Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2008/08

Space prediction models: an application to agricultural data

E. Dreassi, A. Petrucci, E. Rocco

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Space prediction models: an application to agricultural data

Dreassi E, Petrucci A, Rocco E

Dept. of Statistics "G. Parenti", University of Florence, Italy

Abstract

Space prediction, based on individual data, has been widely used in several applications and many advances have appeared in literature. This paper aim is to discuss an application of a hierarchical space Bayesian prediction model at unit level to an agricultural data set, where the geographical location of each unit is known and the response variable is a zero-inflated count variable. The results of our study show that, when a large amount of spatial heterogeneity is present in the data, prediction at unit level may be not suitable.

1 Introduction

In environmental analysis, over-dispersed and zero-inflated count variables are usually present. Space count regression models for individual data with this characteristics have been recently widely developed in literature and used on several application fields (i.e. Stein *et al.*, 1998 and Ridout *et al.*, 1998). The aim of this paper is to investigate the application of these methods to agricultural data. Precisely, the opportunity to use individual data is discussed analyzing the results of the application of a hierarchical space Bayesian model (Banerjee *et al.*, 2004) at individual level, taking into account zero inflated data, to predict surface area allocated to grapevines by farms in the province of Florence (Italy). The opportunity to use individual data is given by the Fifth Italian Agricultural Census driven in the year 2000 which first registered the geographical location of each farm (Bocci *et al.*, 2006).

This paper is organized as follows. Data are described in Section 2. The model used to predict the surface area allocated to grapevines is presented in Section 3. Results and final remarks are reported in Section 4.

2 Data

The Italian Statistical Institute (ISTAT) drives a two-yearly sample Farm Structure Survey (FSS). In this survey the unit of observation is the farm and for each farm are registered the data on the surface areas allocated to different crops. In this study we use the data on the farms of the Florence

province collected by the FSS driven in 2003 in order to test an individual prediction model of the surface area allocated to grapevines. We focus on its prediction for all the farms of province included in the Fifth Italian Agricultural Census driven in the year 2000. The most important feature of the Fifth Italian Agricultural Census is the registration of the geographical location of each farm which allows us to define a space model at unit level. Before the 2000 census the more detailed geographical information available for each farm was the name of the municipality (44 in the province of Florence). Thus only aggregate models at municipal level were usable.

Although the response variable, surface area allocated to grapevines, refers to one of the main cultivation of the Florentine area, it includes many structural zeros. The reason is that the morphological characteristics of the territory are not encouraging in all the province's area for grapevine cultivation and sometimes in specific sub areas the type of cultivated crop depend on the traditions, thus many farms do not cultivate grapevines. About the 25% of the farms in the 2003 FSS sample shows zero grapevine surface area; most of them can be considered as structural zeros. Not only the presence of many farms with zero surface area allocated to grapevine but also the presence of many farms with modest extent of it and few farms with a large grapevines surface made the variable highly positively skewed and over dispersed.

Besides the geographical location many other variables are available for each farm. Among them we select the surface area allocated to grapevines at 2000 census time, the surface area allocated to grapevines at 1990 census time and the European size unit (UDE) at 2000 census time as covariates for the specified model. The latter variable is a stratification variable in the FSS sample design and is based on the size of the farm. It is the stratification variable for the sample at the provincial level which is composed by a group of self-representative farms and the other sampled farms are arranged in three classes of UDE.

The mean grapevine surfaces, at aggregate municipality level, recorded at Census 2000, are showed in Figure 1. The FSS sample used to estimate the model's parameters results from merging three different sources of individual data and his final size is 214. Figure 2 shows locations about the farms included and not included in the FSS sample.

3 Statistical Model

We suggest a Hierarchical Bayesian model considering a ZIP formulation. The likelihood is a mixture of two distinct processes governing respectively the presence, or not, of area allocated to grapevines in the farm and, conditionally on being positive, the grapevines mean surface area at 2003. The first is a Bernoulli process and the second a truncated Poisson process.

The likelihood is re-parameterized as a mixture of two Poisson random variables. Following Lambert (1992) we define

$$L(Y_i) = \pi_i \; p_1 + (1 - \pi_i) \; p_2$$

where $i = 1, \ldots, 214$ indexes farm present in the FSS sample; p_1 is discrete with mass point at zero, p_2 is Poisson(λ_i) with mixing probability $1 - \pi_i$. For the Bernoulli process the logit of the probability having grapevines area is modelled considering grapevines area in 1990 and 2000 censuses

$$\text{logit}(\pi_i) = \alpha_b + \beta_{1990b} \times \text{area1990}_i + \beta_{2000b} \times \text{area2000}_i$$

For the Poisson process the log of the area is modelled as follow:

$$\log(\lambda_i) = \alpha_p + \beta_{2000p} \times \text{area2000}_i + \beta_{ude\ i} \times \text{ude2000}_i$$

where we specify a spatially structured Gaussian Exponential varying coefficient for ude2000 measures. $\beta_{ude\ i}$ is the component of the vector $\boldsymbol{\beta}_{ude}$ which is assumed to follow a MVN($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$) where MVN stands for Multivariate Normal distribution with vector mean $\boldsymbol{\mu}$, whose elements are assumed to follow a flat Normal$(0.0, 1000)$ distribution.

We defined a parametric distance function for the variance-covariance matrix $\boldsymbol{\Sigma}$. A common assumption (Cressie, 1993) is the exponential decay function:

$$\Sigma_{lj} = \sigma^2 \exp(-\phi \; d_{lj})^k$$

where $\sigma^{-2} \sim \text{Gamma}(0.1, 0.1)$ controls the overall variability, $\phi \sim \text{Uniform}$ $(0, 10)$ controls the rate of decline of correlation with distance d_{lj}, the Euclidean distance between pairs of farms locations l, j, and k controls the amount of spatial smoothing. We opted for a pure exponential model choosing $k = 1$ (see Diggle *et al.* 1998, page 323). Informative priors are specified on ϕ and σ^{-2} in order to get proper posteriors (Banerjee *et al.*, 2004: page 131).

Spatial interpolation is a problem of prediction in space. In particular, when we treat point data, the idea is to predict a new value of β_{ude} (say $\beta_{ude\ 0}$) at a new point location (i.e. the locations of 12058 farms not present on the FSS sample). This is straightforward in a Bayesian framework since we have just to figure out the predictive distribution

$$\begin{aligned} p(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \text{area2000}) &= \int p(\beta_{ude\ 0}, \boldsymbol{\theta} \mid \boldsymbol{\beta}_{ude}, \text{area2000}) \; d\boldsymbol{\theta} = \\ &= \int p(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \boldsymbol{\theta}, \text{area2000}) \times \\ &\times \; p(\boldsymbol{\theta} \mid \boldsymbol{\beta}_{ude}, \text{area2000}) \; d\boldsymbol{\theta} \end{aligned}$$

MCMC methods can be used taking advantage of the posterior sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(G)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{\beta}_{ude}, \text{area2000})$ and the conditional normal distribution $p(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \boldsymbol{\theta}, \text{area2000})$ arising from the joint multivariate distribution of $\boldsymbol{\beta}_{ude}$ and $\beta_{ude\ 0}$ (Banerjee *et al.*, 2004: page 132). The predictive

integral may be computed as a Monte Carlo mixture of the form

$$\hat{p}(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \text{area2000}) = \frac{1}{G} \sum_{g=1}^{G} p(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \boldsymbol{\theta}^{(g)}, \text{area2000})$$

In practice bypassing mixture calculation on use composition sampling drawing $\beta_{ude\ 0}^{(g)}$ from $p(\beta_{ude\ 0} \mid \boldsymbol{\beta}_{ude}, \boldsymbol{\theta}^{(g)}, \text{area2000})$.

We used WinBugs software for the MCMC estimation algorithm (see Spiegelhalter *et al.*, 2000). Convergence has been assessed using Gelman and Rubin (1992) convergence test.

4 Results and final remarks

For each farm on the sample we estimated mixing probability. ZIP model for sampled farms shows about 23% of zero grapevines area according to descriptive analysis on the data.

Exponentialized posterior mean for parameters of the proposed model are very close to one for β_{2000p} and 1.95 for α_p.

Posterior mean for parameter of the Gaussian spatial Exponential model are 0.022 for ϕ and 257.43 for σ. The low value for $\hat{\phi}$ and the high value for $\hat{\sigma^2}$ suggest the presence of a great spatial heterogeneity. This obviously affects the spatial distribution of the estimated $\boldsymbol{\beta}_{ude\ 0}$ coefficients; which posterior distribution means are reported in Figure 3.

The great spatial heterogeneity affects dramatically the results of our study and the scenario could even get worse considering aggregated data. Moreover, the implementation of a model with aggregated data does not satisfy the demand of estimates at small area level which is the primary advantage of using individual data. In fact the estimates can be given only at the municipality (or groups of municipalities) level.

We are aware that the results show a not completely good performance of the applied model. In any case, the study copes with a practical situation where the main challenge is to investigate the spatial heterogeneity of the data. The spatial heterogeneity is a critical features of the data and affects the effective prediction model even in presence of detailed auxiliary information.

Acknowledgement The research was partially supported by COFIN-2005 (PRIN 2005132407) and COFIN-2006 (PRIN 2006131039).

References

Banerjee S, Carlin BP, Gelfand AE (2004) *Hierarchical modeling and analysis of spatial data.* Chapman & Hall/CRC, New York.

Bocci C, Petrucci A, Rocco E (2006) Geographically Weighted Regression for Small Area Estimation: An Agricultural Case Study. Paper presented at the Italian Statistical Society 43^{th} Scientific Meeting. Torino. Italy.

Cressie NAC (1993) *Statistics for spatial data*. John Wiley & Sons, Inc.

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society C (Applied Statistics)*, **47**, 299–350.

Gelman A, Rubin DR (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Lambert D (1999) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturings. *Technometrics*, **34**, 1, 1–14.

Ridout M, Dometrio CGB, Hinde J (1998) Models for cout data with many zeros. Invited paper presented at the Ninetheenth International Biometric Conference. Cape Town. South Africa. 179–190.

Spiegelhalter DJ, Thomas A, Best NG, Gilks WR (2000) *WinBugs*. Medical Research Council Biostatistics Unit: Cambridge.

Stein A, Van Groenigen JW, Jeger MJ, Hoosbeek MR (1998) Space-time statistics for environmental and agricultural related phenomena. *Environmental and Ecological Statistics*, **5**, 2, 155–172.

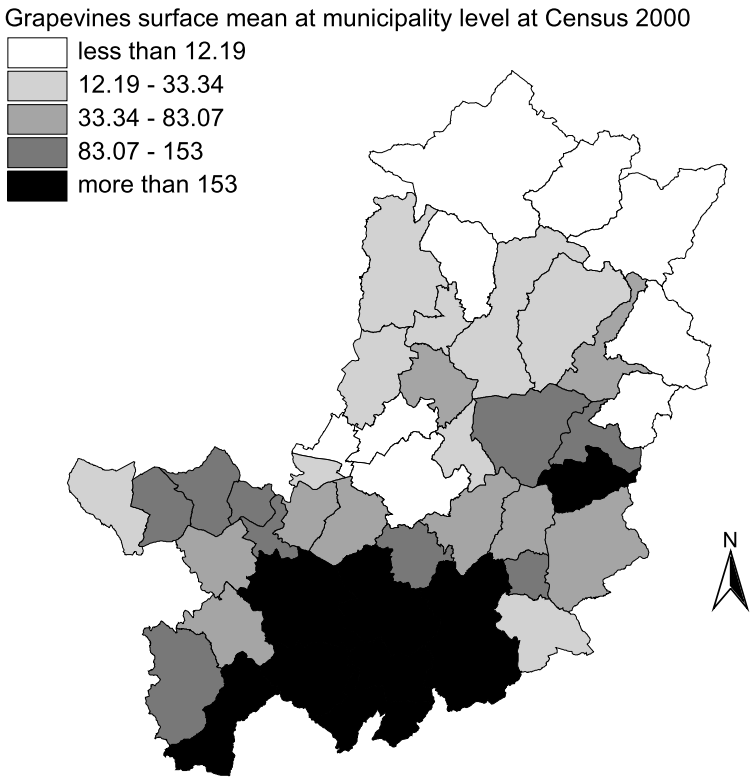

Figure 1. Data on grapevines surface mean at municipality aggregate level at Census 2000

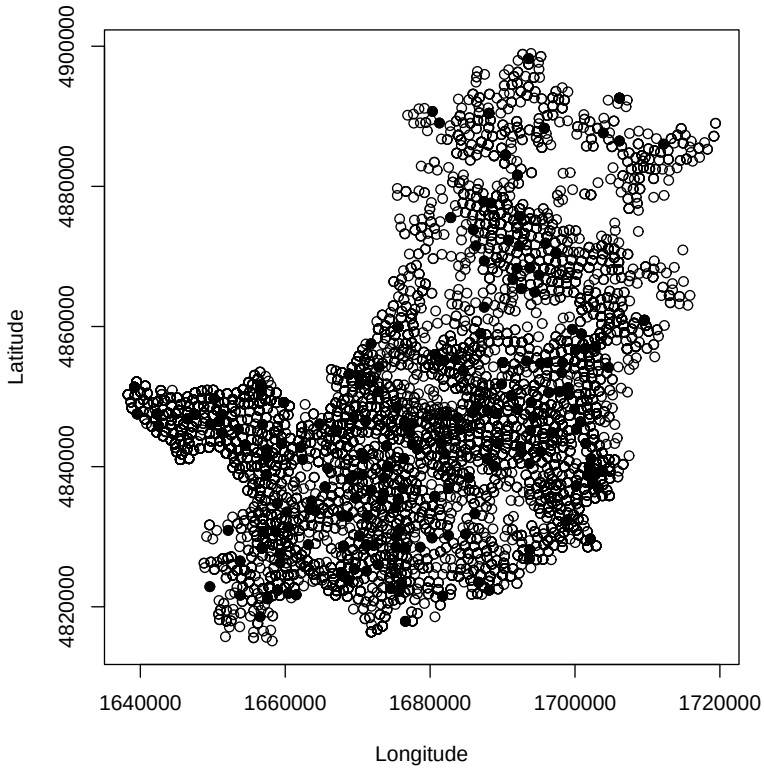

Figure 2. Data: • farms present and ○ farms not present on 2003 FSS sample

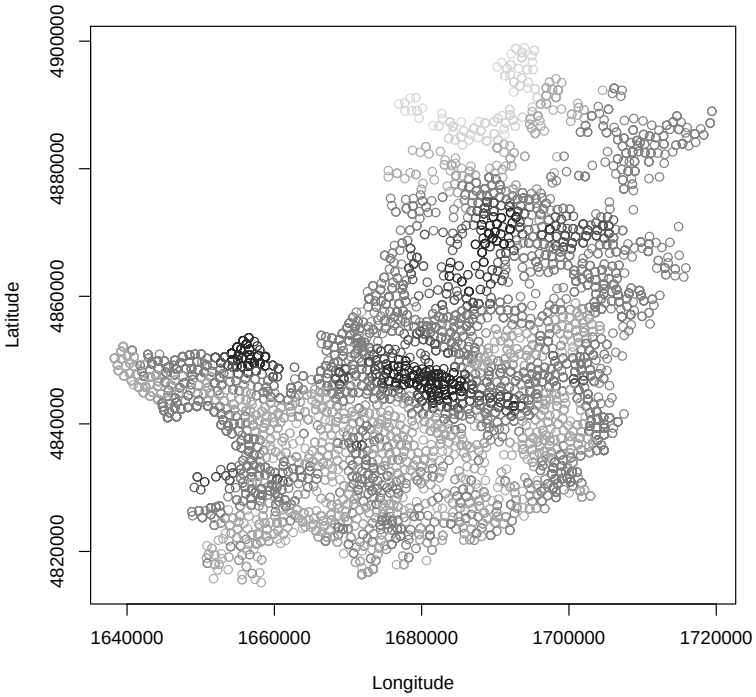

Figure 3. Predicted spatially varying coefficient $\boldsymbol{\beta}_{ude\ 0}$