A CASE STUDY: ORDINAL RESPONSES WITH SPATIO-TEMPORAL DEPENDENCIES

Rogelio Ramos-Quiroga and Graciela González-Farías

Comunicación Técnica No I-05-04/01-02-2005
(PE/CIMAT)

A Case Study: Ordinal Responses With Spatio-Temporal Dependencies

Rogelio Ramos-Quiroga *
Department of Probability and Statistics
CIMAT, Guanajuato, México

and
Graciela González-Farías
Department of Probability and Statistics
CIMAT, Guanajuato, México.

Abstract: Data structures with spatial and temporal dependencies are not uncommon in environmental and agronomic fields. We consider the modeling and estimation problem for these type of structures, in particular we consider proportional odds models with spatio-temporal covariables with estimation via maximum pseudlikelihood. We end by presenting a testing problem on treatment effects on data from a field experiment on *agave tequilana*.

Keywords and phrases: Pseudolikelihood; Ordinal responses; Spatio-temporal dependencies; Markov random fields.

1 Introduction

Treatment comparison is a common objective in many agronomic field experiments. An interesting situation for these comparisons is when we have to account for longitudinal and spatial dependencies. In this work we consider this setting when the response variable is an ordinal observation on the damage level on Agave plants due to a root infection disease. Our main objectives are to present a modeling and inference approach for data from a field experiment on agave plants.

Regression models with ordinal responses have been extensively studied. McCullagh (1980) presents a discussion on regression with ordinal data for the case of independent observations; extensions to some of the ideas presented in that paper have been treated in, for example, Anderson (1984) and Peterson and Harrel (1990) where the partial proportional odds models are discussed. Holtbrugge and Schumacher (1991) propose the stereotype model in the analysis of a clinical trial study. Many more alternative models, in this direction, have been worked out and the list is endless.

*This work was partially supported by research project 39017E of CONACYT-México

The inclusion of spatial dependencies has been pioneered by J. E. Besag (see the 1972, 1974 and 1975 references), who has also considered spatio-temporal dependencies for the binary case. The basic modeling proposal considers the construction of models based on conditionally specified distributions; the classical approach to solve the estimation problem for this kind of conditional probability models has been through the composite likelihood based on estimating function theory, as given by Godambe (1960), Godambe and Heyde (1987), Lindsay (1988), Godambe and Kale (1991) and Arnold and Strauss (1991) among others.

Many different algorithms and methodologies have been worked out for spatial models with categorial and ordinal responses, see for example Gumpertz *et al* (1997), Heagerty and Lele (1998) and Bartolucci and Besag (2002). In particular, Markov random field (MRF) model theory has shown to be very useful when approaching these type of models. Kutsyy (2001) proposes a model assuming a latent Gaussian process with a local dependency structure via a MRF, and then building an ordinal spatial process as an indicator of slices of the latent process. Wu and Huffer (1997) consider Monte Carlo Markov Chain (MCMC) methods in a Bayesian setting. Among other Bayesian approaches we can mention Besag (1974,1986), Knorr-Held and Besag (1998) and Heikkinen and Högmander (1994).

Similar models that include not only spatial dependencies but also time, use similar arguments, in particular Gumpertz *et al* (2000), show an application of a logistic regression model with spatial and temporal autocorrelations. In a series of papers, Cressie and Davidson (1998), Davidson *et al* (1999) and Huang and Cressie (2000), define the concept of partially ordered Markov models (POMMs), and obtain the asymptotic properties of maximum composite likelihood estimation, using the partial order when spatial dependencies are present. POMMs were also used in Peraza-Garay (2004), where a particular model is presented, to handle both, spatial an temporal dependencies in a study of ordinal data. Defining only Markovian time structure through transition probabilities and conditioning on the past values for the spatial dependencies, the full likelihood approach can be established and a modified version of the partial order (POMM's), allows for conditions on the existence and uniqueness of the MLE, as well as its asymptotic normality.

1.1 Case study

In this paper we discuss a model that considers spatio-temporal dependencies in the following way. The agave (*agave tequilana Weber* or simply Blue Agave) is a plant which grows in certain regions of México to produce tequila liquor. By 1998, more than twenty percent of the total plantation started to show symptoms of a sickness that causes the plant to whiter, leaving it unusable for production. The agave plants take 5 to 8 years to reach maturity; consequently the loss of plants represented a critical problem. After the identification of the microorganisms that cause the premature whiter of the Blue Agave plant, which was a difficult task due to the longevity of the plant, a Somatic Embryogenesis method was used to identify a toxin known as *Erwinia* and a fungus denominated *Fusarium* as the main agents of this disease. Other genetic studies were also performed, finding the existence of small variation, making the plant very susceptible to diseases like the one that was affecting the plants. As part of the many efforts made, we present here a data set resulting from a study implemented in a research station of a tequila company.

The objective of the study was to understand the spread of the fungus that produces the premature fading of the agave leaves and to test a pair of pesticides in its effect to control the illness. This type of fungus enters the plant through the root, and then the contagious would be logically expected in the rows of the cultivated area. On the other hand, the tool used in the cleaning of plants can transmit the contagious; therefore in a natural way the spread should be within rows and perhaps on a lesser degree among adjacent rows due to the physical separation between them. Once the fungus is present, it is difficult to eradicate it; therefore, the effect of the chemicals should be manifested mainly in a delay of the severity of leaf damage. In practice the damage is measured from 1 (healthy leaf), 2 (medium level of infection) and 3 (severe infection). Other important piece of information will be to predict the percentage of the damage of a field, mainly the percentage of plants in level 2 and 3 in order to evaluate when the plants should be harvested in order to minimize economical losses.

Field information was taken at different periods of time (total of 5, approximately every two months) to verify the advance of the disease for all the treatments, Figure 1 shows four panels with images of the same experimental field at four different times (labeled time 0 trough time 3), data on time 4 was reserved to compare predictions. The data has been either modified or partially concealed to preserve the confidentiality of the information. The field accommodates 28 rows of plants with a depth space of approximately 100 plants in each row. Three sections were marked, the top with 7 rows was assigned to treatment 1 (dose of 10 times a basic amount of milligrams of fungicide per liter of water), the 14 middle rows were for treatment 2 (a half concentration of treatment 1) and the 7 rows at the bottom of the field were controls with no applications of fungicides. Dark areas in Figure 1 (mainly at the top and at the right) are sections with no plants, light gray areas represent healthy plants, white represents plants with medium infection level and in dark gray we have plants with severe damage.

Among the hypotheses of interest we have, first of all, the treatment comparisons, followed by the contrast of the degree of spatio-temporal dependency versus a model with only independent observations. In the spatial aspect, we would like to verify if the contamination can be credited only to the dependence of rows or if other structure of neighbors must be taken into account. Finally, the generation of a prediction of damage in the field for the next period of time, to decide the appropriate harvesting time of the plant even if it has not reached its level of optimal maturity.

Since we want to model fungus transmission, and the distance between plants among rows and columns are different by design, the lattices will be non-regular, to this effect we will include the effect of neighbor distance, defining different coefficients in the model that capture the effect of dependencies by rows and by columns.

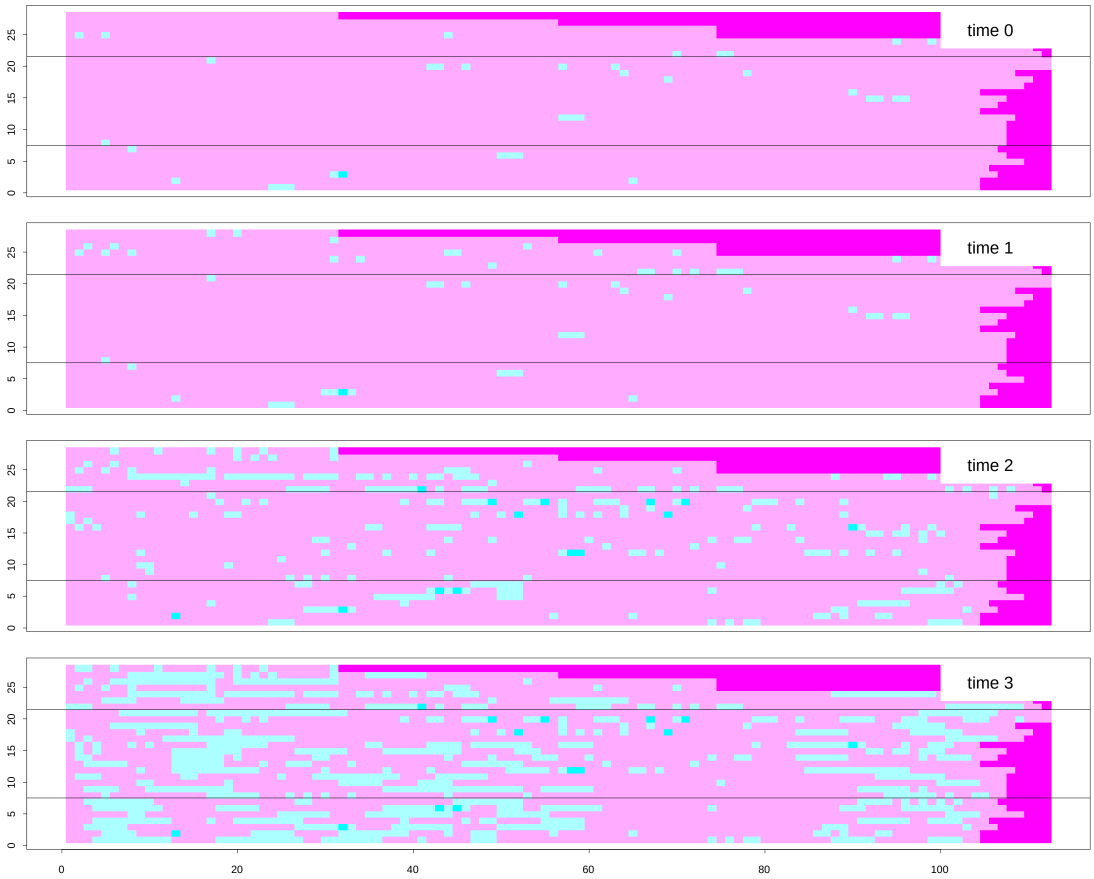

Figure 1: Spreading disease on *agave* plants

A factor that could be of interest is to determine the nature of spreading of the disease, whether it is in the direction of rows or if there is contamination to adjacent rows, this is of concern because weeding practices dictate cleaning by rows and this can compound the contamination since weeders that are not properly cleaned can carry the fungus from plant to plant; testing row and column effects will provide some light on this issue.

Section 2 provides the necessary nomenclature and establish the model, in Subsection 2.1 we provide a method of estimation, testing procedure and selection model. Once we select the proper model, in Section 3 we provide a way to generate predictive maps for a future time and evaluate its fit. Relevant conclusions and further work is discussed in Section 4.

2 The Model

Let y_{ijt} be the random variable defined by the status of the plant at row i, column j, and time t, $(i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad t = 0, 1, \cdots, T)$. We have plants on a lattice $I \times J$ which has been under observation at times $0, 1, \cdots, T$. The observed damage level in each plant is an ordinal response variable with levels $1, 2, \cdots, S$ where level 1 refers to healthy plants, and S is the highest degree of damage due to infection. We will assume a proportional odds model for the responses; that is, if y is the response of a given plant, and if $\gamma_d(x) = P(y \leq d \,|\, x), \ d = 1, \cdots, S-1$, is the cumulative probability of damage up to level d, then the proportional odds model postulates that [see McCullagh (1980)]

$$logit[\gamma_d(x)] = \theta_d - x^T \beta, \qquad d = 1, \cdots, S-1$$

where x is a vector of covariables associated with the given plant, and the θ's are the so-called "cut-point" parameters for the damage levels $(\theta_1 \leq \theta_2 \leq \cdots \theta_{S-1})$. Thus, the (conditional) probability distribution of y is given by

$$\pi_d(x) \equiv P(y = d \,|\, x) = \begin{cases} \gamma_1(x) & d = 1 \\ \gamma_d(x) - \gamma_{d-1}(x) & d = 2, \cdots, S-1 \\ 1 - \gamma_{S-1}(x) & d = S \end{cases}$$

The covariables to be considered include average status of neighboring plants, time, and treatment effects; that is, we are considering models for the conditional distribution of y_{ijt} given a set of values $y_{i'j't'}$ with indices on some neighborhood of (i, j, t). This defines a Markovian random field over the ensemble time × location. The concept of neighbor of a site s is related to those sites around the site, Besag (1972) calls them "closest neighbors", which for a lattice structure they are referred as neighbors of first order and include only the adjoining four quadrants, two in the same row and the other two in the adjoining rows; second order neighbors include the four neighbors in the diagonal plus the neighbors of first order. For a given site (i, j, t) we will use first order neighborhoods augmented by one "site" defined by $(i, j, t-1)$.

The consistency of models specified by conditional distributions has been extensively studied, see for example Arnold *et al* (1999); in our case the linear fashion in which the parameters enter in the logit model, together with multinomial variation, imply the stucture of a multivariate generalized model (see McCullagh (1980)) and thus the proposed model is guaranteed to be consistent, that is, there exists an underlying joint distribution with precisely the same conditionals as postulated in our model.

Now, for our particular problem on disease transmission, let x_1 be the average of damage levels of neighbors within the same row, x_2 the corresponding average for neighbors in adjacent rows, x_3 the damage level of the plant one period of time earlier, and let w_k be an indicator variable for treatment k, $(k = 1, \cdots, K)$; then, the vector of covariables is

$$x^T = (x_1, x_2, x_3, w_1, \cdots, w_K)$$

Thus, the proportional odds model is

$$\gamma_d(x_{ijt}) = P(y_{ijt} \leq d \,|\, x_{ijt}), \qquad d = 1, \cdots, S-1$$

with

$$logit[\gamma_d(x_{ijt})] = \theta_d - x_{ijt}^T\beta, \qquad \theta_1 \leq \theta_2 \leq \cdots \theta_{S-1}.$$

and

$$\beta^T = (\tau_1, \tau_2, \tau_3, \alpha^T)$$

is the $(K+3) \times 1$ vector of regression parameters. The implicit assumption of proportional odds implies the somewhat restriction that regardless of the severity, d, of the disease, the relative impact of the covariables on the response remains the same.

Define the neighbors of y_{ijt} as

$$N_{ijt} = \{\ y_{i-1,jt}\ ,\ y_{i+1,jt}\ ,\ y_{i,j-1,t}\ ,\ y_{i,j+1,t}\ ,\ y_{ij,t-1}\}$$

and for each time $t = 1, 2, 3$ split the data in two groups G_{1t} and G_{2t} so that if a given observation for time t is in group G_{ut}, then its time t neighbors belong to G_{vt}, $(u \neq v)$. Using this split we can write a conditional likelihood for each group. The contribution of y_{ijt} to its corresponding conditional likelihood is obtained assuming a multinomial model

$$\prod_{d=1}^{S} \pi_{ijtd}^{\delta_{ijtd}}(x_{ijt})$$

where $\delta_{ijtd} = 1$ if $y_{ijt} = d$ and 0 otherwise; thus, the composite likelihood or pseudolikelihood is

$$L = C \prod_{t=1}^{3} \prod_{g=1}^{2} \prod_{(i,j,t) \in G_{gt}} \prod_{d=1}^{S} \pi_{ijtd}^{\delta_{ijtd}}(x_{ijt})$$

Properties of the pseudolikelihood estimator (PLE) have been studied in Lindsay (1988), Arnold and Strauss (1991) and Heagerty and Lele (1998), among others; in particular, large sample properties of PLE's include consistency and asymptotic normality.

2.1 Data analysis

Straightforward maximization yields parameter estimates, of interest are those associated with treatment effects, α_1 and α_2, since they are related to the odds of having a healthy plant (level 1) for plants under treatment i, $i = 1, 2$, versus the corresponding odds for control plants, these odds ratios are $\exp(-\alpha_i)$. For the agave data we find $\hat{\alpha}_1 = -0.23$ and $\hat{\alpha}_2 = -0.09$, which potentially indicate that both treatments have a positive effect in slowing the spread of the disease; however a comparison against a reduced model with no treatment effects results in a p-value of 0.43 and thus treatment effects do not show a significant difference compared with the control plants.

A secondary interest is to test whether or not the disease is really affecting neighboring plants. A pseudolikelihood ratio test to compare a full model with past and spatial effects versus a reduced model without spatial terms shows a strong significance, thus supporting that the disease really spreads to neighboring plants. Tests for the significance of the individual spatial effects (row and column effects) show that both are important; thus our working model is

$$\begin{aligned} logit[P(y \leq 1)] &= \hat{\theta}_1 - \hat{\tau}_1 x_1 - \hat{\tau}_2 x_2 - \hat{\tau}_3 x_3 \\ logit[P(y \leq 2)] &= \hat{\theta}_2 - \hat{\tau}_1 x_1 - \hat{\tau}_2 x_2 - \hat{\tau}_3 x_3 \end{aligned}$$

with $\hat{\theta}_1 = 16.8$, $\hat{\theta}_2 = 27.0$, $\hat{\tau}_1 = 5.2$, $\hat{\tau}_2 = 0.8$, and $\hat{\tau}_3 = 6.9$. From here, for example, plants surrounded by healthy neighbors ($x_1 = 1$ and $x_2 = 1$, average states of row and column neighbors respectively), will have an estimated probability distribution given by

$$\begin{aligned} P(y = 1) &= \frac{e^{10.8-6.9x_3}}{1 + e^{10.8-6.9x_3}} \\ P(y = 2) &= \frac{e^{21.0-6.9x_3}}{1 + e^{21.0-6.9x_3}} - \frac{e^{10.8-6.9x_3}}{1 + e^{10.8-6.9x_3}} \\ P(y = 3) &= \frac{1}{1 + e^{21.0-6.9x_3}} \end{aligned}$$

where x_3 is the state of the plant in the previous time. Since the covariables x_1, x_2 and x_3 are on the same scale, its estimated coefficients measure their relative impact on the response and thus x_3, the state of a plant in the previous state, is, logically, the covariable with the most impact in the future state of a plant.

Table 1 shows the estimated transition probabilities for three scenarios, in the first we assume row neighboring plants with a medium level of infection (level 2) and compute the distributions for level of damage conditional on the damage level present in the previous time; from the table we see that a plant with a healthy past has a high probability of 0.78 to reach a level 2 of infection, also, if a plant already was in level 2 it would have an estimated probability of .11 to worsen its state. In the second scenario we show the estimated effect for the case of column neighbors in level 2 whereas row neighbors are in level 1, for this case we see that the effect of column neighbors is weaker than the effect of row neighbors which is consistent with the fact that a row neighbor of a plant is closer than a column neighbor. Finally, in the third scenario we have the case of plants surrounded by heavily infected plants (level 3), in this extreme scenario it is very likely that a plant would increase its damage level regardless of its previous state.

	1	2	3	1	2	3	1	2	3
1	0.22	0.78	0.00	0.96	0.04	0.00	0.00	0.90	0.10
2	0.00	0.89	0.11	0.02	0.97	0.00	0.00	0.01	0.99
3	0.00	0.01	0.99	0.00	0.42	0.58	0.00	0.00	1.00
	$x_1 = 2,\ x_2 = 1$			$x_1 = 1,\ x_2 = 2$			$x_1 = 3,\ x_2 = 3$		
	Scenario 1			Scenario 2			Scenario 3		

Table 1: Transition Probabilities

3 Prediction

Plants with an advanced degree of infection are not suitable for harvesting, thus it is important to have a tool for the prediction of the degree of infection that will be present in a future time. Prediction for random fields requires the knowledge of the state of neighboring units, which clearly are not available, thus we adopt an iterative procedure in which we initially assume as proxies for future levels the current observations. A procedure for predicting future realization of the states of the random field is as follows:

- *Step 1:* Assume current neighbors will remain without change on their status. Predict status for each plant, based on the estimated model. Choose as predictor, the category with the maximum estimated probability.
- *Step 2:* Update status of neighbors using predictions from Step 1 and do a new prediction exercise.

For the agave data set we estimated the model based on times 1 through 3, we have available observations on disease levels for time 4, so that we can compare the performance of the prediction algorithm. Table 2 shows a crosstabulation of predicted versus observed disease levels for time 4. The diagonal of the table has frequencies of classification successes; similarly, off-diagonal values are misclassifications; the procedure had a total of 132 errors out of 2066 trials, or a 6.4% prediction error rate.

		Observed			
		1	2	3	
Predicted	1	1328	83	0	1411
	2	28	593	10	631
	3	0	11	13	24
		1356	687	23	2066

Table 2: Prediction at time 4.

4 Conclusions

Crops with long times to maturity that suffer damages represent an economical risk not only for the loss of the crop itself but also for the lost time and incurred maintenance costs; thus it is important to have predictive models that can monitor the spread of a particular disease. In this work we presented a model for the probability distribution of the levels of an infectious disease in agave plants. This model incorporates spatial and temporal effects and can easily accommodate other covariables such as treatment effects. In particular, the core of the model is the proportional odds model conditional to the states of neighboring plants. This model proposal defines a Markov random field and for estimation purposes we opted for a pseudolikelihood approach.

The model lends itself to testing of nested submodels to answer the pertinent questions posed by the researcher. In particular, for the agave data set, we found that the treatments had no significative difference with respect with control plants; however the spatial dependence was found to be highly significant.

Pseudolikelihood approaches are particularly useful for spatio-temporal problems but more research is needed to evaluate the loss of efficiency with respect to full likelihood approaches.

Acknowledgements: This work was partially supported by research project 39017E CONACYT-México.

References

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1–30.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (1991). *Conditional Specification of Statistical Models.* Springer-Verlag, New York.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya: The Indian Journal of Statistics, Series B*, **53**, 233–243.

Bartolucci, F and Besag, J. E. (2002). A recursive algorithm for Markov random fields. *Biometrika*, **89**, 724–730.

Besag, J. E. (1972). Nearest neighbour system and the auto logist model for binary data. *Journal of the Royal Statistical Society, Series B*, **34**, 75–83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician*, **36**, 192–236.

Besag, J. E. (1986). On the statistical analysis of dirty pictures (with Discussion).*Journal of the Royal Statistical Society, Series B*, **48**, 259–302.

Cressie, N. and Davidson, J. L. (1998). Image analysis with partially ordered Markov models. *Computational Statistics & Data Analysis*, **29**, 1–26.

Davidson, J. L., Cressie, N. and Hua, X. (1999). Texture synthesis and pattern recognition for partially ordered Markov models. *Pattern Recognition*, **34**, 1475–1505.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1211.

Godambe, V. P. and Kale, C. C. (1991). Estimation functions: an overview. In *Estimating Functions* (Ed., V.P. Godambe), pp. 3–20, Oxford: Claredon Press.

Godambe, V. P. and Heyde, C. (1987). An optimum property of regular maximum likelihood estimation. *International Statistical Review*, **55**, 231-244.

Gumpertz, M. L., Graham, J. M. and Ristaino, J. B. (1997). Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricutural, Biological, and Environmental Statistics*, **2**, 131–156

Gumpertz M. L., Wu, C. T. and Pye J. M. (2000). Logistic Regression for Southern Pine Beetle Outbreaks With Spatial and Temporal Autocorrelation. *Forest Science*, **46**, 95–107.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**,1099-1111.

Heikkinen, J. and Högmander, H. (1994). Fully Bayesian approach to image restoration with application in biogeography. *Applied Statistics*, **43**, 569–582.

Holtbrugge, W. and Scumacher, M. (1991). Quasi likelihood and optimal estimation. *Applied Statistics* **40**, 2, 249–259.

Huang, H. and Cressie, N. (2000). Asymptotic properties of maximum (composite) likelihood estimators for partially ordered Markov models. *Statistica Sinica*, **10**, 1325–1344.

Knorr-Held, L. and Besag, J. E. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, **17**, 2045–2060.

Kutsyy, V. (2001). Modeling and inference for spatial processes with ordinal data. Thesis proposal. Department of Statistics. The University of Michigan: Michigan.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 221-239.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, **42**, 2, 109–142.

Peraza-Garay, F. (2004). A model for longitudinal and ordinal data with spatial dependency. *Disertación Doctoral.* CIMAT, México.

Peterson, B. and Harrel, F. E. (1990). Partial proportional odds models for ordinal responses variables. *Applied Statistics.* **39**, 2, 205–217.

Wu, H. and Huffer, F. W. (1997). Modeling the distribution plant species using the autologistic regression model. *Envirnonmental and Ecolocical Statistics.* **4**, 2, 49-64.