ATTI DELLA XLVI RIUNIONE SCIENTIFICA (2012)

XLVI Riunione Scientifica
(Roma, 20-22 giugno 2012)
ISBN 978-88-6129-882-8 Editore: CLEUP – Padova

M. Gallo
Log-ratios analysis to study the relative information of ordinal variables

G. Boari, G. Cantaluppi
A PLS algorithm version working with ordinal variables

Better integrating geographic information and statistics
C. Bocci, A. Petrucci, E. Rocco
A two part geoadditive model for geographical domain estimation

J. Gallego, M. El-Aydam
Using coarse resolution satellite images for crop area estimation: benchmarking their efficiency

L. Barabesi, S. Franceschi, M. Marcheselli
How to select sample sites onto a study area?

Evaluation and quality in the university system
M. Porcu, I. Sulis
Comparing degree programs using unadjusted performance indicators. Assessing the bias from the Potential Confounding Factors

L. Grilli, C. Rampichini, R. Varriale
University admission test and students' careers: an analysis through a regression chain graph with a hurdle model for the credits

M. Attanasio, M. Enea
Bivariate logistic models for the analysis of the Students University "Success"

L. Masserini, M. Pratesi
University of Pisa and academic performance: a sample survey on students with no exams in 2011

Robustness issues in complex data analysis
M. Riani, A. Cerioli, M.A. Milioli, G. Morelli
A comparison of robust methods with small sample experimental data

C. Agostinelli, R. A. Maronna, V.J. Yohai
Robust estimation for multivariate data under the independent contamination model

P. Coretto
Adaptive robust location-scale estimation

G. Porzio, G. Ragozini, S. Liebscher, T. Kirschstein
Minimum Volume Peeling: a Multivariate Mode Estimator

Approximate likelihood methods for highly-structured statistical models
L. Ippoliti, L. Romagnoli
Kalman Filter for Maximum Likelihood Estimation of GMRFs

S. Bianconcini, S. Cagnone, D. Rizopoulos
A Dimension Reduction Method for Approximating Integrals in Latent Variable Models for Binary Data

M. Minozzo, C. Ferrari
Monte Carlo Likelihood Inference in Multivariate Model-Based Geostatistics

A. Davison, S. Padoan, M. Ribatet
Statistical Modelling of Spatial Extremes

Statistical methods for network data
A. Petrucci, E. Rocco
Statistical models for virtual water network analysis

D. De Stefano
Graph embedding via dissimilarity mapping for network comparison

V. Amati
New Statistics for Estimating the Parameter of the Stochastic Actor-Oriented Model

A. Abbruzzo, E. Wit
Factorial Graphical Lasso and Slowly Changing Graphical Models for Estimating Dynamic Networks

Frequency domain analysis
S. Pollock
Cycles Syllogisms and Semantics: Examining the Idea of Spurious Cycles

Matteo Farne'
Different estimators of the spectral matrix: an empirical comparison. Testing a new shrinkage estimator

M. Donatelli, A. Luati, A. Martinelli
Spectral filtering for trend estimation

Family demography and well being
A. Baranowska-Rataj, A. Matysiak
Which family model makes couples more happy - dual earner or male breadwinner ?

A two-part geoadditive model for geographical domain estimation.

Chiara Bocci, Alessandra Petrucci, Emilia Rocco

Abstract Small area estimation (SAE) based on linear mixed models can be inefficient to deal with semicontinuous variables that have a portion of values equal to zero and a continuous, often skewed, distribution among the remaining values. Moreover when small domains of study are geographical areas an adequate use of geographic information and geographical modeling can provides more accurate estimates for small area parameters. In this paper we suggest a two-part geoadditive small area model that can deal with these issues, frequently addressed in many fields of applied research, among which the agricultural one. The performance of the suggested model is then illustrated by an application to agricultural data: the model is used to estimate the per farm average grapevine production in Tuscany at Agrarian Region level.

Key words: Generalized linear mixed model, Geographic information, Semicontinuous data, Small area estimation.

1 Introduction

In many agricultural applications, a response variable may have a continuous distribution whit a large number of values clustered at zero. In literature the "excesses" zeros" in data are usually described by the zero inflated (ZI) regression models that mix a degenerate distribution with point mass of one at 0 with a simple regression model based on a standard distribution. This is realized considering a pair of regres-

Chiara Bocci
IRPET Istituto Regionale Programmazione Economica Toscana - Villa La Quiete alle Montalve - Via Pietro Dazzi 1 - 50141 Firenze , e-mail: chiara.bocci@irpet.it

Alessandra Petrucci and Emilia Rocco
Dept. of Statistics - University of Florence, Viale Morgagni 59 - 50134 Firenze
email alessandra.petrucci@unifi.it – rocco@ds.unifi.it

sion models: a model, usually logit or probit, for the probability of nonzero response and a conditional linear model for the mean response given that it is nonzero. The ZI models has been originally developed to analyze count data, but then are also extended to situation in which a huge number of zeros occur in continuous data (Holsen and Shafer, 2001; Gosh and Albert, 2009) and their application is common in zero inflated lognormal model with skewed semicontinuous data. Frequently, in the context of semicontinuous data these mixture models are referred to as two-part models. A two-part SAE model may be defined replacing the pair of regression models above mentioned with a pair of classic SAE models that is a pair of generalized linear mixed models that include independent random area effects to account for between area variation beyond that explained by auxiliary variables. Nevertheless classic SAE models assume independence of the area-specific random effects. If the small domains of study are geographical areas, this assumption means that any possible spatial structure of the data is not take into account.Therefore it may be unreasonable and an adequate use of geographic information and geographical modeling may be opportune. A geoadditive model allow to analyzes the spatial distribution of the study variable while accounting for possible covariate effects through a linear mixed model representation (Kammann and Wand, 2003). The linear mixed model structure allows to include the area-specific effect as an additional random components. In particular, a geoadditive SAE model has two random effect components: the area-specific effects and the spatial effects.The aim of this paper is to suggest an approach in order to predict the mean value at some domain of interest of a variable that present a semicontinuous structure and a spatial related pattern. A two-part model composed by a pair of geoadditive small area models is suggested and applied to estimate the per farm average grapevine production in Tuscany at Agrarian Region level.

2 Methods

Let y_{ij} denote a non-negative semicontinuous skewed response variable for the unit j $(j = 1,...,N_i)$ in small area i $(i = 1,...,m\ \sum_{i=1}^{m} N_i = N)$, x_{ij} a vector of p linear covariates associated with the same unit and $\mathbf{s}_{ij}$ $(\mathbf{s} \in R^2)$ the spatial location of the unit. We assume that the response variable has a significant spatial pattern and can be recoded as two variables,

$$I_{ij} = \begin{cases} 1 \text{ if } y_{ij} > 0 \\ 0 \text{ if } y_{ij} = 0 \end{cases} \tag{1}$$

and

$$y'_{ij} = \begin{cases} y_{ij} & \text{if } y_{ij} > 0 \\ \text{irrelevant} & \text{if } y_{ij} = 0 \end{cases} \tag{2}$$

We model these responses by a pair of uncorrelated geoadditive small area models. One for the logit probability of $I_{ij} = 1$ and one for the conditional mean of the logarithm of the response $E[log(y'_{ij})|I_{ij} = 1]$. The two models using the linear mixed model representation, can be written respectively as:

$$\eta = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\mathbf{u} \tag{3}$$

and

$$\log(\mathbf{y}') = \mathbf{X}^*\beta^* + \mathbf{Z}^*\gamma^* + \mathbf{D}^*\mathbf{u}^* + \varepsilon \tag{4}$$

where $\eta_{ij} = log(\pi_{ij}/(1+\pi_{ij}))$, $\pi_{ij} = P(I_{ij} = 1)$, $\mathbf{y}'$ is the vector containing all relevant y_{ij} values, the ones corresponding to $I_{ij} = 1$; $\mathbf{u}$ and $\mathbf{u}^*$ are the the vectors of the area specific random effects in the two model, γ and γ^* are the coefficient vectors for the "spline" portion of the two models, β and β^* are the coefficients vectors for the "parametric" portion of the two models, $\mathbf{X}$ and $\mathbf{X}^*$ are the matrices of covariates relating to the fixed effects, $\mathbf{Z}$ and $\mathbf{Z}^*$ are the matrices of covariates concerning the random effects due to the spline, $\mathbf{D}$ and $\mathbf{D}^*$ are the matrices of covariates concerning the random effects due to the small area, ε is the second model the residuals vector. In our model the same set of covariates may appear in the logit and loglinear parts. Even if the same covariates are used in both parts, it will be not generally true that $\mathbf{X} = \mathbf{X}^*$, $\mathbf{Z} = \mathbf{Z}^*$, $\mathbf{D} = \mathbf{D}^*$ because model (4) applies only when $y_{ij} > 0$.
For independent semi-continuous data all observed zero are unambiguous, they necessarily came from the degenerate distribution, rather than from the nondegenerate continuous distribution. The likelihood for such a model factors into terms for the zero and nonzero data, so that it is equivalent to separately model the nonzero data and an indicator variable for wether or not the response is zero. Unfortunately this simplification may not occur for clustered data because the cluster specific random effects into the two models may be correlated. In a recent paper, Zhang et al. (2006) compared the parameter estimates obtained adopting a two part hierarchical model with a correlated random effects structure with those obtained fitting separately the two models and showed that they are similar. Looking at these results, in the application described in next section we assume that the random effects relative to the two models, one due to the logit probability and the other to the logarithm of the mean conditional response, are uncorrelated.

3 Data, preliminary results and comments

The Italian Statistical Institute (ISTAT) drives an Agricultural Census ten-yearly and a sample Farm Structure Survey (FSS) two-yearly. Both in the Census and in the FSS, the unit of observation is the farm and the data of the surface areas allocated to different crops are registered for each farm. In the FSS, until 2005, the productions of each crop were also observed. The FSS survey is designed to obtain estimates only at regional level. We are interest in producing the mean estimation of grapevine production for the 52 Agrarian Regions in which Tuscany region is par-

titioned. A large number of farms don't cultivate grapevines, and a few produce the majority of the total region production. Moreover the cultivation and consequently the production of grapevines for each farm depends on the characteristics of the territory in which the farm is located. Finally, the quantity of grapevine produced by the same allocated surface may change, depending on the soil productivity and on the production choices of the farms (relative to the typology and quality of the produced grapevine). These practical considerations, confirmed by an explorative analysis of the data, motivate our choice of a two part geoadditive SAE model. The model is applied to produce estimates referred to the 2003 year for which the data of the FSS Survey are available. Auxiliary variables and spatial information for each farms referred to 2000 census time. As main results we obtaine the predicted agrarian regions means which are showed in Figure 1.It is evident a evident geographical pattern, with the higher values in the areas belonging to the provinces of Florence and Siena (the well known zone of Chianti) and the lower values in the north mountainous area of the provinces of Massa Carrara and Lucca.

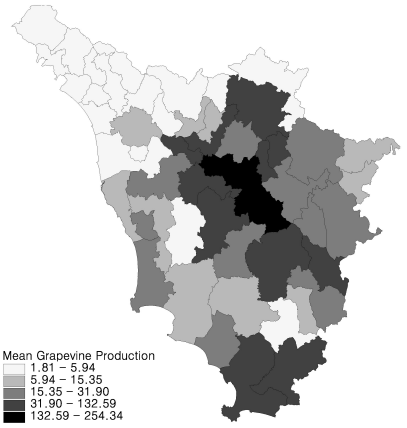

Fig. 1 Agrarian region level estimates of the mean grapevine production.

References

1. Gosh, P and Albert, P.S.: A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. Comput. Stat. Data An. **53**, 699–706 (2009)
2. Kammann, E.E. and Wand, M.P.: Geoadditive Models. Applied Statistics, **52**, 1–18 (2003)
3. Olsen, M.K. and Schafer, J.L.: A two-part random-effects model for semicontinuous longitudinal data. J. Am. Statist. Assoc. **96**, 730–745 (2001)
4. Zhang, M., Strawederman, R. L., Cowen, M.E., Wells, M.E.: Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. J. Am. Statist. Assoc. **101**, 934–945 (2006)