coe <- t(sapply(aut.reais, coef))
emod <- t(sapply(aut.reais, function(x) summary(x)$coef[,2]))
eboot <- t(sapply(autol.11av, function(x)
                  apply(x$gam[-(1:100), ], 2, sd)))
coe
emod
eboot
p.val <- 2*pnorm(abs(coe)/eboot, lower=F)
p.val
```