Autologistic model with an application to the citrus sudden death disease

Elias Teixeira Krainski *Paulo Justiniano Ribeiro Jr †
Luziane Franciscon ‡Renato Beozzo Bassanezi §

1 Introduction

Brazilian citrus fields are responsible for about 53% of the worldwide orange juice production and for 80% of the concentrated form. Citrus producers, industry and scientists are constantly aiming for higher productivity, control of the production process and capacity. Such targets are threaded by various diseases among which is the *Citrus sudden death* (MSC). This disease affects commercial varieties on *limoeiro Cravo*, which represents the majority of the commercial fields causing reduction in size, weight and number of fruits, combined with rapid decline and death of the trees. The suspicion it was caused by a virus transmited by a efficient flying vector (Bassanezi, Fernandes & Yammamoto 2003) has recently being confirmed.

In a citrus fields the trees are usually arranged in a grid with possibly different distances between rows and columns. There is an interest in assessing spatial patterns of the disease. Methods for characterizing the pattern as aggregated, regular or random are currently used. However such methods are note design to quantify the effects of spatial covariates since they do not assume an explicitly model relating such covariates with the presence of the disease. One alternative investigated here is the adoption of an autologistic model which relates the probability of a unit become diseased given the status of neighboring plants in space and/or time, taken as covariates and therefore having a associated coefficient parameter. The regular arrangement favors for the adoption of autoregressive models for the analysis which allows for the detection of usual covariate effects as well as the assessment of spatial effects. The latter are particularly useful for the description and hypothesis tests on the patterns of the disease, which may suggests propagation mechanisms and control strategies. For instance, for binary data such as presence/absence of the disease the autologistic model describes the probability of a tree become infected given the status of the neighboring trees. The model parameters has an straightforward interpretation, incorporating explicitly the dependence structure. In agricultural applications the model has being initially adopted the study the incident of *Phytophthora* in bell pepper (Gumpertz, Graham & Ristano 1997). Here we further explore the model considering the particular aspects of the MSC. The model is presented in Section 2 and Section3 reports the analysis of data collected at 11 different time points in a field with presence of MSC. The conclusions and discussion are presented in Section 4.

*LEG/UFPR e UFMG, elias@leg.ufpr.br
†LEG/UFPR, paulojus@ufpr.br
‡LEG/UFPR e ESALQ/USP, luziane@leg.ufpr.br
§Fundecitrus, rbbassanezi@fundecitrus.com.br

2 Methodology

The logistic regression model is currently widely used to the analysis of binary outcomes such as presence or absence of a certain characteristic of interest. For presence of plant disease it is particularly relevant to consider possible spatial dependence given it is reasonable to assume that neighbouring trees are more likely to have similar status. The autologistic model (Besag 1972) extends the usual logistic regression accounting for such spatial structure.

2.1 Autologistic model

The autologistic model describes the probability of a plant have the disease given the status of the neighbouring plants using through a covariate connected to the outcome through the link function,

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

with p_{ij} being the probability for the plant in the i^{th} row and j^{th} column; $y_{i-1,j}$ and $y_{i+1,j}$ are the status in the adjacents rows which are combined to produce the *row covariate*; $y_{i,j-1}$ and $y_{i,j+1}$ are the status of plants in adjacent columns producing the *column covariate*; γ_1 are γ_2 the respective parameters measuring the effect of such spatial covariates. The separation of row and column effects accommodates the fact the spacing are typically different within and between rows, allowing to study directional effects.

A naïve method to obtain parameter estimates for $\{\gamma_1, \gamma_2\} = \gamma$ is based on the maximisation of the pseudo-likelihood (Besag 1977)

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

where $f(.)$ is the density of the Bernoulli probability distribution. This methods provides consistent parameter point estimates, however underestimates the associated standard errors. Intuitively this is due to the fact that an observation is used as a responser variable as well as providing information for the covariates in the model.

One possible workaround is to use resampling methods. However within the context of spatial patterns this is not straightforward given the need to preserve the spatial structure. This can be achieved by block resampling (Cressie 1993) for instance using a Gibbs sampler (Gumpertz et al. 1997). The basic idea is to sample from the distribution of each observation y_{ij} conditioning on the current status of the neighbours, with probabilities given b the autologistic model (1). This is a sequential algorithm which goes as follows. We start with observed values $y^{(0)}$ from which we obtain via pseudo-likelihood parameter estimates $\hat{\gamma}^{(0)}$. Next we generate bootstrap resamples $(y^{(1)}, ..., y^{(n)})$ and for each of them estimates $(\hat{\gamma}^{(1)}, ..., \hat{\gamma}^{(n)})$ are obtained. The bootstrap sample are obtained through the following steps:

1. starting from an arbitrary location (tree) update the status by sampling from the Bernoulli distribution with probability given by the fitted model parameters and current status of the plants $f(\hat{\gamma}^{(0)}, y^{t'})$. This is done for all units in a random sequence until the cycle is complete, i. e. all the status are updated.

2. once a cycle is completed obtains parameter estimates by maximising pseudo-likelihood function $\tilde{L}(\hat{\gamma^{(0)}}, y^{(t)})$,

3. repeat steps 1 and 2 N times, the number of bootstrap samples.

The variance of the estimate $\hat{\gamma}$ is then given simply by the variance of the estimates $(\hat{\gamma}^{(1)}, ..., \hat{\gamma}^{(n)})$. It is also advisable to disregard a certain number m of initial resamples and also trimming the simulations taking one at each k steps. This allows for convergence of the chain and to reduce the number of stored simulations.

These procedures were implemented by us in a add-on package **Rcitrus** (Krainski & Ribeiro Jr. 2005) to the R statistical environment for statistical analysis (R Development Core Team 2007).

2.2 Models

The data considered here comes from 11 visits to the citrus field at different calendar dates. Three models were considered for the analysis. The first model ($m1$) consider as spatial covariates the neighbouring observations within and between rows, at the present time:

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

Model $m2$, considers the same neighbourhood, however with data reflecting the status of the plants at the previous observation time:

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

Finally, model $m3$ considers contemporary and previous times in the covariates.

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

The significance test for the regression parameters is based on the usual approximation in generalised linear models that $\hat{\gamma}/\sqrt{Var(\hat{\gamma})} \sim N(0,1)$. For $m1$, the significance test for the coefficients allows for detecting spatial effects as well as distinguish between the close neighbours effects given by the within row covariate and more distant neighbours given by the between rows covariate. Model $m2$ assess the predictive ability of the model through the lagged information built in the covariate given the fact the present status of the trees would allow to predict the probability of trees become infected at the next observation time. The different covariate effects assess patterns in the spread of the disease. A last model $m3$, can combines contemporary and lagged covariates attempt to check whether this improves the model fit.

The Akaike Information Criteria (AIC) is a measure used to assess and compare model fits and is given by the penalization of the log-likelihood by model complexity and is given by $2 * log(L(Y, \theta)) + 2 * k$, where k is the number of parameters included in the model.

3 Results

The data considered here were collected on a citrus field with presence of MSC within the farm Vale Verde, municipality of Comendador Gomes, Minas Gerais State, Brazil. The trees are arranged in 20 rows of 48 plants with spacing of $7,5\ m$ between rows and $4\ m$ within rows. Data were collected at 11 time points between 05/11/2001 and 07/10/2002. The incidence ranged from $14,9\%$ at the first visit to $45,73\%$ on the final data. The response variable used here is the presence/absence of MSC on each tree.

Tabela 1: Incidence, parameter estimates and p-values for models $m1$, $m2$, $m3$

		Model $m1$		Model $m2$		Model $m3$			
						Present time		Previous time	
Av.	Incidence	$\hat{\gamma_1}$	p-value	$\hat{\gamma_1}$	p-value	$\hat{\gamma_1}$	p-value	$\hat{\gamma_1}$	p-value
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

The models presented in Section 2 were fitted to the data. Table 1 shows significant effects only for the covariate number of neighbours within row for models $m1$ e $m2$ and the spatial covariate was not significant for the first and second data collections. Overall similar results were found for model $m2$.

Model $m3$ includes two spatial covariates: S_1 is number of within rows neighbours at present time and S_2 is number at previous time. Estimated coefficients and p-values are also shown in Table 1. Some combinations of relevant results are as follows. Both spatial covariates are significant at 5% significance level for times 3, 5 e 6; for times 2, 4, 7, 8 e 11, just S_1 was significant; and only S_2 for times 9 e 10. It is important to notice a potential (nearly) collinearity effect since the values of the two covariates can be similar, specially when the incidence is nearly the same between two consecutive observations in time.

Table 2 shows the Akaike Information Criteria (AIC) which is used to assess the fitted models. This criteria points that model $m1$ is the preferable one for most of the observation periods (2,4,5,6,7,8 e 11), Model $m3$ is better supported for time 3 and $m2$ for times 9 and 10.

4 Conclusion

The autologistic model provides a tool to further explore and describe spatial patterns of plant diseases beyond methods currently adopted, allowing to better unders-

Tabela 2: AIC values for the tree fitted models

	Model $m1$	Model $m2$	Model $m3$
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

tand mechanisms of the spread of the disease, not only detecting spatial patterns but also quantifying effects of presence of disease in different neighbourhood structures through the associated coefficients. An important feature of the autologistic model applied to individual trees is the objectivity when analysis original data, without the need of some sort discretizations, as for instance needed by methods based in *quadrats*.

The results found here for MSC points to the presence of spatial patterns in the disease for which evidence becomes clearly as the incidence rises. In general evidence of aggregation for levels of incidences higher than 20%. From the third data collection time onwards there was a noticeable increase of the probability of a plant become diseased in the presence of neighbours with the disease. Model fit for $m2$ shows evidence of infective pressure. Notice however the detection can be influenced by the time interval between observations. Overall the within row effect is stronger, reflecting the spacing adopted in the field and supporting the conjecture of spatial pattern, i.e. the closer the plants the higher the infective pressure.

Our conclusion at this stage is that the autologistic model has a potential do be widely adopted to investigate spatial patterns. It requires a extra computational burden compared with usual generalised linear models which we have overcomed with our computational implementation. Further attempts to explore more flexible and general descriptions of the spatial patterns, ways to combine a sequence of time observations are steps to be followed in our investigation. Also the methodology suggests a way to objectively combine data from different fields, allowing for investigation of effects of choices of spacing between trees, age, type of citrus and tree combinations and other properties which can vary between different fields.

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