Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 66(1) 2012 135-144

Small Area Methods for Agricultural Data: A Two-Part Geoadditive Model to Estimate the Agrarian Region Level Means of the Grapevines Production in Tuscany

C. Bocci, A. Petrucci*, and E. Rocco
Department of Statistics "G. Parenti", University of Florence,
Viale Morgagni, 59 - 50134 Firenze, Italy

Received 02 May 2011; Revised 31 August 2011; Accepted 01 September 2011

SUMMARY

In applications involving agricultural data, it is common to encounter semicontinuous variables that have a portion of values equal to zero and a continuous, often skewed, distribution among the remaining values. Moreover, these variables often show a spatial pattern. We develop a two-part geoadditive small area model that can deal with these issues. In particular, we are interested in predicting the mean of a target variable with these characteristics for a collection of subsets of the population. Direct estimation using only the survey data is inappropriate as it yields to estimates with unacceptable levels of precision. A study of the Tuscan Agrarian Region (Italy) level means of the grapevines production illustrates this method.

Keywords : Generalized linear mixed model, Penalized splines, Semicontinuous data, Spatial dynamics, Zero-inflated data.

1. INTRODUCTION

In small area estimation the interest is usually on the estimation of a parameter of a small area such as the mean or the total of a variable *y*. Traditional area specific (direct) estimates may not provide acceptable precision for small areas because sample sizes are seldom large enough in many small areas of interest. This makes it necessary to borrow information across related areas through indirect estimation based on models, using auxiliary information such as recent census data and current administrative data. The most popular class of models for small area estimation (SAE) is linear mixed models that include independent random area effects to account for between area variation beyond that explained by auxiliary variables (Fay and Herriot 1979; Battese *et al*. 1988).

Under the classic SAE model we make the assumption of independence of the area-specific random effects. If the small domains of study are geographical areas, this assumption means that we don't take into account any possible spatial structure of the data. It is reasonable, recalling the "first law of geography", to suppose that close areas are more likely to have similar values of the target parameter than areas which are far from each other. Thus an adequate use of geographic information and geographical modeling can provide more accurate estimates for small area parameters. In addition, Pratesi and Salvati (2008) pointed out that geographical small area boundaries are generally defined according to administrative criteria without considering the eventual spatial interaction of the variable of interest. Given all these notes, it is reasonable to assume that the random effects between the neighboring areas (defined, for example, by a contiguity criterion) are correlated and the correlation decays to zero as distance increases.

* *Corresponding author :* A. Petrucci
E-mail address : alex@ds.unifi.it

The first studies that connect spatial relationships and SAE methods are Cressie (1991) and Pfeffermann (2002). In the following years, many papers have been published showing how the use of geographical information improves the estimation of the small area parameters, both increasing efficiency and diminishing bias. We refer, among others, to Petrucci and Salvati (2006), Singh *et al*. (2005) and Pratesi and Salvati (2008). In all these studies, the classical hypothesis of independence of the random effects is overcome by considering correlated random area effects between neighboring areas modeled through a simultaneously autoregressive (SAR) process with spatial autocorrelation coefficient ρ and a proximity matrix $\mathbf{W}$ (Anselin 1988). However, following this approach, only the spatial structure of the data at the area level is considered and the information about spatial contiguity of the small areas is used to built the proximity matrix of the SAR process. When the spatial location is available for every unit, the geographical coordinates can be utilized as covariates of the SAE model.

A geoadditive model 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 current Italian agricultural surveys fall in this general framework where the "small area" commonly refers to a local geographical area such as province, municipality or municipality aggregations. Sample sizes are usually too small to provide reliable direct estimates for these small areas as the surveys are usually planned considering regions as the more detailed estimation level. The use of models for small area estimation becomes crucial. Moreover, the spatial location of the statistical unit (i.e. farm) to which the agricultural variables (i.e. land by type of cultivation, amount of breeding, productions, structure and amount of farm employment) are referred is relevant in the analysis. The spatial location of each farm has been collected for the first time during the fifth Italian Agricultural Census driven in year 2000 introducing a new challenge for the statistical analysis.

Often agricultural variables present a semicontinuous structure, which means that a variable has a fraction of values equal to zero and a continuous, often skewed, distribution among the remaining values. In literature the "excess 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 regression 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 developed to analyze count data, examples include regression models for zero inflation relative to a Poisson (ZIP), zero inflated negative binomial (ZINB) and zero inflated binomial (ZIB). Lambert (1992), Hall (2000), Ridout *et al*. (2001) among others, had largely studied these models.

ZI models are also suggested when 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.

In this paper we present a two-part geoadditive SAE model to estimate the per farm average grapevine production in Tuscany at Agrarian Region level. The two part-model and the geoadditive small area model are considered separately in the literature, here we combine them considering the grapevine production as a semicontinuous skewed variable.

The article is organized as follows. The modeling approach is presented in Section 2 while the application of the model to agricultural real data is described in Section 3. The conclusions in Section 4 point out some problems found in the application and the future research goals.

2. TWO-PART GEOADDITIVE SMALL AREA MODEL

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, we consider a two-part model composed by a pair of geoadditive small area models. The geoadditive model allows us to analyze the spatial distribution of the study variable while accounting for possible covariate effects through a linear mixed model representation, that permits to include the area-specific effect as an additional random component. The geoadditive small area model, which corresponds to a particular specification of the non-parametric SAE model introduced by Opsomer *et al*. (2008), has two random effect components: the area-specific effects and the spatial effects.

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

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

and

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

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 $E[\log(y'_{ij}) \mid I_{ij} = 1]$.

The logit model is

$$\eta_{ij} = \alpha + \mathbf{x}_{ij}^T \boldsymbol{\beta}_x + h(\mathbf{s}_{ij}) + u_i \tag{1}$$

where $\eta_{ij} = \log(\pi_{ij} / (1 - \pi_{ij}))$, $\pi_{ij} = \mathrm{P}(I_{ij} = 1)$, h is an unspecified bivariate smooth function and u_i is the area specific random effect. Representing $h(.)$ with a low rank thin plate spline (Ruppert *et al*. 2003, p.253) with K knots

$$h(\mathbf{s}) = \beta_{0s} + \mathbf{s}^T \boldsymbol{\beta}_s + \sum_{k=1}^{K} \gamma_k b_{tps}(\mathbf{s}, \boldsymbol{\kappa}_\mathrm{k})$$

model (1) can be written as a mixed model (Kammann and Wand 2003)

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} \tag{2}$$

with $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_K)$and $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{I}_K)$ and where

- $\mathbf{X} = \left[1, \mathbf{x}_{ij}^T, \mathbf{s}_{ij}^T\right]$ is the fixed effect matrix for the N population units;
- $\boldsymbol{\beta} = \left[\beta_0, \boldsymbol{\beta}_x^T, \boldsymbol{\beta}_s^T\right]$ with $\beta_0 = \alpha + \beta_{0s}$ is the coefficients vector for the "parametric" portion of the model;
- $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_K]$ is the coefficients vector for the "spline" portion of the model;
- $\mathbf{u} = [u_1, ..., u_m]$ is the vector of the area specific random effect;
- $\mathbf{D} = [d_1, ..., d_N]^T$ with $d_i = (d_{i1}, ..., d_{im})^T$ and d_{ij} an indicator taking value 1 if observation j is in small area i and 0 otherwise;
- $\mathbf{Z} = [b_{tps}(\mathbf{s}, \boldsymbol{\kappa}_k)]_{N \times K} = \left[C(\mathbf{s}_{ij} - \boldsymbol{\kappa}_k)\right]_{\substack{1 \le ij \le N \\ 1 \le k \le K}} \times \left[C(\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k)\right]_{1 \le h, k \le K}^{-1/2}$ where $C(v) = \|v\|^2 \log\|v\|$ and $\boldsymbol{\kappa}_1, ..., \boldsymbol{\kappa}_K$ are the knots location of the spline function.

The model for the continuous response is

$$\log(\mathbf{y}') = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{Z}^* \boldsymbol{\gamma}^* + \mathbf{D}^* \mathbf{u}^* + \boldsymbol{\varepsilon} \tag{3}$$

where

- $\mathbf{y}'$ is the vector of length N^* containing all relevant y'_{ij} values, the ones corresponding to $I_{ij} = 1$;
- the residuals $\boldsymbol{\varepsilon}$ are assumed to be distributed as $\mathrm{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N^*})$;
- $\boldsymbol{\gamma}^* = \left[\gamma_1^*, ..., \gamma_K^*\right]$ is the coefficient vector for the "spline" portion of the model;
- $\mathbf{u}^* = \left[u_1^*, ..., u_m^*\right]$ is the vector of the area specific random effect;
- $\boldsymbol{\gamma}^* \sim \mathrm{N}\left(0, \sigma_{\gamma^*}^2 \mathbf{I}_K\right)$ and $\mathbf{u}^* \sim \mathrm{N}\left(0, \sigma_{u^*}^2 \mathbf{I}_K\right)$.

$\boldsymbol{\beta}^*$ is the coefficients vector for the "parametric" portion of the model, $\mathbf{X}^*$, is the matrix of covariates relating to the fixed effects and $\mathbf{Z}^*$ and $\mathbf{D}^*$are the matrices of covariates concerning the random effects due to the spline and to the small area respectively. 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 $\mathbf{X}^* = \mathbf{X}$, $\mathbf{Z}^* = \mathbf{Z}$ and $\mathbf{D}^* = \mathbf{D}$ because model (3) applies only when $y_{ij} > 0$.

The loglikelihood under independence for semicontinuous not clustered response variables is

$$l = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \log\left[\pi_{ij}(\theta)^{I_{ij}} \; (1 - \pi_{ij}(\theta))\}^{1 - I_{ij}}\right] + \sum_{i=1}^{m} \sum_{j=1}^{N_i} I_{ij} \log(f_2(\theta^*))$$

where θ and θ^* denote the parameters for the two models and f_2 is the generic model assumed for the nonzero elements (not necessarily lognormal). The ML estimation for such a model can be accomplished by separately fitting a binomial regression model to the indicator variable $I_{ij}(i = 1,\ldots m;\ j = 1,\ldots., N_i)$ and a model based on f_2 to the nonzero y_{ij} elements. The simplification respect to the general ZI models for count data can be done in the presence of semicontinuous independent data. In this case a continuous distribution has a null probability of yielding at zero and the distribution in the mixture of each response is simply defined by its own value.

Unfortunately this simplification may not occur for clustered data (Olsen and Shafer 2001; Berk and Lachenbruch 2002) because the cluster specific random effects into the two models may be correlated. In a recent paper, Zhang *et al.* (2006) apply a two part hierarchical model with a correlated random effects structure to analyze profiling providers in managed health care. In order to evaluate the model assumptions a comparison between the results and those obtained fitting separately the two models is carried out showing that the parameters estimated are similar (with the exception of the correlation parameter that cannot be estimated in the second case). Looking at these results, 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.

Now suppose to observe a sample from the population, with the covariates and the spatial location known for all the population units, the small area mean can be estimated using the model-based mean estimator

$$\hat{\bar{y}}_i = \frac{1}{N_i}\left[\sum_{j\in S_i} y_{ij} + \sum_{j\in R_i} \hat{y}_{ij}\right] \tag{4}$$

where S_i and R_i indicate the sets of the sampled and non-sampled units belonging to region i.

The predicted values are

$$\hat{y}_{ij} = \hat{\pi}_{ij}\tilde{y}_{ij} \text{ with } \tilde{y}_{ij} = \hat{\lambda}_{ij}^{-1}\exp\left(\hat{\phi}_{ij} + \frac{\hat{v}_{ij}}{2}\right) \tag{5}$$

where $\hat{\phi}_{ij} = x_{ij}^{*T}\hat{\beta}^* + z_{ij}^{*T}\hat{\gamma}^* + u_i^*$, $\hat{v}_{ij} = \hat{\sigma}_e + \mathbf{d}_i^{*T}\hat{\sigma}_u^*\mathbf{d}_i^* + \mathbf{z}_i^{*T}\hat{\sigma}_\gamma^*\mathbf{z}_i^*$, $\mathbf{z}_i^*$ is the i-th row of matrix $\mathbf{Z}^*$ and $\exp\left(\hat{\phi}_{ij} + \frac{\hat{v}_{ij}}{2}\right)$ is the back log transformation. Since $\exp\left(\hat{\phi}_{ij} + \frac{\hat{v}_{ij}}{2}\right)$ is a design biased estimator, we introduce the factor $\hat{\lambda}_{ij}$ that is the bias adjustment suggested by Chandra and Chambers (2005) defined as

$$\hat{\lambda}_{ij} = 1 + \frac{1}{2}\left\{\hat{a}_{ij} + \frac{1}{4}\hat{V}\left(\hat{v}_{ij}\right)\right\}$$

where $\hat{a}_{ij} = \mathbf{x}_{ij}^{*T}\hat{V}\left(\hat{\beta}^*\right)\mathbf{x}_{ij}^*, \hat{V}\left(\hat{\beta}^*\right)$ is the usual estimator of $\text{Var}\left(\hat{\beta}^*\right)$ and $\hat{V}\left(\hat{v}_{ij}\right)$ is the estimated asymptotic variance of $\hat{v}_{ij}$. Under ML and REML of the variance components of (4), the estimated asymptotic variance is obtained from the inverse of the information matrix.

3. APPLICATION

3.1 Data and Model Estimation

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, therefore to obtain estimates at sub-regional levels it is necessary to employ indirect estimators that “borrow strength” from related areas. The indirect estimators can be based on regression models that use the variables collected at the census time as auxiliary variables, known for all the population units, and that can incorporate specific random area effects to account for the residual between area variation. Moreover, the Fifth Agricultural Census driven in year 2000 registered the farms location on the territory and this geographical location can be a particularly useful information for the analysis of many phenomena concerning the agricultural field (Bocci *et al.* 2006).

As mentioned before, we are interest in producing the mean estimation of grapevine production for the 52 Agrarian Regions in which Tuscany region is partitioned. The agrarian regions are sub-provincial aggregations of municipalities homogenous respect to natural and agricultural characteristics. The estimates are 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. Due to the high correlation values observed over sampled data between the explicative variables at years 2000 and 2003 (about 90% for the grapevines surface), we suppose that the time lag between the response and the explicative variables should have a negligible effect. The available spatial information consists in the universal transverse Mercator (UTM) geographical coordinates of each farm's administrative centre.

The nature of the study variable does not allow the use of classic small area methods that assume a linear mixed model and don't take into account the spatial structure of the data. 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 model: a logit model for the probability of nonzero grapevine production and a conditional log-linear model for the mean of nonzero grapevine production. The selection of the covariates among several socioeconomic variables (including land use information) available at the census time follows the indications obtained from a stepwise regression analysis of the data. For the logit model two auxiliary variables are considered: the surface allocated to grapevines in logarithmic scale and a dummy variable that indicate the selling of grapevine related products, both at 2000 census time. In the conditional log-linear model we include the same two variables plus the number of working days done by farm family members in the 2000 year.

Moreover, since both the choice to produce or not produce grapevines ($I_{ij} = 1$ or $I_{ij} = 0$) and the conditioned level of production depend on the characteristics of the farm's location, in both the models the response is assumed dependent on a smooth function of the UTM geographical coordinates of each farm's administrative center, that is we adopt a geoadditive model. Regarding the possibility to include into the model the specific small area random effect, it results significant only in the loglinear model. Therefore, recalling (2) and (3), our chosen models are

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

The splines knots are selected setting $K = 50$ and using the *clara* space ûlling algorithm of Kaufman and Rousseeuw (1990). The two models are estimated separately, with, the logit one fitted through the Penalized Quasi-Likelihood method using all the 2450 farms in the 2003 FSS sample, and the loglinear one fitted by maximizing the restricted log-likelihood and using only the 961 farms with a strictly positive value of grapevines production.

The resulting spatial smoothing of the probability of production π_{ij} and of the nonzero log-production $\log(\mathbf{y}')$ is presented in Fig. 1. From these maps, it is evident the presence of a spatial dynamic in the probability of grapevine production (first map) and of both a spatial dynamic and small area level effect in the level of grapevine production (second map).

The estimated models parameters (presented in Table 1) are combined with the census values of the 136817 non sampled farms using (5) to obtain the grapevine production predictions. Since in two agrarian regions there are no sampled farms with strictly positive value of grapevines production, in this regions the $\hat{\phi}_{ij}$ are calculated using the synthetic predictor $\hat{\phi}_{ij} = x_{ij}^{*T}\hat{\beta}^{*} + z_{ij}^{*T}\hat{\gamma}^{*}$. Finally, expression (4) is applied to obtain the predicted agrarian regions means showed in Table 2 and Fig. 2.

The map of the estimated agrarian region means presents an 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, confirming the pattern of the expert's estimate means produced by ISTAT. [Statistics are produced using expert information. Data are provided by local authorities that collect experts evaluations on area and yield of different crops. The auxiliary information could be included in expert's estimate, such as verifying the availability of external sources (*e.g.*, professional bodies or associations of producers, administrative sources, auxiliary sources of data related to the cultivation being estimated). (Source: "ISTAT Information system on quality of statistical production processes", http://siqual.istat.it/)]. These expert's estimates are

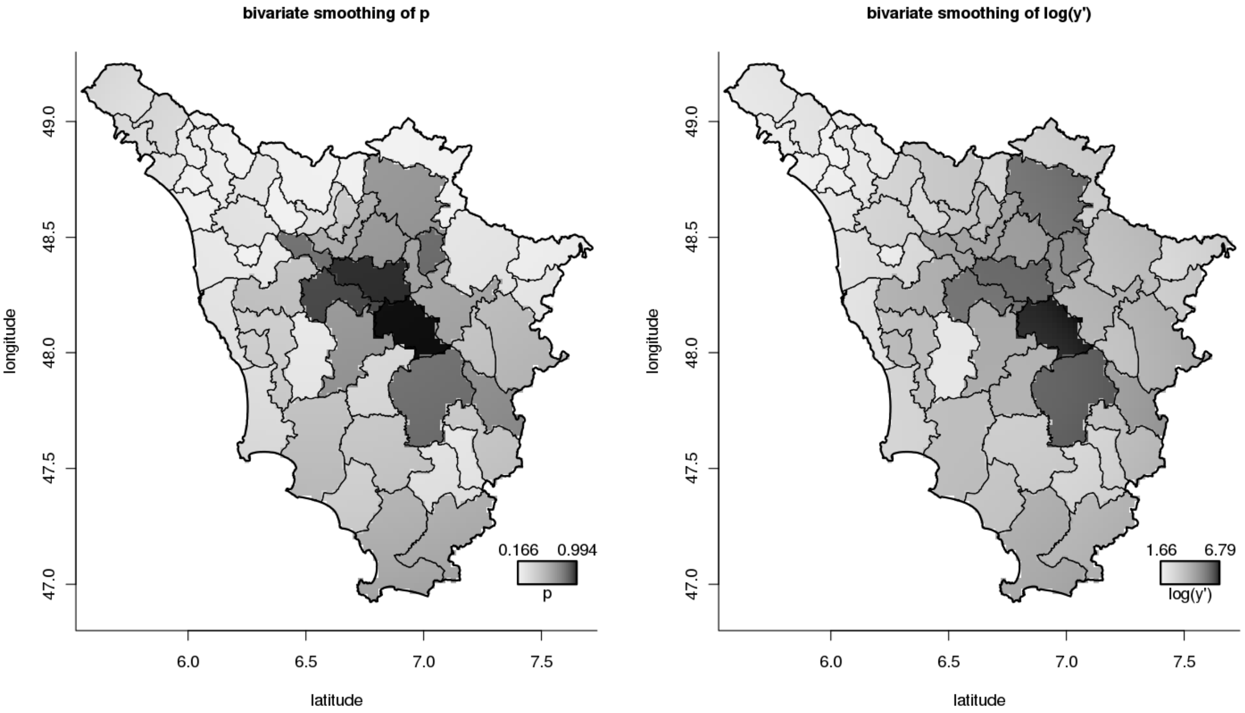

Fig. 1. Spatial smoothing of the probability of production π_{ij} and of the nonzero log-production log ($\mathbf{y}'$)

Table 1. Estimated parameters of the logit model and of the conditional log-linear model*.

Logit model			Log-linear model		
Parameters	Estimate	Confidence Interval	Parameters	Estimate	Confidence Interval
Fixed Effects			Fixed Effects		
Intercept	17.2292	(–23.818 ; 58.276)	Intercept	–0.5709	(–0.2501 ; 23.866)
X coordinate	0.0710	(–0.8436 ; 0.9857)	X coordinate	0.4730	(–0.0130 ; 0.9591)
Y coordinate	–0.3965	(–1.1956 ; 0.4026)	Y coordinate	–0.0081	(–0.5179 ; 0.5018)
log(grapevine surface)	1.9745	(0.9118 ; 3.0372)	log(grapevine surface)	1.2694	(1.2059 ; 1.3328)
grapevine products selling	1.0636	(0.0358 ; 2.0915)	grapevine products selling	0.6701	(0.5163 ; 0.8239)
			family members working days	0.0004	(0.0002 ; 0.0006)
Random Effects			Random Effects		
σ_γ	0.2124	(0.0204 ; 2.2059)	σ_γ^*	0.2394	(0.0795 ; 0.7204)
σ_ε	2.9930	(2.9102 ; 3.0781)	σ_u^*	0.2189	(0.1242 ; 0.3855)
			σ_ε^*	0.8973	(0.8570 ; 0.9396)

* Intercepts and coordinates coefficients are not significant, but required by the model structure.

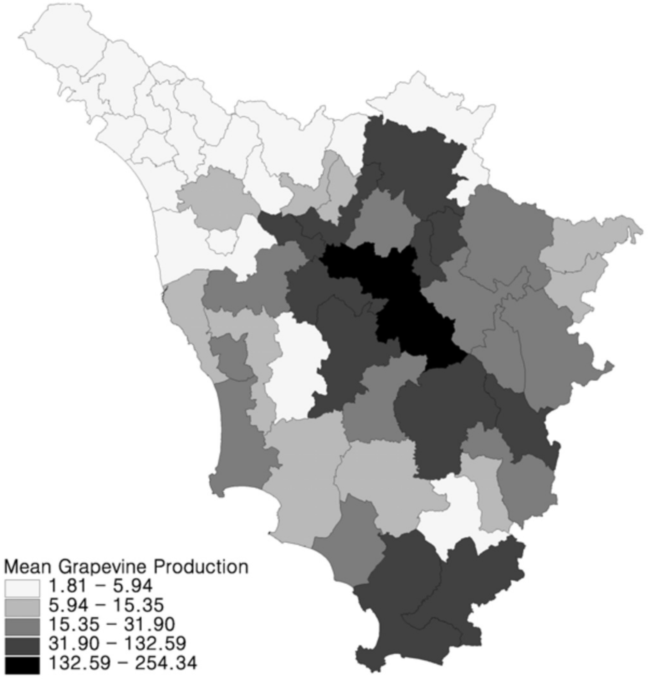

Fig. 2. Agrarian region level estimates of the mean grapevine production.

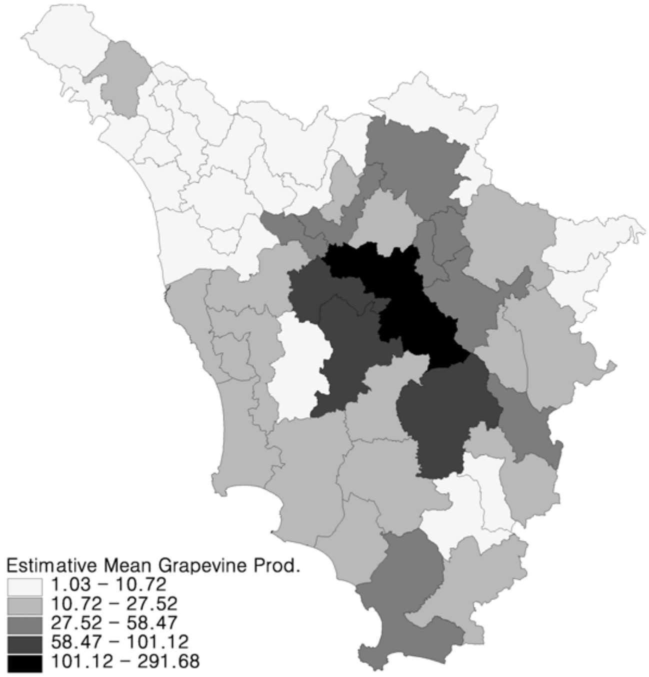

Fig. 3. Expert's estimates of the mean grapevine production at agrarian region level.

obtained by determination of a crop specific coefficient of soil productivity and are released at provincial level. To better compare them with our results, we calculate the agrarian region level expert's estimate (shown in Fig. 3) by multiplying the agrarian region grapevine surfaces at year 2000 with the coefficient of soil productivity at regional level. The comparison of Fig. 2 and Fig. 3 confirms that our estimates present the same pattern of the expert's estimates.

3.2 Variability Measure

In order to give a first evaluation of the uncertainty associated with our predictions we computed the predicted mean-squared error (PMSE) for the small area estimates by using a parametric bootstrap with 1000 replications.

Bootstrap replicate observations are generated as

$$\boldsymbol{\eta}^b = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^b$$

$$\log(\mathbf{y}^b) = \mathbf{X}^*\hat{\boldsymbol{\beta}}^* + \mathbf{Z}^*\boldsymbol{\gamma}^{*b} + \mathbf{D}*\mathbf{u}^{*b} + \boldsymbol{\varepsilon}^b \qquad (6)$$

where $\boldsymbol{\gamma}^b$, $\boldsymbol{\gamma}^{*b}$, $\mathbf{u}^{*b}$ and $\boldsymbol{\varepsilon}^b$ are bootstrap replicates of the random components in the model. In principle there are various possibilities to draw such replicates. A natural way to do this is to make use of the stochastic models

$$\boldsymbol{\gamma}^* \sim \mathrm{N}\left(\mathbf{0}, \hat{\sigma}^2_{\gamma^*}\mathbf{I}_K\right)$$

$$\boldsymbol{\gamma} \sim \mathrm{N}\left(\mathbf{0}, \hat{\sigma}^2_{\gamma}\mathbf{I}_K\right) \text{ and } \mathbf{u}^* \sim \mathrm{N}\left(\mathbf{0}, \hat{\sigma}^2_{u^*}\mathbf{I}_K\right)$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \hat{\sigma}^2_{e}\mathbf{I}_{N^*}\right)$$

with fitted variance parameters.

Once the bootstrap random components and errors have been generated, the linear predictors $\boldsymbol{\eta}^b$ and $\log(\mathbf{y}^b)$ are constructed by using equation (6) and then $\boldsymbol{\pi}^b$ and y^b are obtained using inverse logit transformation and unbiased log-back transformation Finally the values of the indicator variable $\mathbf{I}^b$ are generated performing for each unit a Bernoulli experiment with probability of success equal to the corresponding π^b_{ij}, $\mathbf{y}'^b$ is obtained as $y'^b = \mathbf{I}^b\, y^b$ and the mean of its values for each area $\tilde{\mathbf{y}}'^b_i$ are evaluated.

Drawing B bootstrap samples obtained in this way, the PMSEs for the small areas estimates (showed in Table 2) are obtained by

$$\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{\mathbf{y}}'^b_i - \bar{\tilde{\mathbf{y}}}'^b_i\right)^2.$$

Table 2. Agrarian region level estimates of the mean grapevine production with predicted root mean squared error (PRMSE) and coefficient of variation (CV%). (Regions in order of increasing sample size)

Agrarian Region	n_i	Estimate	PRMSE	CV %
04605 – Montagna Litoranea della Versilia	4	1.81	0.75	41.53
05004 – Colline Litoranee del Monte Pisano	4	3.35	0.73	21.65
04601 – Garfagnana Occidentale	5	2.71	1.20	44.46
10001 – Alto Bisenzio	6	3.39	0.85	24.96
04604 – Montagna della Val di Lima Lucchese	10	2.77	0.76	27.46
04808 – Colline di Incisa in Val d’Arno	10	50.18	8.53	17.00
05104 – Colline dell’Alta Valle Tiberina	11	12.23	3.35	27.37
05102 – Alto Tevere	12	7.65	2.14	28.03
05201 – Versante Orientale dell’Amiata	12	7.31	2.30	31.50
05005 – Colline Litoranee del Medio Cecina	14	18.35	2.63	14.33
04802 – Montagna di Vallombrosa	16	44.75	7.08	15.82
04602 – Garfagnana Centrale	18	3.04	1.12	36.90
04603 – Garfagnana Orientale	18	2.71	1.13	41.73
04504 – Colline della Lunigiana Sud-occidentale	20	2.91	1.99	68.43
04801 – Alto Santerno e Alto Lamone	20	5.61	1.49	26.50
04503 – Montagna Litoranea di Massa	21	2.38	1.53	64.40
10002 – Colline di Prato	21	12.73	02.12	16.68
05002 – Colline tra Era e Fine	26	12.04	2.09	17.39
04809 – Pianura di Fucecchio	27	50.33	9.03	17.94
04502 – Montagna della Lunigiana Sud-orientale	29	4.53	2.75	60.60
04804 – Colline del Medio Valdarno	29	61.09	11.62	19.01
05003 – Colline dell’alto Cecina	29	5.64	1.08	19.14
05101 – Casentino	30	17.72	3.11	17.55
04501 – Montagna della Lunigiana Settentrionale	32	4.72	2.99	63.39
05204 – Colline di Siena	33	19.63	4.46	22.72
04805 – Colline di Firenze	34	19.79	5.08	25.69
05203 – Colline del Chianti	34	254.34	63.31	24.89
04803 – Colline del Mugello	40	48.00	7.92	16.49
05301 – Versante Occidentale dell’Amiata	40	5.94	2.46	41.48
04901 – Colline Litoranee di Livorno	45	7.17	2.11	29.45
04806 – Colline della Val d’Elsa Inferiore	48	132.59	19.61	14.79
05105 – Colline della Media Val di Chiana	48	18.65	5.28	28.32
05207 – Colline di Val d’Orcia	50	23.49	8.61	36.65
05202 – Colline dell’Alta Val d’Elsa	51	82.49	23.92	29.00
04807 – Colline del Greve e del Pesa	56	244.39	50.82	20.79
05006 – Pianura di Pisa	59	4.62	1.12	24.31
05307 – Pianura di Grosseto	60	18.90	5.79	30.63
05103 – Colline del Valdarno superiore	62	31.90	7.99	25.06

Agrarian Region	n_i	Estimate	PRMSE	CV %
05206 – Colline dell’Alta Val di Chiana	63	2.25	20.42	28.26
05001 – Colline del Valdarno Inferiore	65	24.28	3.91	16.12
04703 – Colline dell’Ombrone Pistoiese	71	8.63	1.98	22.97
05205 – Colline di Val d’Arbia	77	79.36	17.78	22.40
05302 – Colline dell’Ombrone	77	10.55	3.66	34.69
05304 – Colline Litoranee di Follonica	80	15.35	3.72	24.22
05303 – Colline del Fiora	92	46.70	15.74	33.71
04902 – Colline Litoranee di Piombino	95	20.34	4.41	21.66
04606 – Pianura della Versilia	108	2.45	0.79	32.30
04702 – Colline della Val di Nievole	110	3.14	0.97	30.97
04607 – Pianura di Lucca	112	7.14	2.12	29.68
05106 – Colline di Arezzo	114	22.36	5.58	24.96
05305 – Colline Litoranee dell’Albenga	117	55.22	15.30	27.71
04701 – Montagna di Pistoia	185	3.28	1.00	30.34

As noted by Opsomer *et al*. (2008) a drawback of this parametric bootstrap approach is that it could lead to biased inference if the distributions for the random components are misspecified. Moreover a double-bootstrap procedure able to capture also the variability of the variance components should be a better choice. Research in a more accurate method to estimate the uncertainty of our predictions would certainly be warranted.

4. FINAL REMARKS AND FORTCOMING ISSUE

The interest in spatial data analysis is increased in every area of statistical research. Particular interest is given to the possible ways in which spatially referenced data can support local policy makers. Geographical information is frequently available in many areas of observational sciences, and the use of specific techniques of spatial data analysis can improve our understanding of the studied phenomena. Moreover, it is recurrent, not only in agricultural field but also in many other applications such us environmental and biomedical ones, to encounter variables that have a proportion of values equal to zero and a continuous, often skewed, distribution among the remaining values. The two part model represents the leading suggested in literature for this sort of variables. However, there seems to be no studies which combine jointly small area estimation (SAE), models for overdispersed or zero-inflated data and spatial data.

We have developed a two-part geoadditive model under the framework of small area estimation (SAE) and we demonstrate its practical usefulness by estimating the per farm average grapevine production in Tuscany (Italy) at agrarian region level.

While the two-part model provides the flexibility to model data in accordance with a scientifically plausible data generating mechanism and the results are encouraging, further research is necessary to better analyze the variability and to develop a better method for estimating the mean square error of the mean predictor. Another aspect that we should investigate is the use of a two-part small area geoadditive model with a correlation between the random effects of the two parts of the model. Such a situation leads to a likelihood that does not factor in two separate components, that is the two models cannot be fitted separately. Finally we would like to underline that in literature the application of the two-part model mainly concern biomedical data, however our results show that this kind of model could be usefully employed in other application fields.

ACKNOWLEDGEMENTS

This work is supported by the project PRIN2007 ‘Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: Methodological aspects and applications for producing official statistics’(PRIN 2007RHFBB3_004) awarded by the Italian Ministry for Education, University and Research to the universities of Perugia, Cassino, Firenze, Pisa and Trieste.

REFERENCES

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

Berk, K.N. and Lachenbruch, P.A. (2002). Repeated measures with zeros. *Stat. Methods Med.* Res., **11**, 303-316.

Bocci, C., Petrucci, A. and Rocco, E. (2006). Geographically weighted regression for small area estimation: An agricultural case study. *Proceedings of the XLIII Scientific Meeting of Italian Statistical Society*, 615-618.

Chandra, H. and Chambers, R. (2006). Small area estimation with skewed data. *Working Paper M06/05, Southampton Statistical Sciences Research Institute*, University of Southampton.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, 93-105.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Gosh, P. and Albert, P.S. (2009). A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Comput. Statist. Data Anal.*, **53**, 699-706.

Hall, D.B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, **56**, 1030-1039,

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *J. Appl. Stat.*, **52**, 1-18.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Olsen, M.K. and Schafer, J.L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.*, **96**, 730-745.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *J. Roy. Statist. Soc.*, **B70**, 265-286.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Agric. Biol. Environ. Stat.*, **11**, 169-182.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Internat. Statist. Rev.*, **70**, 125-143.

Pratesi, M. and Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statist. Method. Appl.*, **17**, 113-141.

Ridout, M., Hinde, J. and Demétrio, G.G.B. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219-223.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Singh, B., Shukla, G., Kundu, D. (2005) Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183-195.

Zhang, M., Strawederman, R.L., Cowen, M.E. and Wells, M.E. (2006). Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *J. Amer. Statist. Assoc.*, **101**, 934-945.