

Composite likelihood estimation in spatial logistic regression

Stima di massima verosimiglianza composita per modelli logistici spaziali

Manuela Cattelan and Cristiano Varin

Abstract A regression model for binary spatial data with dependence expressed in terms of cross-ratios is discussed. In order to avoid difficulties associated with the specification of multivariate binary distributions, a conditional pairwise likelihood method is developed for inferential purposes. The methodology is illustrated through the analysis of data about the effect of acid rains on trout populations in Norwegian lakes.

Abstract *Si propone un modello di regressione per dati spaziali binari in cui la dipendenza è modellata attraverso la specificazione dei cross-ratios. Al fine di evitare le difficoltà associate alla specificazione di una distribuzione congiunta per dati binari multivariati, l'inferenza è basata sulla verosimiglianza a coppie condizionata. La metodologia è illustrata attraverso l'analisi di dati concernenti gli effetti della pioggia acida sulla popolazione di trote presenti nei laghi norvegesi.*

Key words: binary data, conditional pairwise likelihood, geostatistics, logistic regression, spatial data

1 Introduction

The analysis of data deriving from spatial applications is often cumbersome because these data are naturally characterized by spatial dependence that should be taken into account in the model. It is even harder to analyze binary data, since the specification of a full multivariate distribution of the observations may be cumbersome and the

Manuela Cattelan

Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova, Italy
e-mail: manuela.cattelan@unipd.it

Cristiano Varin

D.A.I.S., Ca' Foscari University - Venice, San Giobbe Cannaregio 873, 30121 Venice, Italy e-mail: cristiano.varin@unive.it

calculation of the likelihood usually poses noticeable computational difficulties [3, 8, 10].

A model for spatial binary data, in which dependence is accounted for through the specification of cross-ratios for couples of observations is proposed. We suggest to use conditional pairwise likelihood for making inference in the model, hence only bivariate distributions need to be specified. Section 2 introduces the model proposed and the composite likelihood estimation. Section 3 illustrates the model through an application to the reduction in the population of trouts in Norwegian lakes and Section 4 closes the paper.

2 Regression model for binary spatial data

Let Y_i , $i = 1, \dots, n$, denote a binary random variable that can assume values 0 or 1. Suppose that, for each observation a vector of p explanatory variables x_i is available. The marginal probabilities are described through the logistic regression model

$$\pi_i = P(Y_i = 1) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)},$$

where β is a p -dimensional vector of regression parameters.

Dependence in binary observations is often described through the cross-ratio between the random variables Y_i and Y_j

$$\psi_{ij} = \frac{P(Y_i = 1, Y_j = 1)P(Y_i = 0, Y_j = 0)}{P(Y_i = 1, Y_j = 0)P(Y_i = 0, Y_j = 1)}.$$

Spatial dependence between observations decreases with distance. Accordingly, we propose to model the cross-ratio as

$$\psi_{ij} = \exp \left\{ \frac{\alpha}{d_{ij}} \right\},$$

where d_{ij} denotes the distance between observations Y_i and Y_j and α is a given constant. This specification ensures that the cross-ratio is non negative. If α is zero, then all cross-ratios are equal to one, and the data are independent. Otherwise, the cross-ratio decreases as observations become more and more distant, and α allows to determine the distance at which observations can be considered essentially independent.

Given univariate marginal probabilities and cross-ratios, the bivariate marginal probabilities are

$$\pi_{ij} = P(Y_i = 1, Y_j = 1) = \frac{1 + (\pi_i + \pi_j)(\psi_{ij} - 1) - G(\psi_{ij}, \pi_i, \pi_j)}{2(\psi_{ij} - 1)},$$

where $G(\psi_{ij}, \pi_i, \pi_j) = \sqrt{[1 + (\pi_i + \pi_j)(\psi_{ij} - 1)]^2 - 4\psi_{ij}(\psi_{ij} - 1)\pi_i\pi_j}$. Likelihood inference for a model that describes the multivariate binary distribution of all observations would require to specify a large set of higher-order cross ratio parameters and to solve complex high-order polynomials, see [6].

As a feasible alternative, we suggest to employ an estimating method that requires only the specification of univariate and bivariate distributions. This method is an instance of composite likelihood methods [4, 9]. Composite likelihood is a type of pseudo likelihood constructed from the combination of likelihoods for subsets of marginal or conditional events. Following the suggestions in [5], we consider the conditional pairwise likelihood

$$PL(\theta; Y) = \prod_{i=1}^n \prod_{j \neq i} P(Y_i | Y_j)^{w_{ij}},$$

where $\theta = (\beta, \alpha)$ denotes the vector of model parameters, w_{ij} are non negative weights and $Y = (Y_1, \dots, Y_n)$. The computation of conditional probabilities is straightforward exploiting the formula given in [1]

$$\text{logit} P(Y_i = 1 | Y_j = y_j) = \gamma_{ij} y_j + \log \frac{\pi_i - \pi_{ij}}{1 - \pi_i - \pi_j + \pi_{ij}},$$

where $\gamma_{ij} = \log \psi_{ij}$. Accordingly, the conditional probabilities become

$$P(Y_i | Y_j; \theta) = P(Y_i = y_i | Y_j = y_j; \theta) = \frac{\exp(\gamma_{ij} y_i y_j) \frac{\pi_i - \pi_{ij}}{1 - \pi_i - \pi_j + \pi_{ij}}}{1 + \exp(\gamma_{ij} y_i y_j) \frac{\pi_i - \pi_{ij}}{1 - \pi_i - \pi_j + \pi_{ij}}}.$$

Thus, the conditional pairwise log likelihood is

$$\begin{aligned} pl(\theta; Y) &= \sum_{i=1}^n \sum_{j \neq i} \left(\frac{\alpha}{d_{ij}} y_i y_j + y_i \log \frac{\pi_i - \pi_{ij}}{1 - \pi_i - \pi_j + \pi_{ij}} - \log \left\{ 1 + \frac{\pi_i - \pi_{ij}}{1 - \pi_i - \pi_j + \pi_{ij}} e^{\frac{\alpha}{d_{ij}} y_j} \right\} \right) w_{ij} \\ &= \sum_{i=1}^n \sum_{j \neq i} w_{ij} \log P(Y_i | Y_j). \end{aligned}$$

Under suitable regularity conditions, the maximum pairwise likelihood estimator is asymptotically normally distributed with variance $G(\theta)^{-1} = H(\theta)^{-1} J(\theta) H(\theta)^{-1}$, where $H(\theta) = E(-\nabla^2 pl(\theta; Y))$ and $J(\theta) = \text{var}(\nabla pl(\theta; Y))$, see [7].

3 Trouts data

The data considered here have already been analyzed in [8]. These data concern the status of the trout population in 542 Norwegian lakes in 1986. After interviewing local fishermen, the population status is recorded as unaffected or decreased/extinct,

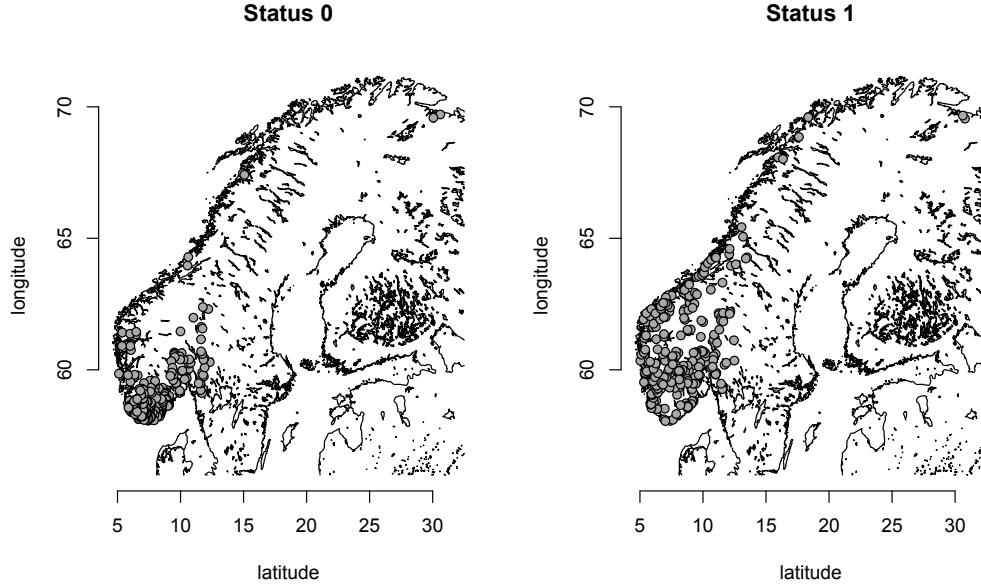


Fig. 1 Trouts data. Lakes where the population of trouts is unaffected (left panel) and lakes where the population is decreased/extinct (right panel).

coded as 0 and 1, respectively. Fig. 1 shows the lakes that are unaffected and those which suffer a reduction in the trout population. Trouts are particularly sensitive to acidification of lakes and streams which may be caused by acid deposition from long range transportation of air pollutants. It is of interest to investigate whether the measured acid neutralizing capacity (ANC) affects the probabilities of observing a decrease in the population of trouts. The covariate ANC reflects local properties of soils and the load from current and historic acid deposition.

It is not numerically efficient to consider all the $\binom{n}{2}$ possible couples of observations in the construction of the conditional pairwise likelihood since it is plausible that couples formed by closest observations are more informative, hence the first step of the data analysis is their identification. For this purpose, the empirical lorelogram [2] of the data is computed. The data with distance up to 350 kilometers are divided into 10 bins and the logarithm of the empirical cross-ratio is computed. The empirical lorelogram decreases as the distance between couples of observations increases. The model based log cross-ratio is α/d_{ij} . We consider as distance the center of each bin and estimate the parameter α using weighted least squares with weights equal to the number of observations in each bin. The log cross-ratio approaches zero in the bin between 105 and 140 kilometers. Accordingly, in the conditional pairwise likelihood, couples of observations which are more than 140 kilometers apart are excluded ($w_{ij} = 0$).

Table 1 Estimates, standard errors, z -values and p -values of the proposed model for the trouts data.

	Estimate	Standard error	z -value	p -value
Intercept	0.798	1.743	-0.458	0.647
ANC	-14.909	2.568	-5.806	$6 \cdot 10^{-9}$
$\log(\text{area})$	-0.275	0.168	-1.636	0.102
$\log(\text{altitude})$	0.104	0.296	0.350	0.727

Further covariates on the area and the altitude of the lakes are available, so we fit the model

$$\text{logit}\{E(Y_i|x_i)\} = \beta_0 + \beta_1 \text{ANC} + \beta_2 \log(\text{area}) + \beta_3 \log(\text{altitude}),$$

Table 1 reports the estimates and their standard errors. The covariate ANC appears highly significant in keeping the number of trouts stable. The other covariates considered, area and altitude of the lake, are not significant. The estimate of the parameter α is 0.189, with standard error 0.061, this means that when the distance between two lakes is larger than 387 kilometers, their cross-ratio is lower than 1.05, so the related observations may be considered essentially independent.

4 Discussion

This paper proposes a model for the analysis of binary spatial data, in which dependence is accounted for though the specification of the cross-ratio that depends on the distance between observations. Conditional pairwise likelihood inference has the clear benefit of avoiding the specification of a joint model for the vector of binary spatial data.

The model is applied to the data on the reduction of the population of trouts in 542 Norwegian lakes. The analysis shows that the acid neutralizing capacity of the soil seems to increase the probability that the population of trouts in the lake remains stable.

Acknowledgements The first author acknowledges the “Progetto Giovani Ricercatori” of the University of Padova for providing financial support for a research project carried out at the Department of Statistical Sciences.

References

1. Carey, V., Zeger, S. L., Diggle, P.: Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526 (1993)

2. Heagerty, P. J., Zeger, S. L.: Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *J. Am. Statist. Ass.* **93**, 150–162 (1998)
3. Lin, P.-S., Clayton, M. K.: Analysis of binary spatial data by quasi-likelihood estimating equations. *Annals of Statistics* **33**, 542–555 (2005)
4. Lindsay, B. G.: Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239 (1988)
5. Lindsay, B. G., Yi, G. Y., Sun, J.: Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* **21**, 71–105 (2011)
6. Molenberghs, G., Lesaffre, E.: Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Am. Statist. Ass.* **89**, 633–644 (1994)
7. Molenberghs, G., Verbeke, G.: *Models for discrete longitudinal data*, Springer (2005)
8. Varin, C., Høst, G., Skare, Ø.: Pairwise likelihood inference in spatial generalized linear mixed models. *Comp. Stat. & Data Anal.* **49**, 1173–1191 (2005)
9. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42 (2011)
10. Zhang, H.: On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58**, 129–136 (2002)