

UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA “LUIZ DE QUEIROZ”

PLANO DE DISSERTAÇÃO

CURSO DE PÓS-GRADUAÇÃO EM: Estatística e Experimentação Agronômica

NÍVEL: Mestrado

CANDIDATA: Luziane Franciscon

ORIENTADORA: Prof. Dr. Paulo Justiniano Ribeiro Júnior

TÍTULO: Modelo Autologístico: Uma aplicação a dados de citros

PIRACICABA
Estado de São Paulo - Brasil
Maior de 2007

SUMÁRIO

	Página
1 INTRODUÇÃO	1
2 OBJETIVOS	3
3 REVISÃO DE LITERATURA	4
3.1 Morte Súbita em Citros	4
3.2 Leprose dos Citros	5
3.3 Modelos Lineares Generalizados	5
3.3.1 Especificação do Modelo	6
3.4 Modelo de Regressão Logística	7
3.4.1 Especificação do Modelo	7
3.4.2 Inferência	8
3.5 Modelo Autologístico	9
3.5.1 Estrutura de Vizinhança	10
3.5.2 Especificação do Modelo	12
3.6 Inferência	13
3.6.1 Estimação por Pseudo-verossimilhança	13
3.6.2 Método de Bootstrap	14
3.6.3 Método de Monte Carlo	15
3.6.4 Método de Monte Carlo com Cadeias de Markov (MCMC)	16
3.6.4.1 Cadeias de Markov:	16
3.7 Dependência Temporal no Modelo Autologístico	17
3.7.1 Covariáveis de vizinhança no período de tempo anterior	17
3.7.2 Covariáveis de vizinhança no tempo atual e anterior	18
4 MATERIAL E MÉTODOS	19
4.1 Material	19
4.2 Métodos	20

	3
4.2.1 Especificação do modelo autologístico	20
4.2.2 Inferência para o modelo autologístico	22
REFERÊNCIAS BIBLIOGRÁFICAS	23

1 INTRODUÇÃO

A produção citrícola brasileira tem notável importância na economia nacional, sendo 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 transita pelo mercado internacional. Citricultores, indústrias e cientistas brasileiros criaram um setor de ponta na agroindústria nacional e 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 (TRIBESS; TADINI, 2001). Algumas doenças, como a Leprose dos Citros e a Morte Súbita dos Citros (MSC), podem causar a erradicação de talhões inteiros de plantas, na fazenda produtora, comprometendo a quantidade e a qualidade das frutas cítricas.

A MSC representa uma ameaça potencial para a citricultura paulista e nacional, uma vez que afeta todas as variedades comerciais enxertadas sobre limoeiro *Cravo*. Provoca diminuição de tamanho, peso e quantidade de frutos, além de rápido definhamento e morte das plantas (BASSANEZI; FERNANDES; YAMMAMOTO, 2003). A leprose dos citros tem sido considerada, na última década, a mais importante virose na citricultura nacional, pois reduz a produção e o período de vida das plantas de citros (RODRIGUES, 2000). Os dados de plantas cítricas são, em geral, coletados ao longo do tempo, as plantas estão dispostas dentro de um talhão, em linhas e colunas com certa estrutura de vizinhança. A disposição das plantas permite estudar o padrão espacial de doenças cítricas.

Em estudos para detectar o padrão espacial nos dados foram, inicialmente, propostas técnicas descritivas, como por exemplo: análise por quadrat, variograma, índices de associação. Mais recentemente além de detectar o padrão espacial dos dados, também foram introduzidos modelos estatísticos espaciais, tais como: o modelo CAR (*Conditional Auto Regressive*), o modelo estocástico espacial e o modelo autologístico. Enquanto que os dois primeiros modelos requerem técnicas de Monte Carlo com Cadeias de Markov (MCMC), o modelo autologístico possui simples interpretabilidade e incorpora explicitamente a estrutura de dependência espacial.

O modelo autologístico é flexível para modelar a incidência de doenças em

plantas, considera o *status* da planta como resposta binária, e através de covariáveis de vizinhança, modela a dependência espacial ou autocorrelação espacial dos dados. Os coeficientes de regressão dão a estimativa do acréscimo na probabilidade da presença, ou não, da doença. Na agricultura, esse modelo foi estudado para a análise da incidência de *Phytophthora* em pimentas de sino por Gumpertz, Graham e Ristiano (1997), também foi estudado por Czermainski (2006) para análise de incidência de leprose dos citros em um talhão de plantas cítricas e por Viola (2007) na análise de dados de um experimento com pimentões e a dados simulados.

Na estratégia de modelagem de dados reais algumas questões podem ser incorporadas no estudo quando ajustado o modelo autológico aos dados: a estrutura de covariáveis de vizinhança, a dimensão temporal dos dados e o efeito defasado das covariáveis de vizinhança.

2 OBJETIVOS

Devido ao interesse em modelar o padrão espacial da incidência de doenças em plantas cítricas, essa pesquisa tem por objetivos:

- i) Fazer uma revisão de literatura sobre o modelo autologístico, comumente utilizado na modelagem de doenças de plantas;
- ii) Ajustar o modelo autologístico a dois conjunto de dados de citros: um com incidência de MSC e outro com incidência de leprose dos citros;
- iii) Comparar diferentes métodos de estimação dos parâmetros (Pseudo-Verossimilhança, Método de Monte Carlo e Método de Monte Carlo com Cadeias de Markov);
- iv) Propor uma extensão do modelo na dimensão temporal dos dados. A fim de incorporar informações adicionais das diversas avaliações coletadas em diferentes períodos de tempo;
- v) Estudar diferentes formas de considerar a informação da covariável de vizinhança no modelo.

3 REVISÃO DE LITERATURA

3.1 Morte Súbita em Citros

Essa doença, MSC, representa uma ameaça potencial para a citricultura nacional, uma vez que afeta diversas variedades comerciais de citros, causando o definhamento generalizado das plantas dentro da área produtora.

Sua incidência foi primeiramente observada na região sul do Triângulo Mineiro e norte do Estado de São Paulo. Desde 1999, tem causado a morte súbita de laranjeiras (*Citrus sinensis*), tangerinas (*C. reticulata*) enxertadas em limoeiros Cravo (*C. limonia*), (BASSANEZI; FERNANDES; YAMMAMOTO, 2003). Por ser transmitida por enxertia de tecidos e pela sua disseminação no pomar suspeita-se que a doença seja causada por um vírus, transmitido por vetor aéreo. Dois vírus (vírus da doença da tristeza dos citros e outro da família *Tymoviridae*) parecem estar associados a doença.

Seu diagnóstico é realizado com base nos sintomas apresentados pelas plantas doentes, pois não é possível diagnosticar a MSC antes do aparecimento deles na planta. De início, ocorre perda generalizada do brilho das folhas das plantas doentes, perdendo a coloração verde-escuro, tendendo para o amarelo-esverdeado. Com a evolução da doença, ocorre a desfolha parcial da planta e seu sistema radicular apresenta grande quantidade de raízes podres e mortas. A morte do sistema radicular é consequência do bloqueio dos vasos do floema, que transportam a seiva elaborada das folhas para as raízes.

O sintoma característico da doença é o aparecimento de coloração amarela, tendendo para o alaranjado na parte interna da casca do porta-enxerto, logo abaixo da zona de enxertia (BASSANEZI; AMORIM; BERGAMIN, 2003a). O tempo entre o aparecimento dos primeiros sintomas e a morte da planta é bastante variável em função da idade da planta, da época do ano e da condição de carga dos frutos.

O avanço da doença pelas regiões produtoras causa preocupação aos produtores e pesquisadores, que buscam conhecer o seu comportamento no tempo e espaço. O interesse está nas relações espaciais entre plantas com sintomas de MSC. Essas informações servirão de suporte para as inferências sobre o seu período de incubação, a velocidade e os mecanismos de disseminação (Bassanezi et al. 2003).

3.2 Leprose dos Citros

A leprose dos citros é uma doença causada pelo vírus *Citrus leprosis vírus* (CiLV), transmitido pelo ácaro *Brevipalpus phoenicis*, medidas de controle da doença visam reduzir a população do vetor através de aplicações de acaricidas (BASSANEZI; LARANJEIRA, 2007).

Endêmica nas regiões produtoras do estado de São Paulo, a doença causa sérios prejuízos à produção. Na citricultura nacional, tem sido considerada a mais importante virose, pois reduz a produção e o período de vida das plantas de citros (RODRIGUES, 2000). Os principais sintomas da doença na planta são manchas em frutos, folhas e ramos, provocando queda prematura, seca de ramos, e levando a planta ao definhamento (BASSANEZI; LARANJEIRA, 2007).

Nas folhas, as lesões são inicialmente pouco salientes na face superior, apresentando coloração verde-pálida no centro e amarelada nas bordas. Nos frutos maduros, a lesão mostra-se na forma de uma depressão de coloração mais escura na casca, indo de verde a marrom-escura ou preta. Quando as lesões são abundantes, há queda intensa de folhas e frutos, reduzindo a capacidade fotossintética da planta. Nos ramos, a ocorrência de muitas lesões pode ocasionar a morte dos ponteiros. Os ramos mais grossos descascam transformando-se em porta de entrada para uma série de patógenos oportunistas.

O movimento e a distribuição espacial de vetores que transmitem a virose, afetam a disseminação nas plantas e o progresso da epidemia. Existem relativamente poucos estudos dos processos temporais e espaciais, relacionando a dinâmica da doença na população de plantas hospedeiras e a dinâmica populacional do vetor (CZERMAISNKI, 2006).

3.3 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG's), também denominados modelos exponenciais lineares, foram introduzidos por Nelder e Wedderburn (1972). A idéia básica consiste em estender os modelos lineares usuais para respostas gaussianas. Permite-se, então, diferentes opções para a distribuição da variável resposta, assumindo que a mesma pertença à família exponencial de distribuições e tenha maior flexibilidade para a relação funcional

entre a média da variável resposta e o preditor linear.

3.3.1 Especificação do Modelo

A estrutura de um MLG é dada por três componentes:

- i) um componente aleatório composto de uma variável aleatória Y com n observações independentes de uma distribuição pertencente à família exponencial e um vetor de médias μ ;
- ii) um componente sistemático composto por variáveis explicativas x_1, \dots, x_p tais que, produzem um preditor linear;
- iii) e uma função $g(\cdot)$ monótona diferenciável, conhecida como função de ligação, que relaciona os dois componentes.

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

$$f(y_i; \theta_i, \phi) = \exp[\phi\{y_i\theta_i - b(\theta_i) + c(y_i)\} + a(y_i, \phi)] , \quad (1)$$

em que $E[Y_i] = \mu_i$ e $\text{Var}[Y_i] = \phi^{-1}V_i$, $V_i = d\mu_i/d\theta_i$ é a função de variância, sendo $i = 1, 2, \dots, n$, $\theta = \theta(\beta)$ é o parâmetro canônico e ϕ é o parâmetro de dispersão.

Os modelos lineares generalizados são definidos por (1) e pelo componente sistemático,

$$g(\mu_i) = \eta_i , \quad (2)$$

em que $\eta_i = x_i\beta$ é o preditor linear, $\beta = (\beta_1, \dots, \beta_p)^\top$, $p < n$, é o vetor dos parâmetros da regressão a serem estimados, $x_i = (x_{i1}, \dots, x_{ip})$ 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, o que, 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 podem-se mencionar as distribuições normal, normal inversa, Poisson e binomial, dentre outras. A partir dessas distribuições podem-se obter os modelos de regressão normal inversa, regressão Poisson, regressão logística entre outros.

Desde sua proposição os MLG's tem sido largamente usados, maiores detalhes sobre o assunto podem ser encontrados em Nelder e Wedderburn (1972), Demétrio (2001) e McCullagh e Nelder (1989).

3.4 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, ($i = 1, \dots, n$). O modelo logístico é atrativo devido à facilidade de interpretação dos parâmetros do modelo. A partir dos parâmetros pode-se medir o aumento na chance de sucesso com o acréscimo de uma unidade da covariável associada enquanto as demais permanecem constantes.

3.4.1 Especificação do Modelo

Seja Y variável aleatória com distribuição binomial, escrita na forma da família exponencial por,

$$f(y; \theta) = \exp \left\{ \binom{m}{y} + y \ln \theta + (m - y) \ln(1 - \theta) \right\} I_A(y). \quad (3)$$

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

$$\theta(x) = P(Y = 1|x) = \frac{\exp\{\beta_0 + \sum_{k=1}^p \beta_k x_k\}}{1 + \exp\{\beta_0 + \sum_{k=1}^p \beta_k x_k\}}, \quad (4)$$

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

Observe que como:

$$\ln \left(\frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \sum_{k=1}^p \beta_k x_k , \quad (5)$$

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 MLG's, o logito é a função de ligação canônica para a regressão logística.

3.4.2 Inferência

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

$$L(\beta) = \prod_{i=1}^n (\theta(x_i)^{y_i} (1 - \theta(x_i))^{1-y_i}) , \quad (6)$$

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

O princípio da máxima verossimilhança é estimar o valor de β que maximiza $L(\beta)$. Algebricamente, é mais fácil trabalhar com o logaritmo dessa função,

$$l(\beta) = \ln[L(\beta)] = \sum [y_i \ln \left(\frac{\theta(x_i)}{1 - \theta(x_i)} \right) + \ln(1 - \theta(x_i))] . \quad (7)$$

Para encontrar o valor que maximiza $l(\beta)$, basta diferenciar a respectiva função com respeito a cada parâmetro β_j ($j = 0, 1, \dots, p$), obtendo-se o sistema de $p + 1$ equações, que, igualadas a zero, produzem como solução as estimativas de máxima verossimilhança de β .

As $p + 1$ equações são chamadas de equações de verossimilhança e, por não serem lineares nos parâmetros β_j , requerem métodos numéricos para obtenção da solução. Os métodos iterativos de Newton-Raphson e o escore de Fisher são algoritmos numéricos comumente utilizados com essa finalidade. Os valores ajustados para o modelo de regressão logística são, portanto, obtidos substituindo-se as estimativas de β em (4).

3.5 Modelo Autologístico

Uma vez ajustado um MLG aos dados, quer-se verificar a qualidade do ajuste, uma forma é observar a deviance residual e os graus-de-liberdade do resíduo do modelo. O valor da deviance dever estar próximo ao valor dos graus-de-liberdade do resíduo do modelo. Quando o valor da deviance é maior há superdispersão, mais detalhes Demétrio (2001). Em um contexto em que a superdispersão seja parcialmente explicada pelo padrão espacial. A análise exploratória busca o padrão espacial nos resíduos e havendo indicação, redefine-se o modelo. Uma proposta é usar o modelo autologístico.

O modelo autologístico é uma forma apresentada na literatura para que a dependência espacial dos dados seja explicitamente considerada na modelagem. É um modelo auto-regressivo espacial para dados binários, considera a observação em determinado local dependente das observações em locais vizinhos na forma de covariáveis. Grande parte da teoria estatística do modelo autologístico foi introduzida por Besag (1972), com alguns exemplos envolvendo doenças de plantas. Desde então, diversos autores como Gumpertz, Graham e Ristiano (1997), Huffer e Wu (1998), Gu e Zhu (2001) e Wu e Huffer (2005) desenvolveram suas pesquisas usando e fazendo generalizações do modelo.

O modelo é aplicável em diferentes contextos. Na modelagem de distribuições de espécies de plantas considerando covariáveis climáticas (WU; HUFFER, 1997), no estudo do efeito das informações de covariáveis de vizinhança na sobrevivência de árvores tropicais (HUBBEL; CONDIT; FOSTER, 2001). Em um conjunto de dados epidemiológicos espaciais envolvendo câncer de fígado (SHERMAN; APANASOVICH; CARROLL, 2006).

Gumpertz, Graham e Ristiano (1997) usaram o modelo para a análise da incidência de *Phytophthora* em pimentas de sino, dispostas em um *lattice*. O interesse estava em usar o modelo autologístico para estudar a probabilidade de um particular *quadrat* ter planta com a doença, a partir das informações de vizinhança e covariáveis de solo. Czermainski (2006) estudou a incidência de leprose dos citros utilizando o modelo autologístico, Viola (2007) aplicou o modelo a dados de um experimento com pimentão.

Em seus estudos sobre a distribuição de duas diferentes florestas no Canadá He, Zhou e Zhu (2003) observaram que o modelo autologístico conseguiu explicar a dependência espacial dos dados, levando os autores a obterem resultados satisfatórios.

Wu e Huffer (2005), em seu artigo, adicionaram um componente temporal ao modelo autológico para analisar dados binários do tipo espaço-temporal. O modelo, então, captura a relação entre a variável binária e possíveis covariáveis, além de simultaneamente ajustar a dependência espacial e a dependência temporal para um campo aleatório Markoviano espaço-temporal. Para ilustrar o método, o modelo foi ajustado a dados de rupturas causadas por besouros em pinus na Carolina do Norte/USA. As observações foram feitas em diferentes períodos de tempo.

A essência do modelo autológico é utilizar covariáveis obtidas a partir da variável resposta em posições vizinhas da localização considerada. Portanto, o elemento chave do modelo é a estrutura de vizinhança que será discutida na próxima Seção.

3.5.1 Estrutura de Vizinhança

Alguns tipos de dados estão dispostos regularmente em linhas e colunas em uma determinada região, ou seja, dados formando uma malha regular de pontos. A estrutura descrita permite construir covariáveis de vizinhança, que farão parte do modelo autológico.

A estrutura de vizinhança pode ser construída de diferentes formas, como ilustrado na Figura 1, considerando simplesmente a ordem de vizinhança ou considerando o efeito da planta nas linhas e colunas separadamente (GUMPERTZ; GRAHAM; RISTIANO, 1997).

Considerando as ordens de vizinhança, uma possibilidade é construir uma covariável dentro de cada ordem de vizinhança. A covariável de primeira ordem é a soma das observações de quatro vizinhos, dois vizinhos na mesma linha em colunas adjacentes e dois vizinhos na mesma coluna em linhas adjacentes, e a de segunda ordem é a soma das observações 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 observações dos dois vizinhos na mesma linha em colunas adjacentes, posições $(k, l + 1)$ e $(k, l - 1)$, enquanto que 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, são somadas as

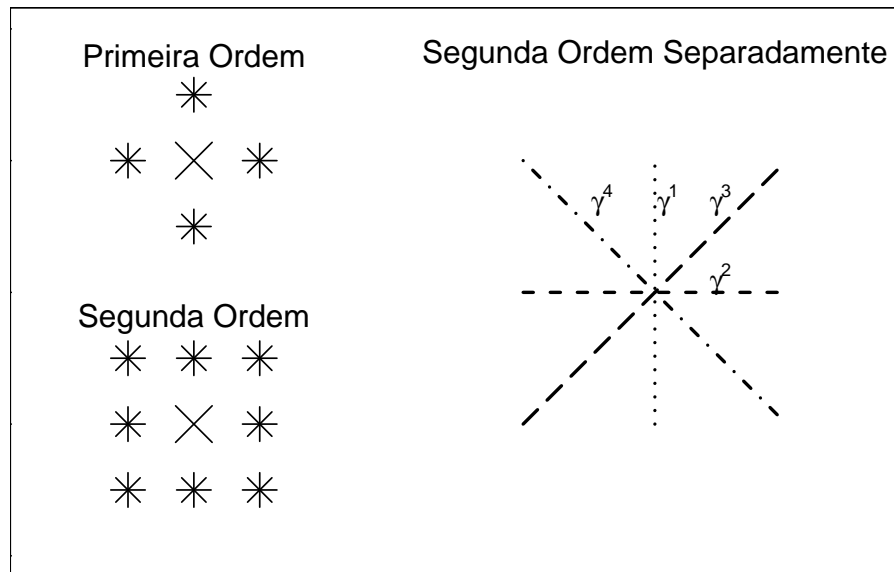


Figura 1 - Estruturas de vizinhança: primeira ordem, segunda ordem e segunda ordem separadamente.

covariáveis de primeira ordem a uma terceira covariável, a qual inclui os dois vizinhos da diagonal $(-1, 1)$, posições $(k - 1, l + 1)$ e $(k + 1, l - 1)$ e a uma quarta covariável que inclui os dois vizinhos da diagonal $(1, 1)$, posições $(k - 1, l - 1)$ e $(k + 1, l + 1)$ (HE; ZHOU; ZHU, 2003). De forma análoga pode-se considerar vizinhança de maior ordem.

A estrutura de segunda ordem com efeitos considerados é mais flexível e particularmente útil no caso com doenças de plantas que possuem espaçamento diferente entre linhas e colunas.

A informação da covariável de vizinhança pode ser considerada de duas formas no modelo autológico: a informação de existência ou não de vizinhas doentes ou, o número ou proporção de vizinhas doentes. No primeiro caso, plantas com uma vizinha doente serão consideradas de forma igual às plantas com mais de uma vizinha doente. Essas duas possibilidades de considerar a informação das vizinhas devem ser observadas na análise. Isso porque, ambas contemplam diferentes formas de contágio. No primeiro caso assume-se que o efeito de uma planta doente é o mesmo de duas plantas doentes.

Quando for considerado a contagem de plantas vizinhas doentes as expressões para o logito no caso de nenhuma vizinha doente, uma vizinha doente e duas vizinhas doentes

são, respectivamente:

$$\begin{aligned} \text{logit}(p_{kl}) &= \beta_0 \\ \text{logit}(p_{kl}) &= \beta_0 + \gamma_1 \\ \text{logit}(p_{kl}) &= \beta_0 + 2\gamma_1 \end{aligned}$$

Quando for considerado a covariável de vizinhança na forma binária, presença ou ausência de vizinhas doentes, as expressões são, respectivamente:

$$\begin{aligned} \text{logit}(p_{kl}) &= \beta_0 + \gamma_1 \\ \text{logit}(p_{kl}) &= \beta_0 \end{aligned}$$

3.5.2 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:

$$P(Y = y_i | x_{ji}, y_{(i)}) = \text{logit}(p_{kl}) = \sum_{j=1}^p \beta_j x_{ji} + \sum_{k=1}^q \gamma_k y_{(i)}, \quad (8)$$

em que $y_{(i)}$ 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 .

Um *lattice* D é uma região com n posições, cada uma descrita pelas coordenadas (k, l) especificando a linha e a coluna em que a observação é alocada. Em cada posição é observada uma resposta binária $y_{(k,l)}$, que tem valor 1 na presença da variável de interesse e zero, caso contrário, e um vetor $p \times 1$ de covariáveis $x_{(k,l)}$. 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; ZHOU; ZHU, 2003).

A expressão do modelo autologístico é dada por,

$$Pr(Y_{k,l} = y_{k,l} | x_{k,l}, y_{(k,l)}, (k, l) \in D) = \frac{\exp\{\sum_{j=0}^p \beta_j x_{k,l} + \sum_{t=1}^q \gamma_t y_{(k,l)}\}}{1 + \exp\{\sum_{j=0}^p \beta_j x_{k,l} + \sum_{t=1}^q \gamma_t y_{(k,l)}\}}, \quad (9)$$

em que, os β 's são parâmetros de regressão e os γ 's são os parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariáveis usuais e $y_{(k,l)}$ representam as covariáveis de vizinhança, com preditor linear $\eta = g(\mu) = \ln\left(\frac{\theta(x)}{1-\theta(x)}\right)$.

3.6 Inferência

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 $s_{k,l}$ é condicional ao *status* das outras observações $v_{k,l}$. Isso faz com que as observações sejam dependentes entre si, e ao ajustar um MLG usual, este não levará em conta a autocorrelação dos dados. Diversos métodos de estimação foram propostos na literatura.

O método de pseudo-verossimilhança proposto por Besag (1977) e utilizado por Gumpertz, Graham e Ristiano (1997) maximiza uma função de pseudo-verossimilhança para o modelo ajustado. A reestimação dos erros-padrão das estimativas de vizinhança é feita usando-se o método de Bootstrap com o Amostrador de Gibbs, pois, é preciso preservar a estrutura de vizinhança dos dados originais e o método de Bootstrap sozinho não permite isso. Em Hubbell, Condit e Foster (2001), o método de estimação usado foi de máxima verossimilhança com simulação Monte Carlo com Cadeias de Markov, utilizando o Amostrador de Gibbs, semelhante ao procedimento utilizado por Gumpertz, Graham e Ristiano (1997).

He, Zhou e Zhu (2003) compararam três métodos de estimação dos parâmetros: o método de pseudo-verossimilhança, proposto por Besag (1977), o de máxima verossimilhança com simulação Monte Carlo, introduzido por Geyer e Thompson (1992), e a aproximação estocástica Monte Carlo com cadeias de Markov proposto por Gu e Zhu (2001). Wu e Huffer (2005) estimaram os parâmetros do modelo por pseudo-verossimilhança e obtiveram a predição de respostas futuras no *lattice* por meio de Amostrador de Gibbs.

Sherman, Apanasovich e Carrol (2006) demonstraram porque o método de máxima verossimilhança não é adequado para estimação dos parâmetros no ajuste de dados binários espaciais usando o modelo autologístico. Discutiram métodos alternativos de estimação: máxima pseudo-verossimilhança, máxima pseudo-verossimilhança generalizada e máxima verossimilhança com simulação Monte Carlo. Esses três métodos de estimação foram aplicados a um conjunto de dados reais e a dados simulados.

3.6.1 Estimação por Pseudo-verossimilhança

O método de pseudo-verossimilhança foi proposto por Besag (1977) e estudado por muitos autores desde então. Besag sugeriu um método de estimação que maximize uma

função de pseudo-verossimilhança para o modelo ajustado, dado por:

$$l_{ps}(\theta) = \sum_{(k,l) \in D} \ln[P(Y_{k,l} = y_{k,l} | todos os outros valores)] ,$$

$$l_{ps}(\theta) = \sum_{(k,l) \in D} \{y_{k,l} f_{k,l}(\theta) - \ln[1 + \exp(f_{k,l}(\theta))]\} ,$$

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

O estimador $\hat{\theta}$, em geral, é consistente e assintoticamente normal. Entretanto, os desvios-padrão do estimador obtido pela regressão logística usual são inválidos e precisam ser recalculados de outra maneira (WU; HUFFER, 2005). Uma forma de recalculá-los proposta por Gumpertz, Graham e Ristiano (1997) é usar o método de reamostragem Bootstrap paramétrico.

3.6.2 Método de Bootstrap

A metodologia Bootstrap consiste em gerar amostras a partir dos dados originais para estimar a quantidade de interesse com essas amostras. 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 aplicada aos valores da amostra Bootstrap, condicional aos dados observados, é definida como distribuição Bootstrap dessa estatística.

Para o presente estudo, os passos de Bootstrap consistem em simular N lattices ($\tilde{y}^{(1)}, \dots, \tilde{y}^{(N)}$) dado as estimativas dos parâmetros de vizinhança do modelo ajustado aos dados observados. Obtém-se, então, a estimativa de $y_i^{(l)}$, para cada $\tilde{y}^{(l)}$ simulado.

Devido à configuração espacial dos dados, o método de Bootstrap necessita usar o algoritmo do amostrador de Gibbs (GEMAN; GEMAN, 1984). O Amostrador de Gibbs permite que o padrão espacial de vizinhança original seja preservado na reamostragem Bootstrap.

O amostrador de Gibbs é útil, no sentido de ser um procedimento de amostragem baseado em distribuições condicionais, que permite retirar amostras sucessivas da distribuição de cada dado y_i , condicionado aos seus vizinhos.

O algoritmo de Gibbs para obter N amostras $\hat{\gamma}$ é feito 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 aos 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 aos 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 $P(y_i^t | y_{i-1}^{t-1})$, em que

$$y_{i-1}^{t-1} = (y_0^t, \dots, y_{i-1}^t, y_{i+1}^{t-1}, \dots, y_n^{t-1}),$$

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.

3.6.3 Método de Monte Carlo

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

O método de Monte Carlo é mais simples do 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.

3.6.4 Método de Monte Carlo com Cadeias de Markov (MCMC)

Wu e Huffer (1995) e, posteriormente, outros autores concluíram que usar máxima verossimilhança com simulação MCMC é superior aos outros métodos usados para estimação dos parâmetros, como por exemplo, método de pseudo-verossimilhança. Em estudos de simulação as estimativas encontradas pelo método são aproximadamente normais e usando a informação de Fisher as estimativas de variância obtidas podem ser usadas para construir intervalos de confiança.

Os métodos de Monte Carlo com Cadeias de Markov são uma alternativa aos métodos não iterativos em problemas complexos. A idéia é obter uma amostra da distribuição a posteriori e calcular estimativas de características desta distribuição. As técnicas usadas são de simulação iterativa, baseadas em cadeias de Markov, e assim os valores gerados não serão mais independentes. Os métodos de MCMC mais usados são Amostrador de Gibbs e o algoritmo de Metropolis-Hastings.

3.6.4.1 Cadeias de Markov: Uma cadeia de Markov é um processo estocástico $\{X_0, X_1, \dots\}$ tal que a distribuição de X_t dado todos os valores anteriores X_0, \dots, X_{t-1} depende apenas de X_{t-1} .

Os métodos MCMC requerem que a cadeia seja:

- i) homogênea, isto é, as probabilidades de transição de um estado para outro são invariantes;
- ii) irredutível, isto é, cada estado pode ser atingido a partir de qualquer outro número finito de interações;
- iii) e aperiódica.

Uma questão de ordem prática é como os valores iniciais influenciam o comportamento da cadeia. A idéia é que conforme o número de iterações aumenta, a cadeia gradualmente esquece os valores iniciais e, eventualmente, converge para uma distribuição de equilíbrio. Assim, em aplicações práticas é comum que as iterações iniciais sejam descartadas, como se formassem uma amostra de aquecimento.

3.7 Dependência Temporal no Modelo Autologístico

No artigo de Wu e Huffer (2005) os dados do *lattice* analisado estavam disponíveis em diferentes períodos de tempo; ou seja, para uma dada localização s_i e uma dada observação no tempo t , a variável resposta é $Y_{i,t} = Y(s_i, t)$, ($i = 1, \dots, n$) e ($t = 1, 2, \dots$). Nesse caso, é necessário o acréscimo da dependência temporal no modelo autologístico.

A formulação do modelo tem uma similaridade com Besag (1972) e com Preisler (1993). Em Besag (1972), o modelo autologístico espaço-temporal foi proposto com a suposição de estacionaridade. Preisler (1993) considerou a regressão autologística espaço-temporal, mas supôs a independência entre as observações em diferentes tempos. A contribuição de Wu e Huffer (2005) foi de melhorar o modelo de Preisler (1993) de modo que a dependência temporal esteja contida no modelo. Além disso, o método pode ser estendido para Campos aleatórios Markovianos com dependência espacial e temporal.

3.7.1 Covariáveis de vizinhança no período de tempo anterior

Nas situações em que as observações são medidas em diferentes períodos de tempo pode-se 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 covariáveis de vizinhança no período de tempo anterior $t - 1$.

Essa abordagem permite explorar o potencial preditivo do modelo. Por exemplo, ajusta-se o modelo em cada avaliação do tempo t , acessando as covariáveis do tempo anterior $t - 1$, assim o modelo pode ser utilizado para prever a incidência no tempo futuro, baseado nas covariáveis de vizinhança do período atual.

O modelo é dado por,

$$Pr(Y_{k,l} = y_{k,l}^t | x_{k,l}, y_{(k,l)}^{t-1}, (k, l) \in D) = \frac{\exp\{\sum_{j=0}^p \beta_j x_{k,l} + \sum_{t=1}^q \gamma_t y_{(k,l)}^{t-1}\}}{1 + \exp\{\sum_{j=0}^p \beta_j x_{k,l} + \sum_{t=1}^q \gamma_t y_{(k,l)}^{t-1}\}}, \quad (10)$$

em que, os β 's são parâmetros de regressão e γ 's são parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariáveis e $y_{(k,l)}^{t-1}$ representam as covariáveis de vizinhança no tempo anterior. Com preditor linear $\eta = g(\mu) = \ln\left(\frac{\theta(x)}{1-\theta(x)}\right)$.

Nessa situação, a estimação dos parâmetros pode ser feito por pseudo-verossimilhança, e as estimativas dos erros-padrão fornecidas pelo MLG usual estão corretas,

pois não há autocorrelação das observações.

3.7.2 Covariáveis de vizinhança no tempo atual e anterior

A partir do modelo ajustado com covariáveis do tempo anterior, pode-se testar se ainda restou algum efeito possível de ser explicado pelas covariáveis no tempo atual. Para isso, constrói-se o modelo autologístico considerando covariáveis no período de tempo atual e anterior em um mesmo modelo. Tem-se, para esse caso,

$$Pr(Y_{k,l} = y_{k,l}^t | x_{k,l}, y_{k,l}^t, y_{(k,l)}^{t-1}, (k, l) \in D) = \frac{\exp\{\eta^*\}}{1 + \exp\{\eta^*\}}, \quad (11)$$

sendo que,

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

Ao fazer a estimação por pseudo-verossimilhança 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-padrão usuais não são apropriados.

4 MATERIAL E MÉTODOS

4.1 Material

Os dados com incidência de MSC e os dados com incidência de leprose dos citros a serem analisados nesse estudo foram cedidos pelo Fundecitrus (Fundo de Defesa da Citricultura).

Os dados de MSC foram obtidos na Fazenda Vale Verde, localizada no município de Comendador Gomes, estado de Minas Gerais. O talhão possui 20 linhas com 48 plantas em cada uma. O espaçamento entre linhas é de 7,5 metros e entre as plantas na linha é de 4 metros. Na época da coleta dos dados as plantas tinham 12 anos.

Foram coletadas observações em 12 avaliações feitas entre os dias 05/11/2001 e 07/10/2002. A coleta dos dados resumia-se a diagnosticar o *status* de cada planta dentro do talhão. Em cada avaliação, se a planta observada apresentasse sintomas da doença atribuíam-se o valor 1, se estivesse sadia atribuíam-se o valor 0.

O mapa da incidência da doença em cada avaliação está apresentado na Figura 2. A incidência da doença no talhão variou de 14,9% na primeira avaliação até 45,73% na 11ª avaliação. A 12ª avaliação foi descartada devido à alta incidência da doença.

Os dados com incidência de leprose de citros são de um talhão de citros, no município de Santa Cruz do Rio Pardo/SP. O espaçamento entre linhas é de 7,5 metros e entre plantas na linha é de 4 metros com 20 linhas de 58 plantas.

Foram feitas 45 avaliações entre os períodos de janeiro de 2002 e novembro de 2004. As avaliações foram feitas em intervalos variáveis de tempo, foram coletadas as seguintes informações: contagem de frutos, ramos e folhas com sintomas de leprose, dentro do talhão; a partir dos registros de infestação e de infecção nas plantas, foram obtidas as incidências binárias (0 para ausência do evento ácaro ou sintoma e 1 para presença do evento), correspondentes aos órgãos avaliados nas plantas (quanto a sintomas) e à planta como um todo (quanto a sintomas e ácaros).

Os talhões a serem analisados no trabalho permitem desenvolver uma metodologia de análise que poderá ser estendida para avaliar os demais talhões com incidência doenças cítricas que foram monitorados espacialmente.

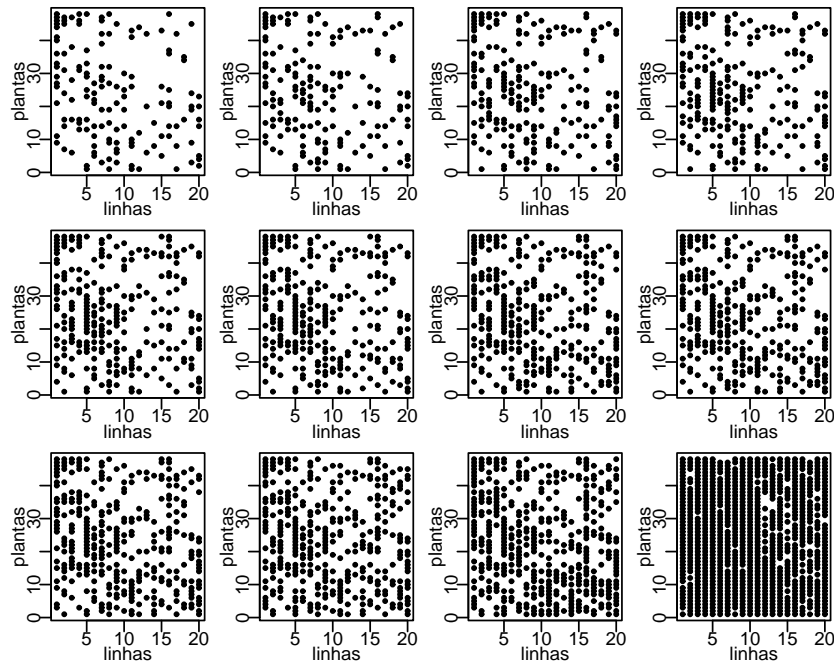


Figura 2 - Incidência de MSC em cada avaliação.

4.2 Métodos

Devido ao fato que, presença, ou não, da doença em determinada planta está altamente correlacionada com a presença, ou não, da doença em plantas vizinhas, espera-se que plantas próximas tenham características similares, particularmente, que a localização das plantas doentes apresente algum padrão espacial. Para os dados apresentados na Seção anterior a proposta é considerar a dependência espacial dos dados na modelagem estatística através do modelo autológico.

4.2.1 Especificação do modelo autológico

Pelo modelo autológico calcula-se a probabilidade de determinada planta estar doente como combinação linear do *status* da doença (doente/sadia) nas plantas vizinhas e de covariáveis usuais de um modelo de regressão. Tem-se, então,

$$P(Y = y_{kl} | x_{kl}, y_{(kl)}) = \text{logit}(p_{kl}) = \sum_{j=1}^p \beta_j x_{kl} + \sum_{k=1}^q \gamma_k y_{(kl)}, \quad (12)$$

sendo que $y_{(kl)}$ é o vetor de observações sem a kl -ésima observação, o índice kl indica a posição da observação na linha k e na coluna l . Os β_j 's são parâmetros de regressão e γ_k 's são os parâmetros de autocorrelação espacial, $x_{k,l}$ representam as covariáveis e $y_{(k,l)}$ representam as covariáveis de vizinhança.

Quer-se encontrar a melhor forma de considerar a informação das covariáveis de vizinhança no modelo ajustado aos dois conjuntos de dados. Diante desse fato serão comparadas duas possibilidades: considerar a informação como a existência ou não de vizinhas doentes (binária), e considerar o número de vizinhas doentes (contagem).

O fato de se terem avaliações dos dados em diferentes períodos de tempo, permite que seja explorada a dimensão temporal dos dados estudados. Uma proposta é modelar os dados de um período de tempo t , utilizando as covariáveis de vizinhança no período de tempo anterior $t - 1$, expresso pelo Modelo 2. Outra possibilidade é construir o modelo autologístico considerando covariáveis no período de tempo atual e período de tempo anterior expresso pelo Modelo 3.

Os modelos propostos aos dados são:

$$\begin{aligned} \text{Modelo1 : } \text{logit}(p_{kl}^t) &= \beta_0 + \gamma_1 L_{kl}^t + \gamma_2 C_{kl}^t + \gamma_3 Da_{kl}^t + \gamma_4 Db_{kl}^t \\ \text{Modelo2 : } \text{logit}(p_{kl}^t) &= \beta_0 + \gamma_1 L_{kl}^{t-1} + \gamma_2 C_{kl}^{t-1} + \gamma_3 Da_{kl}^{t-1} + \gamma_4 Db_{kl}^{t-1} \\ \text{Modelo3 : } \text{logit}(p_{kl}^t) &= \beta_0 + \gamma_1 L_{kl}^{t-1} + \gamma_2 C_{kl}^{t-1} + \gamma_3 Da_{kl}^{t-1} + \gamma_4 Db_{kl}^{t-1} + \\ &\quad \gamma_5 L_{kl}^t + \gamma_6 C_{kl}^t + \gamma_7 Da_{kl}^t + \gamma_8 Db_{kl}^t \end{aligned}$$

O teste da dependência espacial consiste em testar a significância dos coeficientes associados as covariáveis de vizinhança. Para isso, utiliza-se o seguinte resultado assintótico

$$\frac{\hat{\gamma}}{ep_{\gamma}} \sim N(0, 1) . \quad (13)$$

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 covariáveis 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 desse modelo. É interessante saber se será 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 coeficiente da covariável de vizinhança na linha for significativo, pode-se dizer que a doença tendeu a se propagar na linha.

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

4.2.2 Inferência para o modelo autologístico

A estimação do modelo será feita usando diferentes métodos propostos na literatura: Método de Pseudo-Verossimilhança, Método de Monte Carlo e Método de Monte Carlo com Cadeias de Markov. O *software* estatístico usado é o R, e o pacote Rcitrus, disponível para *download* em <http://www.leg.ufpr.br/Rcitrus>.

REFERÊNCIAS BIBLIOGRÁFICAS

- BASSANEZI, R. B. ; FERNANDES, N. G. and YAMMAMOTO, P. T. Morte súbita do citros. **Fundecitrus** São Paulo, Araraquara, SP, Brasil, 2003.
- BASSANEZI, R. B. ; AMORIN, L. and BERGAMIN FILHO, A. Spatial and Temporal Analyses of Citrus Sudden Death as a Tool to Generate Hypotheses Concerning Its Etiology. **Phytopathology.**, v. 93, n. 4, 502-512, 2003a.
- BASSANEZI, R. B. and LARANJEIRA, F. F. Spatial patterns of leprosis and its mite vector in commercial citrus groves in Brazil. **Plant Pathology.**, v. 56, 97-106, 2007.
- BESAG, J. Nearest-Neighbour Systems and the Auto-logistic Model for Binary Data, **Journal of the Royal Statistics Society**, Series B, n. 34, p. 75-83, 1972.]
- BESAG, J. Efficiency of Pseudo Likelihood Estimators for Simple Gaussian Fields, **Biometrika**, n.64, p.616-618, 1977.
- CZERMAINSKI, A. B. C. **Dinâmica espaço-temporal de populações do patossistema leprose dos citros em condições naturais de epidemia.** 2006. 91p. (Tese de Doutorado), ESALQ/USP, Piracicaba, 2006
- DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados em Experimentação Agrônoma** . Piracicaba: ESALQ. Departamento de Ciências Exatas, 2001. 113p.
- FRIEL, N. and PETTITT, A. N. Likelihood Estimation and Inference for the Autologistic Model, **Journal of Computational and Graphical Statistics**, vol.13, n.1, p.232-246, 2004.
- GELMAN, A. and MENG, X.L. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling, **Statistical Science**, v.13, n. 2, p.163-185, 1998

- GEYER, C. J. and THOMPSON, E. A. Constrained Monte Carlo Maximum Likelihood for Dependent Data, **Journal of the Royal Statistics Society**. Series B, n. 54, p. 657-699, 1992.
- GU, M. G. and ZHU, H. T. Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation, **Journal of the Royal Statistics Society**, Series B, n. 63, p. 339-355, 2001.
- GUMPERTZ, M. L. ; GRAHAM, J. M. and 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**, v.2, n.2, p.131-156, 1997.
- HE, F. ; ZHOU, J. and ZHU, H. Autologistic Regression Model for the Distribution of Vegetation, **Journal of Agricultural, Biological, and Environmental Statistics**, v.8, n.3, p.205-222, 2003.
- HUBBELL, S. P. ; CONDIT, R. and FOSTER, R. B. Local Neighborhood Effects on Long-term Survival of Individual Trees in a Neotropical Forest, **Ecological Research**, n. 16, p. 859-875, 2001
- PREISLER, H. K. Modelling Spatial Patterns of Trees Attacked by Bark Beetles, **Applied Statistics**, v.42, p.501-14, 1993.
- GEMAN, S. and GEMAN, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v.6, p.721-41, 1984.
- HUFFER, F. W. and WU, H. Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plants Species,(submitted for publication), 1995.
- McCULLAGH, P; NELDER, J. A. **Generalized linear models** 2nd ed. London: Chapman & Hall. 1989.511p.

- RODRIGUES, J.C.V. **Relações patógeno-vetor-planta no sistema leprose dos citros.** 2000. 168p. (Tese de Doutorado) Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2000.
- SHERMAN, M.; APANASOVICH, T. V. and CARROLL, R. J. On Estimation in Binary Autologistic Spatial Models, **Journal of Statistical Computation and Simulation**, v. 76, n. 2, p. 167-179, 2006.
- TRIBESS, T. B. e TADINI, C.C. **Suco de laranja natural minimamente processado: uma alternativa para ampliar o mercado de suco de laranja no Brasil** . Disponível em: <pqi.poli.usp.br/lea/docs/egna2001>. Acesso em: 18/01/2007.
- VIOLA, D. N. **Detecção e modelagem de padrão especial em dados binários e de contagem.** 2007. p. (Tese de Doutorado), ESALQ/USP, Piracicaba, 2007
- WU, H. and HUFFER, F. W. Modelling the Distribution of Plant Species using the Autologistic Regression Model, **Environmental and Ecological Statistics** , v. 4, p. 49-64, 1997.
- ZHU, J.;HUANG H. and WU, J. Modeling Spatial-Temporal Binary Data Using Markov Random Fields, **Journal of Agricultural, Biological, and Environmental Statistics**, v.10, n.2, p.212-225, 2005.