Modelo autologístico:

Uma aplicação a dados de morte súbita dos citros

Elias Teixeira Krainski [*] Luziane Franciscon [†] Paulo Justiniano Ribeiro Jr [‡]

1 Introdução

A produção citrícola brasileira responde por 53% do suco de laranja produzido no mundo e por 80% do suco concentrado. Citricultores, indústrias e cientistas brasileiros buscam aumento de produtividade, controle e manutenção da capacidade produtiva. Esta é constantemente ameaçada por doenças, como a *Morte Súbita dos Citrus* (MSC). A doença afeta variedades comerciais enxertadas em limoeiro *Cravo*, que representa grande parte dos pomares. Provoca diminuição de tamanho, peso e quantidade de frutos, além de rápido definhamento e a morte das plantas. Suspeita-se que seja causada por um vírus transmitido de forma bastante eficiente por um vetor aéreo, (Bassanezi, Fernandes & Yammamoto 2003).

As plantas estão dispostas dentro de um talhão, em linhas e colunas. Essa disposição em forma de reticulado permite o uso de modelos autoregressivos espaciais para a análise dos dados. No caso de dados binários, de incidência, o modelo autologístico é utilizado para estudar a probabilidade de doença de uma planta em

[*]LEG/UFPR e UFMG, ekrainski@ufmg.br

[†]LEG/UFPR e ESALQ/USP, lfrancis@esalq.usp.br

[‡]LEG/UFPR, paulojus@ufpr.br

função do *status* de suas vizinhas. Esse modelo possui simples interpretabilidade e incorpora explicitamente a estrutura de dependência espacial. Na agricultura este modelo foi inicialmente estudado para a análise da incidência de *Phytophthora* em pimentas de sino (Gumpertz, Graham & Ristano 1997).

O modelo é apresentado na seção 2. Na seção 3 faz-se uma análise de 11 avaliações de um talhão com MSC. As conclusões estão na Seção 4.

2 Metodologia

O modelo de regressão logística é utilizado para analisar respostas binárias. No caso de dados de incidência de doenças em plantas, torna-se necessário considerar a dependência espacial, pois espera-se que plantas próximas tenham características similares. O modelo utilizado neste caso é o modelo autologístico (Besag 1972).

2.1 Modelo autologístico

O modelo autologístico descreve a probabilidade de uma planta estar doente dado o *status* das plantas vizinhas, utilizando a função de ligação logito,

$$logit(p_{ij}) = \beta_0 + \gamma_1(y_{i-1,j} + y_{i+1,j}) + \gamma_2(y_{i,j-1} + y_{i,j+1}) \ , \qquad (1)$$

em que p_{ij} é a probabilidade da planta na linha i e na coluna j estar doente; $y_{i-1,j}$ e $y_{i+1,j}$ são as vizinhas das linhas adjacentes, formando a covariável de vizinhança entre linha; $y_{i,j-1}$ e $y_{i,j+1}$ são as vizinhas das colunas adjacentes, formando a covariável de vizinhança dentro da linha; γ_1 e γ_2 são os parâmetros que medem o efeito das covariáveis de vizinhança. Esta estrutura é flexível para estudar efeito direcional.

Um método simples de estimação de $\{\gamma_1, \gamma_2\} = \gamma$ é baseado em pseudo-verossimilhança (Besag

1977). Obten-se γ que maximize a função de pseudo-verossimilhança

$$\tilde{L}(\gamma, y) = \prod_i \prod_j f(p_{ij}, y) \ , \tag{2}$$

em que $f(.)$ é a densidade da distribuição Bernoulli. Este método estima bem os efeitos das covariáveis, porém as estimativas de variância dos efeitos são subestimadas.

Uma metodologia sugerida para estimar a variância das estimativas, é utilizar reamostragem. Devida a configuração espacial dos dados, a reamostragem não é simples e uma idéia é reamostrar em blocos (Cressie 1993). Uma alternativa é reamostrar utilizando o algoritmo amostrador de Gibbs (Gumpertz et al. 1997). A idéia é retirar amostras da distribuição de cada observação y_{ij} condicionando ao *status* das vizinhas, segundo a fórmula do modelo autologístico (1).

Inicialmente considera-se os dados observados, $y^{(0)}$ e estima o modelo utilizando o método de pseudo-verossimilhança, obtendo $\hat{\gamma}^{(0)}$. As estimativas $(\hat{\gamma}^{(1)}, ..., \hat{\gamma}^{(n)})$ são obtidas das re-amostras $(y^{(1)}, ..., y^{(n)})$. As re-amostras são obtidas utilizando os passos:

1. fazer um ciclo em ordem aleatória pelas plantas atualizando seu *status* utilizando $f(\hat{\gamma}^{(0)}, y^{t'})$, onde $y^{t'}$ são as observações atuais.

2. ajustar o modelo com a reamostra do passo anterior maximizando $\tilde{L}(\hat{\gamma^{(0)}}, y^{(t)})$,

3. repetir N vezes os passos 1 e 2.

A estimativa da variância de $\hat{\gamma}$ é calculando a variância de $(\hat{\gamma}^{(1)}, ..., \hat{\gamma}^{(n)})$. É comum descartar-se m observações iniciais e utilizar os valores obtidos a cada k passos. Esses procedimentos foram implementados em (R Development Core Team 2007), no pacote **Rcitrus** (Krainski & Ribeiro Jr. 2005).

2.2 Modelos considerados

Os dados analisados são 11 avaliações feitas em um mesmo talhão. Diante disso, consideramos três modelos. O modelo $m1$ para cada avaliação:

$$logit(p_{ij}^t) = \beta_0 + \gamma_1(y_{i-1,j}^t + y_{i+1,j}^t) + \gamma_2(y_{i,j-1}^t + y_{i,j+1}^t) ;$$

O modelo $m2$, com o *status* das plantas vizinhas da avaliação anterior:

$$logit(p_{ij}^t) = \beta_0 + \gamma_1(y_{i-1,j}^{t-1} + y_{i+1,j}^{t-1}) + \gamma_2(y_{i,j-1}^{t-1} + y_{i,j+1}^{t-1}) ;$$

e o modelo $m3$ com o *status* das plantas vizinhas na avaliação anterior e na mesma avaliação da resposta ao mesmo tempo:

$$\begin{aligned} logit(p_{ij}^t) = \beta_0 &+ \gamma_1(y_{i-1,j}^{t-1} + y_{i+1,j}^{t-1}) + \gamma_2(y_{i,j-1}^{t-1} + y_{i,j+1}^{t-1}) + \\ &\gamma_3(y_{i-1,j}^{t} + y_{i+1,j}^{t}) + \gamma_4(y_{i,j-1}^{t} + y_{i,j+1}^{t}) . \end{aligned}$$

O teste da significância dos parâmetros esta baseado em que $\hat{\gamma}/\sqrt{Var(\hat{\gamma})} \sim N(0,1)$. No modelo $m1$, o teste da significância dos coeficientes, verifica a existência da dependência espacial e permitirá ver se a agregação ocorre apenas nas linhas (curto alcance), entre as linhas (longo alcance). O modelo $m2$ avalia a capacidade preditiva deste modelo. É interessante saber se é possível prever o *status* das plantas em um momento futuro. A análise da significância de cada coeficiente separadamente, permitirá estudar a forma de propagação da doença. A partir do modelo $m3$, podemos avaliar conjuntamente o efeito das covariáveis no tempo atual e anterior.

Um critério de escolha de modelos que tem sido bastante utilizado é o critério de informação de Akaike, ou *Akaike information criteria - AIC*. O AIC é simplesmente $2 * log(L(Y,\theta)) + 2 * k$, onde k é o número de parâmetros.

3 Resultados

Os dados analisados são incidência de MSC em um talhão da Fazenda Vale Verde, no município de Comendador Gomes, em Minas Gerais. O talhão possui 20 linhas de plantas com 48 plantas em cada linha. O espaçamento entre linhas é de $7,5$ metros e entre as plantas na linha é de 4 metros. Foram analisadas 11 avaliações feitas entre os dias 05/11/2001 e 07/10/2002. A incidência da doença variou de $14,9\%$ na primeira avaliação até $45,73\%$ na 11^o avaliação. A variável resposta de interesse é a presença ou ausência de MSC em cada planta.

Os modelos apresentados na secção anterior foram ajustados aos dados. Nos modelos $m1$ e $m2$ apenas a covariável número de vizinhos na mesma linha foi significativa, conforme vemos os coeficientes e valores-p na tabela 1. Nas duas primeiras avaliações, notamos que não a covariável de vizinhança não é significativa no modelo $m1$. No modelo $m2$, ajustado para as avaliações 2 a 11, o resultado é similar, embora o valor-p para a segunda avaliação seja menor.

O modelo $m3$ foi ajustado considerando apenas o número de vizinhos na linha no tempo atual e anterior. Os coeficientes e valores-p desse modelo também estão na tabela 1. Nota-se algumas combinações de conclusões possíveis ao nível de 5% de significância: Nas avaliações 3, 5 e 6 ambas são significativas; nas avaliações 2, 4, 7, 8 e 11, apenas o número de vizinhas na mesma linha no tempo atual foi significativa; e, nas avaliações 9 e 10, apenas o número de vizinhos no tempo anterior foi significativa. Nota-se um efeito de colinearideade, induzido pelo uso de duas covariáveis que podem ter valores muito próximos, especialmente quando a incidência de plantas doentes é muito parecida em duas avaliações consecutivas.

O AIC de cada modelo em cada avaliação foi calculado, Tabela 2. Para a maioria

Tabela 1: Incidências, estimativas dos parâmetros dos modelos $m1$, $m2$, $m3$ e valores p

		Modelo 1		Modelo 2		Modelo 3			
						Tempo Atual		Tempo Anterior	
Av.	Incidência	$\hat{\gamma_1}$	valor p	$\hat{\gamma_1}$	valor p	$\hat{\gamma_1}$	valor p	$\hat{\gamma_1}$	valor p
1	0.14895	-2.02052	0.27365						
2	0.17293	-1.97306	0.16735	0.35758	0.0542	0.4166	0.0462	−0.0342	0.4350
3	0.21875	-1.84436	0.00221	0.44081	0.0054	1.0271	0.0000	−0.5060	0.0041
4	0.23840	-1.78096	0.00036	0.63954	0.0000	0.9160	0.0000	−0.2393	0.0600
5	0.26354	-1.68169	0.00126	0.61800	0.0000	0.3901	0.0163	0.2437	0.0246
6	0.27812	-1.63307	0.00025	0.59488	0.0000	0.8874	0.0000	−0.2445	0.0305
7	0.32292	-1.45117	0.00018	0.58248	0.0000	0.5437	0.0006	0.0972	0.1955
8	0.33125	-1.39161	0.00014	0.61067	0.0000	0.5732	0.0002	0.0703	0.2597
9	0.34167	-1.28953	0.00005	0.60794	0.0000	0.1542	0.1672	0.4721	0.0000
10	0.37500	-0.90676	0.00190	0.50256	0.0000	0.0641	0.3341	0.4443	0.0000
11	0.45729	-0.90008	0.00028	0.43635	0.0000	0.6371	0.0000	−0.1196	0.1179

das avaliações (2,4,5,6,7,8 e 11), o modelo mais adequado é o $m1$. O modelo 2 apenas foi melhor para a avaliação 3 e para as avaliações 9 e 10 o modelo adequado foi o 2.

4 Conclusão

Um aspecto importante do modelo autologístico é sua objetividade em analisar dados originais, sem discretização de informação, como é feita na análise por *quadrats*. Os resultados foram objetivos em apontar o padrão espacial da doença. A incidência

Tabela 2: Valores de AIC para os três modelos ajustados aos dados de MSC

	Modelo 1	Modelo 2	Modelo 3
2	725.55	726.76	727.54
3	813.25	824.66	812.33
4	851.58	858.66	853.08
5	908.32	909.09	909.81
6	932.52	936.61	934.17
7	992.94	997.26	994.80
8	1003.70	1004.79	1005.68
9	1019.30	1018.58	1020.50
10	1067.11	1064.87	1066.82
11	1109.49	1121.87	1111.08

ocorre aleatoriamente no talhão quando a incidência é baixa, nas duas primeiras avaliações, e ocorre de forma agregada quando a incidência foi de 21,88% a 45,73%.

A partir da terceira avaliação, as novas plantas contaminadas tem maior probabilidade de ocorrerem na vizinhança de uma planta contaminada. Além disso, dos resultados do modelo $m2$, nota-se que as plantas contaminadas na primeira avaliação exercem uma pressão infectiva quase significativa. Talvez se a avaliação tivesse ocorrido em um intervalo de tempo mais próximo, seria significativa.

Outra conclusão, é que a dependência espacial é de curto alcance, pois apenas o número de vizinhos na mesma linha foi significativo, sendo que o espaçamento dentro da linha é de 4 metros e entre linhas é de 7,5 metros.

Referências

Bassanezi, R., Fernandes, N. & Yammamoto, P. (2003). Morte súbita do citros, *Technical report*, Fundecitrus, Araraquara, SP, Brasil.

Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data, *Journal of the Royal Statistics Society, Series B* **34**: 75–83.

Besag, J. (1977). Efficiency of pseudo likelihood estimators for simple gaussian fields, *Biometrika* **64**: 616–618.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics.

Gumpertz, M. L., Graham, J. M. & Ristano, 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* **2**(2): 131–156.

Krainski, E. & Ribeiro Jr., P. (2005). *Rcitrus: Funções em R para análise de dados de doenças de citros*. R package version 0.3-0.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

*http://www.R-project.org