Dipartimento di Statistica
"Giuseppe Parenti"

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

WORKING PAPER 2012/05

Small area estimation for semicontinuos skewed georeferenced data

Emanuela Dreassi,
Alessandra Petrucci, Emilia Rocco

Università degli Studi
di Firenze

Small area estimation for semicontinuos skewed georeferenced data

Emanuela Dreassi, Alessandra Petrucci, Emilia Rocco

Dept. of Statistics "G. Parenti" University of Firenze
Viale Morgagni, 59 - I 50134 Firenze Italy

Abstract: Semicontinuous random variables combine continuous distributions with point masses at one or more locations. A particular type of semicontinuous variable has been considered in this paper: a mixture of zeros and highly skewed continuously distributed positive values. This kind of variable occurs in economic surveys of individuals or establishments (e.g. specific types of income or expenditures), as well as in agricultural, epidemiological and environmental surveys. Frequently, this type of variable describes phenomena that have a spatial distribution and reliable small area estimates of their means or totals could be required. As in other small area estimation (SAE) problems, the small sample sizes (in at least some of the sampled areas) and/or the existence of non sampled areas need to use model based estimation methods. However, commonly used small area estimation methods, which assume that a linear mixed model can be used to characterize the regression relationship between the response variable and at least one auxiliary variable, are not suitable for this kind of data.

In this paper we propose the use of a two-part random effects SAE model that accounts for excess of zero but also for skewness of the distribution of non zero responses. This is carried out by specifying, in the first part, a logistic regression model for the probability of a non zero occurrence and, in the second part, a gamma regression model for the mean of the non zero values. The model includes a correlation structure on the area random effects that appears in the two parts and incorporates a bivariate smooth function of the geographical coordinates of the units in order to consider the spatial structure in the data within each small area. A hierarchical Bayesian approach is suggested to fit the model, produce the small area estimates of interest, and evaluate their precision. An application to real agricultural survey data from the Italian Statistical Institute demonstrates the satisfactory performance of the method.

Keywords: Geostatistical models; Hierarchical Bayesian models; Linear mixed model; Penalized splines; Two-part random effects models.

1 Introduction

Direct survey estimators for small areas are usually unreliable due to the unduly small size of the sample in the areas. Hence it becomes necessary to use models, either explicit or implicit, to connect the small areas and obtain estimators of improved precision by 'borrowing strength'. The most popular class of models for small area estimation (SAE) are linear mixed models which include independent random area effects to account for the variability between the areas exceeding that explained by auxiliary variables. The response variable can either be observed at the small area level or at a smaller unit or respondent level. Fay and Herriot (1979) studied the area level model and proposed an empirical Bayes estimator for this case. Battese *et al.* (1988) considered the unit level model and constructed an empirical best linear unbiased predictor (EBLUP) for the small area means. Numerous extensions to this set-up have been considered in literature, including cases in which data follow various generalized linear models, or have more complicated random-effects structures. Jiang and Lahiri (2006) provide a general review.

The extension that we propose in this paper is a two-part SAE model at a unit level for variables that describe spatial phenomena, have a portion of values equal to zero and a continuous, skewed distribution, with the variance increasing with the mean among the remaining values. In many fields of applied research, including agricultural, environmental and epidemiological framework, researchers encounter data with these characteristics. The small area estimation approach suggested in order to deal with them arises from putting together several different methodologies developed separately within the framework of small area estimation methods or different contexts.

In literature, the 'excess' zeros in data are usually described by non standard two component mixture models that mix a degenerate distribution with point mass of one at zero and a standard distribution. This is carried out considering a pair of regression models: a model, usually logit or probit, for the mixing proportion and a conditional regression model (linear or Poisson, or others depending on the nature of data) for the mean response given that it is non zero. These models were originally developed to analyze count data and in this context are referred to as zero inflated (ZI) models. Examples include regression models for zero inflation relating to a Poisson (ZIP models), zero inflated negative binomial (ZINB) and zero inflated binomial (ZIB). Lambert (1992), Hall (2000), Ridout *et al.* (2001) among others, have studied these models extensively.

Naturally large numbers of zeros sometimes occur in continuous data as well, but continuous distributions have a null probability of yielding at zero and therefore there is little motivation to model them as a ZI model since it is possible to tell from which distribution in the mixture each response comes simply from its value. Unfortunately, this simplification may not occur for clustered data, and therefore in the presence of cluster correlation the mixture models become interesting. ZI models for clustered semicontinuous data in literature are referred to as *two-part* models and have been developed mainly for analyzing longitudinal response variables in biomedical application (Olsen and

Shafer, 2001; Berk and Lachenbruch, 2002; Tooze *et al.*, 2002; Albert and Shen, 2005; Gosh and Albert, 2009). Moreover, they usually include a cluster specific random effect in both the logit or probit model used for the probability of non zero response and in the conditional regression model used for the mean response given that is non zero, which is commonly assumed to be linear in the log-transformed scale. Even though the lognormal distribution is a popular model in biostatistics and other fields of statistics, Bayesian inference on the mean of the distribution is problematic due to the fact that for many popular choices of the prior for variance (on the log-scale) parameter, the posterior distribution has no finite moments, leading to infinite expected loss on Bayes estimators for the most common choices of the loss function (Fabrizi and Trevisano, 2012). The Gamma distribution could be a valid alternative to the lognormal choice for the non zero response distribution (Grunwald and Jones, 2000 and Hyndman and Grunwald, 2000). In this paper we adopt a Gamma distribution to model the non-zero values of a variable for which we are interested in producing small area estimates. The problem of zero inflated data for SAE has already been discussed in literature. Pfeffermann *et al.* (2008) and Chandra and Sud (2012) dealt with it under a two-part random effects model using a Bayesian and a frequentist approach respectively, but both considered a 'non skewed' distribution for the non zero responses, adopting a Normal distribution to model their means. Here we consider a variable for which the non-zero values have a skewed distribution and in order to deal with this we suggest a two-part random effects model consisting of a logit random effects model for the probability of non-zero values, and a conditional gamma random effects model for the mean response given that is non zero. The model includes a correlation structure on the area random effects that appears in two parts and incorporates a bivariate smooth function of geographical coordinates of the units in order to consider the spatial structure of the data within each small area.

The inclusion of the spatial proximity effect in a random effects model is possible by using penalized splines to represent the smoothing trend. The possibility of incorporating the spatial proximity effects and more generally, the non linear covariate effects into the small area estimation by using penalized splines has already been investigated in literature by Opsomer *et al.* (2008), who in turn exploited the close connection between penalized splines and the linear mixed model shown by Wand (2003). Here we extend this possibility within the framework of SAE two-part random effects models. It must also be noted that, in a different perspective, the possibility of analyzing the spatial distribution of a study variable while accounting for possible linear or non-linear covariate effects was also suggested by Kammann and Wand (2003). They represent these effects by merging an additive model (Hastie and Tibshirani, 1990) - that accounts for the non-linear relationship between the variables - and a kriging model - that accounts for the spatial correlation - and by expressing both as a linear mixed model, which, due to its specific generation process (a fusion of a geostatistical and an additive model) is called geoaddive. The mixed model structure provides a unified and modular framework that allows for easily extending the model to include various kinds of generalization and evolution, and for our purposes in particular, to include the specific cluster/area random effects. Considering this perspective, our proposed model can also be referred to as a geospatial or geoadditive two-part SAE model.

The study is motivated by a real small area estimation problem. We are interested in producing estimates of the per-farm average grapevine production in Tuscany (Italian region) at a subregional level using the Farm Structure Survey data collected by the Italian Statistical Institute. The survey is structured so as to provide reliable estimates at the level of the Administrative Regions - NUTS2 level. Therefore, to produce estimates at a subregional level it is necessary to employ indirect estimators. We intend to produce estimates at an Agrarian Region level (aggregations of municipalities with uniform natural and agricultural characteristics) and moreover, our response variable, the per-farm grapevine production, shows a point mass at zero, a highly skewed distribution of the non zero values and a spatial trend. All these aspects are considered in our proposed model. It is consequently fitted to data and estimates of the per-farm average grapevine production at Agrarian Region level and the corresponding credibility intervals are obtained. The results demonstrate a satisfactory performance of the suggested small area estimation approach.

The paper is organized as follows. Our modelling small area approach is described in Section 2. Section 3 describes the motivating application, followed by the discussion in Section 4.

2 The model

2.1 Basic Setup, Definitions and Assumptions

A small area estimation problem is usually formulated as follows. A finite population U of N units partitioned in m subsets (areas) of size N_i, such that $\sum_{i=1}^{m} N_i = N$ is considered. A sample r of n units is selected from U according to a non-informative sampling design $p(r)$. r may be decomposed as $r = \bigcup_{i=1}^{m} r_i$ where r_i is the area specific sample of size n_i. A response variable y is observed for each unit in the sample; y_{ij} denotes the value of the response variable for the unit $j = 1, \ldots, N_i$ in small area $i = 1, \ldots, m$. Of primary interest is the estimation of the area means $\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ (or area totals $Y_i = \sum_{j=1}^{N_i} y_{ij}$). These means may be decomposed as $\overline{Y}_i = N_i^{-1}(\sum_{j \in r_i} y_{ij} + \sum_{j \in q_i} y_{ij})$ where q_i is the complement of the area specific sample r_i to the area population (thus of size $N_i - n_i$). The sample area sizes n_i are too small to calculate reliable direct estimates. The values of some covariates are available at area and/or unit level for $j \in r_i$ and also in $j \in q_i$ therefore, generally speaking, indirect estimation could be considered. Here we assume that for each unit j in small area i two vectors $\mathbf{x}_{ij}$ and $\mathbf{x}^*_{ij}$ of covariates and the spatial location $\mathbf{s}_{ij}$ ($\mathbf{s} \in R^2$) of the unit are known ($\mathbf{x}_{ij}$ and $\mathbf{x}^*_{ij}$ may coincide, or be partial or completely different). We further assume that the response variable Y is a semicontinuous skewed variable.

2.2 Hierarchical Bayesian non standard mixture model for SAE with semicontinuous skewed and georeferenced data

Due to its semicontinuous nature we decompose the response variable into two variables:

$$I_{ij} = \begin{cases} 1 & \text{if } y_{ij} > 0 \\ 0 & \text{if } y_{ij} = 0 \end{cases} \quad \text{and} \quad y'_{ij} = \begin{cases} y_{ij} & \text{if } y_{ij} > 0 \\ \text{irrelevant} & \text{if } y_{ij} = 0 \end{cases}$$

for which we assume a two-part superpopulation model. The two-part model is specified conditionally on the covariates $(\mathbf{x}_{ij}, \mathbf{x}^*_{ij})$, the geographical coordinates $(\mathbf{s}_{ij})$ and two sets of random area effects $\{u_1, \ldots, u_m\}$ and $\{u^*_1, \ldots, u^*_m\}$. Specifically, for unit j in area i, let $\boldsymbol{\theta}_{ij} = (\mathbf{x}_{ij}, \mathbf{x}^*_{ij}, \mathbf{s}_{ij}, u_i, u^*_i)$ and $\pi_{ij} = P(I_{ij} = 1|\boldsymbol{\theta}_{ij})$, the two 'parts' of the model are the following.

Part one, the mixing proportions π_{ij} are modelled as

$$\eta_{ij} = \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_{0x} + \mathbf{x}^T_{ij}\boldsymbol{\beta}_x + h(\mathbf{s}_{ij}) + u_i \qquad (1)$$

where $h(\cdot)$ is an unspecified bivariate smooth function depending on geographical unit coordinates $\mathbf{s}_{ij}$ and $\{u_i : \forall i = 1, \ldots, m\}$ is a set of area specific random effects.
By representing $h(\cdot)$ with a low rank thin plate spline (Ruppert *et al.*, 2003) with K knots $(\kappa_1, \ldots, \kappa_K)$ that is

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

the model (1) can be written as a mixed model (Kammann and Wand, 2003; Opsomer *et al.*, 2008):

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

where:

- $\mathbf{X} = [1, \mathbf{x}^T_{ij}, \mathbf{s}^T_{ij}]$ is the matrix of covariates pertaining to the fixed effects referring to the N population units;
- $\mathbf{Z} = [C(\mathbf{s}_i - \kappa_K)]_{1\leq i \leq N, 1 \leq k \leq K} [C(\kappa_h - \kappa_K)]^{-1/2}_{1 \leq h,k \leq K}$ with $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log \|\mathbf{r}\|$ is the matrix of the thin plate spline basis functions;
- $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_N]^T$ with $\mathbf{d}_i = [d_{i1}, \ldots, d_{im}]^T$ and d_{ij} an indicator taking value 1 if observation i is in small area j and otherwise 0 is the matrix that model the structure of the area random effects;
- $\boldsymbol{\beta} = (\beta_{0x} + \beta_{0s}, \boldsymbol{\beta}^T_x, \boldsymbol{\beta}^T_s)^T$ is a vector of unknown coefficients;
- $\mathbf{u}$ is the vector of the m area specific random effects;

- $\boldsymbol{\gamma}$ is the vector of the K thin plate spline coefficients (seen as random effects).

Part two, the probability distribution of $(Y_{ij} \mid I_{ij} = 1, \boldsymbol{\theta}_{ij})$ is a Gamma distribution with shape parameter ν and scale parameter δ_{ij}, therefore with a constant coefficient of variation over units $1/\sqrt{\nu}$ (see McCullagh and Nelder, 1989, pages 30 and 49). Considering mean parameterized Gamma distribution, $\mathrm{E}(Y_{ij} \mid I_{ij} = 1, \boldsymbol{\theta}_{ij}) = \mu_{ij} = \delta_{ij}\nu$ and $\mathrm{Var}(Y_{ij} \mid I_{ij} = 1, \boldsymbol{\theta}_{ij}) = V(\mu_{ij}) = \mu_{ij}^2/\nu$, the means μ_{ij} are modelled through a log-link function as

$$\log \mu_{ij} = \beta_{0x}^* + \mathbf{x}_{ij}^{T*} \boldsymbol{\beta}_x^* + h^*(\mathbf{s}_{ij}) + u_i^* \tag{3}$$

where $h^*(\cdot)$ is an unspecified bivariate smooth function depending on geographical unit coordinates $\mathbf{s}_{ij}$ and $\{u_i^* : \forall\ i = 1, \ldots, m\}$ a set of area specific random effects.
By representing $h^*(\cdot)$ with a low rank thin plate spline with K knots, as we did for $h(\cdot)$, the model (3) became

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

where all terms $(\mathbf{X}^*, \boldsymbol{\beta}^*, \mathbf{Z}^*, \boldsymbol{\gamma}^*, \mathbf{D}^*, \mathbf{u}^*)$ have the same meaning as those indicated by the same symbol without an asterisk in model (2). It must be noted that even if the same covariates are used in both parts, $\mathbf{X}^* \neq \mathbf{X}$, $\mathbf{Z}^* \neq \mathbf{Z}$ and $\mathbf{D}^* \neq \mathbf{D}$ as the (4) only concern the population units with positive responses.

Two different assumptions could be adopted on the relationship between the two parts of the model. First, both the random area effects and the spline random effects from the two parts are assumed to be jointly normal and possibly correlated,

$$(u_i, u_i^*)^T \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma_u^2 & \sigma_{uu^*} \\ \sigma_{uu^*} & \sigma_{u^*}^2 \end{bmatrix}\right)$$

and

$$(\gamma_k, \gamma_k^*)^T \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma = \begin{bmatrix} \sigma_\gamma^2 & \sigma_{\gamma\gamma^*} \\ \sigma_{\gamma\gamma^*} & \sigma_{\gamma^*}^2 \end{bmatrix}\right).$$

Second, the two parts of the model are assumed to be independent, that is: $u_i \sim N(0, \sigma_u^2)$, $u_i^* \sim N(0, \sigma_{u^*}^2)$, $\gamma_k \sim N(0, \sigma_\gamma^2)$ and $\gamma_k^* \sim N(0, \sigma_{\gamma^*}^2)$. This last assumption corresponds to estimate separately the two models (hereafter separate two-part model). Otherwise, the first assumption and any other in which at least one of the two random components are assumed correlated, define a full two-part model.

Within a Bayesian framework we assume independent, noninformative priors for the parameters of the whole model given by (2) and (4). More specifically, we assume a noninformative normal distribution for each element of $\boldsymbol{\beta}$ and for each element of $\boldsymbol{\beta}^*$. Noninformative half-Cauchy distributions (as suggested by Gelman, 2006) are assumed as priors for ν, the shape parameters of Gamma distribution (the squared reciprocal of the standard deviation-like coefficient of variation). Under the assumption of correlated random effects, the variance-covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\gamma$ are assumed to follow a noninformative inverse

Wishart. On the contrary, under the assumption of independence, a half-Cauchy distribution is given for each variance parameter.

It must also be noted that the model specified through the expressions (2) and (4) may easily be extended to include other random effects (due, for example, to other non-linear covariate effects or to a clustering process inside the areas). Moreover, even if the area random effects and the spatial proximity random effects are present in both the expressions (2) and (4), there may be situations in which a random effect is relevant for one part of the model but not for the other part and therefore it is only included in one part.

After obtaining estimates for all the parameters via the MCMC sampling procedure (using data on the sample: $j \in r_i$), we extended to $j \in q_i$ all posterior distribution of the parameters and obtained the posterior distribution for y_{ij} when $j \in q_i$ and finally the posterior distribution for the means of each small area $\overline{Y}_i$ with their credibility interval.

$$\widehat{\overline{Y_i}} = N_i^{-1} \left(\sum_{j \in r_i} y_{ij} + \sum_{j \in q_i} \hat{y}_{ij} \right)$$

where the predicted values $\hat{y}_{ij}$ are obtained as: $\hat{y}_{ij} = \hat{\pi}_{ij} \tilde{y}_{ij}$ with $\hat{\pi}_{ij} = \exp(\widehat{\eta_{ij}})/(1+ \exp(\widehat{\eta_{ij}})$ and $\tilde{y}_{ij} = \exp\left(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}} + \hat{u}_i \right)$.

3 A real example on SAE: Tuscany's grapevine production estimation

A ten-yearly Agricultural Census and a two-yearly Farm Structure Survey are routinely conducted by the Italian Statistical Institute (ISTAT). The unit of observation for both the census and the survey is the farm, and surface areas (measured in hectares) allocated to different crops as well as many other socioeconomic variables registered for each unit. In the Fifth Agricultural Census conducted in 2000 (see census2000 below) spatial information was collected for the first time. This consists of the universal transverse Mercator (UTM) geographical coordinates of each farm's administrative centre. Moreover, in the Farm Structure Survey, up until 2005, the productions of each crop (quantity in quintals) were observed.

We are interested in the estimation of the per-farm average grapevine production for the 52 Agrarian Regions the Tuscany is divided into, with reference to 2003 for which the Farm Structure Survey (see FSS2003 below) data are available. The FSS2003 survey is designed to obtain estimates at just a regional level, so when the interest is on an agrarian region we have to consider indirect estimators that 'borrow strength' from related areas using as auxiliary variables (known for all the units in the population) the spatial information and other variables registered at census2000 time. In fact, due to the high correlation values observed over sampled data between the explicative variables in years 2000 and 2003 (about 90% for the grapevines surface), we can assume that the time lag between the response and the explicative variables would have a negligible effect.

Classical small area estimations are not feasible in this study because a large number (1,489 compared to a total of 2,450) of farms in the FSS2003 sample do no't produce grapevines, and a few (961 units) produce the majority of the total production in the region. See Figure 1 for the spatial pattern of zero grapevine production farms and Figure 2 for the elevated skewness of positive grapevine production distribution for the whole region (on the left) and for each agrarian region (on the right). Figure 3 shows the grapevine production coefficients of variation at an agrarian regional level evaluated by only considering the positive values. On observing Figure 2 and Figure 3, the Gamma distribution parameterized with a constant coefficient of variation McCullagh and Nelder (1989) seems to be a valid model.

The peculiarity of the agrarian region, as an aggregation of homogeneous municipalities with respect to natural and agricultural characteristics, makes the random area effects sufficient for explaining the spatial heterogeneity in the probability of zero occurrence. The choice of not considering the geospatial term in the logit part has also been motivated by DIC criteria (Spiegelhalter *et al.*, 2002): comparing different models with various combinations of random effects. While, with regard to the quantity of grapevine produced with the same allocated surface, some spatial heterogeneity persists inside each area (see Figure 4). These considerations, confirmed by an explorative data analysis, induce us to adopt a two-part SAE model in which the random effects due to the smooth function of spatial coordinates is only enclosed in the second part of the model. For this part of the model, the $K = 50$ spline knots are set by means of the *clara* space-filling algorithm of Kaufman and Rousseeuw (2003); using the `clara` function on library `Cluster` on `R package`.

The selection of the covariates to be included in each of the two parts of the model, among several socioeconomic variables available at census2000, was first performed using explorative analysis. For the logit model, two auxiliary variables are considered: the surface allocated to grapevines in logarithmic scale and a dummy variable that indicates the selling of grapevine related products, both at census2000. In the log-linear model for gamma we include the same two variables plus the number of days worked by farm family members in 2000.

We allow for non zero correlation between the random area effects in the two parts, as we expected that a farm located in an area with an elevated per-farm average of grapevine production would have a higher probability of a non zero grapevine production. Figure 5 provides evidence of this possible covariance structure. However, the magnitude of this correlation and the importance of accounting for it when fitting the model, depends on the prediction power of the covariates available for the two parts of the model. Therefore, as our covariates and above all the surface allocated to grapevine, have a strong prediction power (the correlation coefficient among the surfaces allocated to grapevine and the grapevine production, evaluated using the FSS2003 sample data is greater than 85%, when the zero values are included and also when they are excluded) we expected a lower value for the estimated covariance parameter. This same consideration has also led us to consider the ‘separate’ two-part model.

Estimation of both the models (full and separate two-part models) was performed using MCMC simulation methods. We used the library `BRugs` (an interface for `OpenBugs`) on `R package`. Convergence was stated using Gelman and Rubin (1992) convergence diagnostic criteria. The algorithm seems to converge after a few thousand iterations. However, also given the very high number of (non monitored) parameters in the model, we decided to discard the first 200,000 iterations (burn-in) and to store 2,000 samples (one each 100) of the following 200,000 iterations for estimation. All sampled MCMC chains are used on the two-part model to predict grapevine production distribution for each farm not present in the sample FFS2003 but for which we have all the information: covariates and geographical locations at census2000. Therefore, estimates related to 136,817 non sampled on FFS2003 farms are combined with values observed for 2,450 farms on the sample to obtain the grapevine mean production predictions at an agrarian region level. Results on grapevine means production estimates at an agrarian region level and their credibility interval (95%) from the simulated posterior distribution are illustrated in Figure 6 for both models. Coefficient estimates are reported in Table 1. We observed that covariance σ_{uu^*} was not significantly different from zero. As stated above, this could be due to the strong prediction power of the covariates. From DIC values, we observed that the models are comparable: the full two-part model has a higher complexity but a greater goodness of fit than the separate two-part model.

The map of the estimated agrarian region means for both model is shown in Figure 7(a) (the maps for both models are coincident). The map presents an evident geographical pattern, with the higher values in the areas belonging to the Chianti area and the lower values in the northern mountainous area of the provinces of Massa Carrara and Lucca, confirming the pattern of the experts' estimate means produced by ISTAT. These experts' estimates are obtained via determination of a crop specific coefficient of soil productivity and are released at a provincial level. To better compare them with our results, we calculate the agrarian region level experts' estimate by multiplying the agrarian region grapevine mean surfaces at year 2000 by the coefficient of soil productivity at a regional level. This experts' estimates map is illustrated in Figure 7(b). The comparison between the two maps confirms that our estimates are very close to those of the experts.

4 Conclusions

This paper presents a Hierarchical Bayesian small area estimation approach for estimating the small area means of a variable that is semicontinuous, highly skewed and with a relevant spatial structure. Even if literature relating to each of these aspects of the data is plentiful, to the best of our knowledge, the problem as a whole has never been considered before.

The application to real survey data shows on one hand the concrete presence of this sort of problem in real situations, and on the other, the usefulness of the

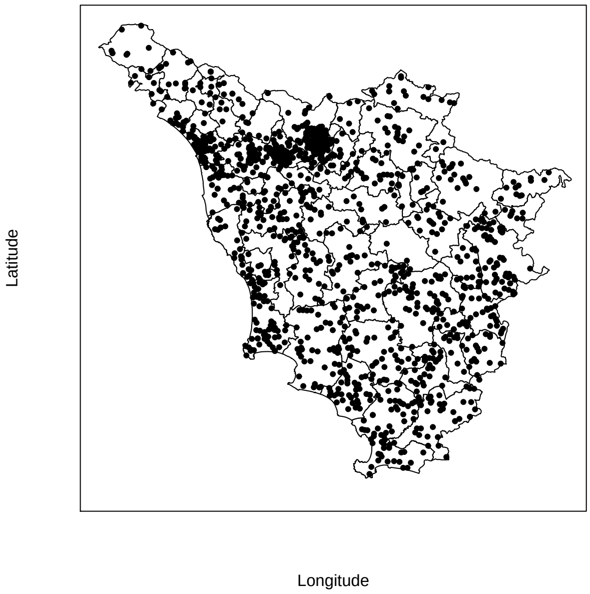
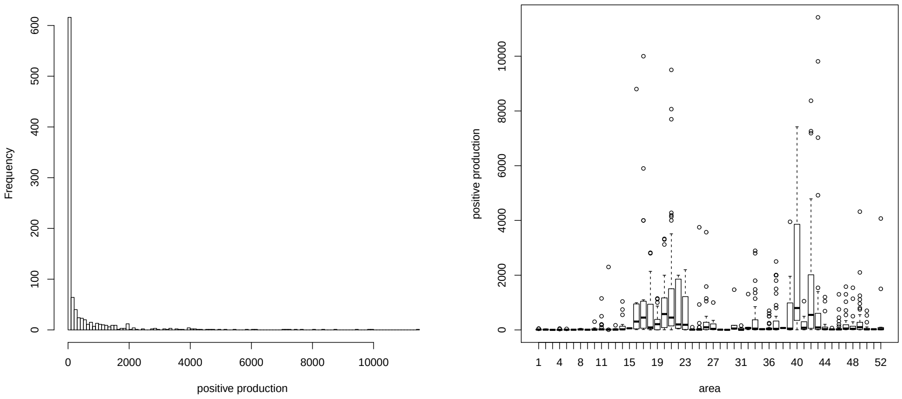


Figure 1: Zero grapevine production' farms location



(a)

(b)

Figure 2: Positive grapevine production distribution: (a) the histogram for the whole region (b) the box-plot for each agrarian region

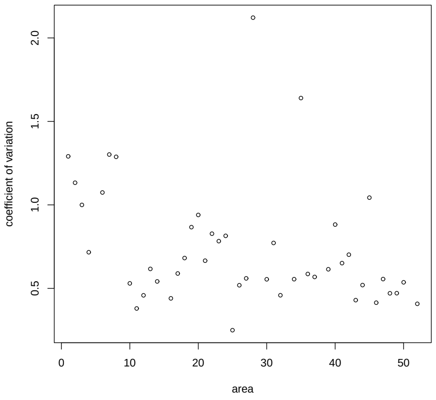


Figure 3: Coefficient of variation for grapevine positive production for each agrarian region

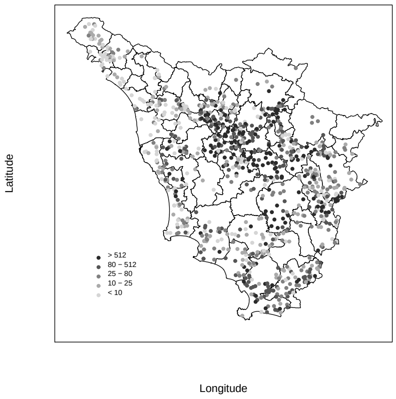


Figure 4: Positive grapevine production and his spatial pattern

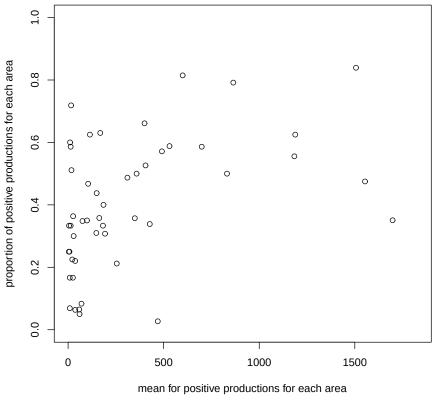
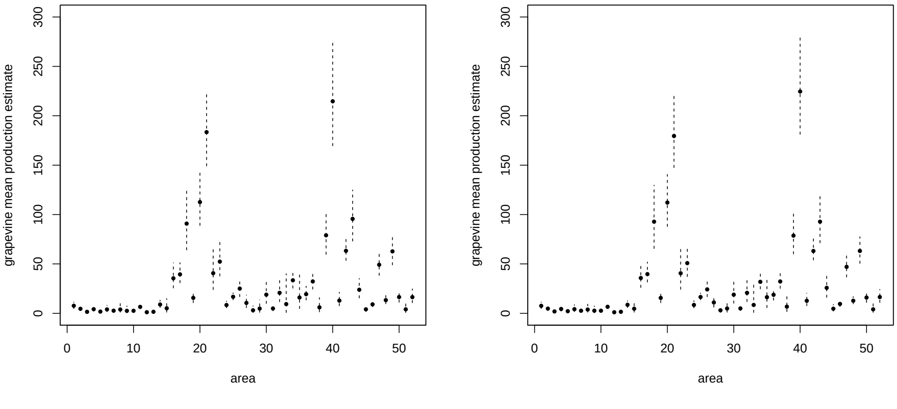


Figure 5: Proportion of positive grapevine production by average of positive grapevine production for each agrarian regions



(a) (b)

Figure 6: Estimated grapevine mean productions and their 95% credibility interval from: (a) full two-part model (b) separate two-part model

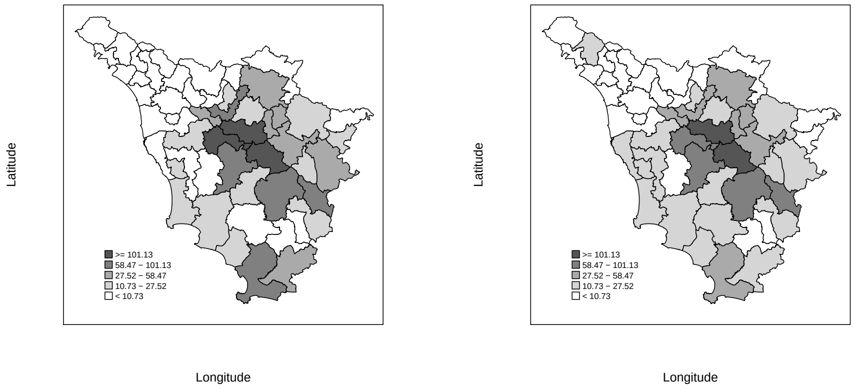
Table 1: Results from full two-part model and separate two-part model: coefficient estimates with their 95% credibility interval and DIC criteria values

		full two-part model		separate two-part model	
	parameter	estimates	IC95%	estimates	IC%95
first	constant	-1.367	-1.644 - -1.082	-1.369	-1.657 - -1.094
	log surface allocated to grapevines	1.908	1.558 - 2.286	1.907	1.559 - 2.271
	selling of grapevines products	1.098	0.743 - 1.488	1.103	0.758 - 1.472
	σ_u	0.868	0.661 - 1.121	0.882	0.680 - 1.139
second	constant	0.369	0.291 - 0.457	-0.234	-0.258 - -0.201
	x coordinate	0.281	0.254 - 0.308	0.044	**-0.0001 - 0.102**
	y coordinate	-0.021	-0.039 - -0.005	0.050	0.040 - 0.060
	log surface allocated to grapevines	1.195	1.139 - 1.253	1.114	1.070 - 1.169
	selling of grapevines products	0.314	0.242 - 0.410	0.580	0.525 - 0.621
	number of days worked	0.0004	0.0002 - 0.0006	0.0004	0.0002 - 0.0006
	σ_u^*	0.397	0.288 - 0.533	0.382	0.255 - 0.535
	σ_γ^*	1.865	1.319 - 2.568	1.056	0.604 - 1.644
	σ_{uu^*}	-0.0005	**-0.170 - 0.181**		
Dbar		12750		12760	
Dhat		12650		12660	
DIC		12850		12850	
pD		98.20		92.45	

suggested approach.

It is well established in literature that by ignoring the accumulation of zeros in fitting a model, the model assumptions are rendered invalid and therefore problems with inference are liable to occur. More specifically, highly biased predictors and wrong coverage rate of credibility intervals may be obtained. Clearly the relevance of these problems depends on the percentage of zeros. In the application considered in this work, where the percentage of zero values for the response variable exceeds 60%, estimates using a linear mixed model were unacceptable.

Another inefficient approach to analyzing zero inflation in data consists of



(a) (b)

Figure 7: Estimates grapevine mean productions at an agrarian region level: (a) using suggested models (b) experts' estimates

only considering the data greater than zero. If only the data greater than zero are used in the analysis, important information about units with zero response is lost, and estimates of means/totals will not include zero values. When relying on estimates to make policy decisions, inaccurate conclusions may be arrived at, thus leading to policies that are inadequate or inappropriate for the population of interest. In addition, this method does not account for the relationship that may exist between the probability of a non zero response and the level of the non zero response.

Fitting the full two-part model, is the only way to take into account all the population units and possible relationship that may exist between the probability of a non zero response and the level of the non zero response.

Clearly, in the choice of the model it is necessary to consider not only the accumulation of zeros but also the distribution of the non zero values, and if it is highly skewed a linear random effects model to model the mean of the positive responses may be inopportune. This justifies our choice of the gamma model in the second part, the effectiveness of which is confirmed by the results. To our knowledge, within the context of SAE, the skewness of data is usually treated using the lognormal distribution. The aim of this paper is also to stress how the Gamma distribution with its high flexibility could be a valid alternative (see for example Firth, 1988).

The last message of the paper is that when small areas of study are geographical areas, and the study variable shows a spatial trend, an adequate use of geographic information and geographical modelling is able to provide more accurate estimates for small area parameters.

Even if the suggested approach provides the flexibility to model the data in accordance with a scientifically plausible data generating mechanism and the results are encouraging, further research is still necessary. An accurate evaluation of the conditions that make the full two-part model actually preferable to the separate estimation of the two components represents a future topic of research. Moreover, in this paper we adopted a Bayesian approach, however we intend to investigate the possibility of developing a similar SAE method with a frequentist perspective in the future.

References

Albert, P.S., and Shen, J. (2005). Modelling longitudinal semicontinuous emesis volume data with serial correlation in an acupuncture clinical trial. *Journal of the Royal Statistical Society: Series C* **54**,707–720.

Battese, G.E., and Harter, R.M., and Fuller, W.A. (1988). An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association* **83**, 28–36.

Berk, K.N., and Lachenbruch, P.A. (2002). Repeated measures with zeros. *Statistical Methods in Medical Research* **11**, 303–316.

Chandra, H., and Sud, U.C. (2012). Small Area Estimation for Zero-Inflated Data. *Communications in Statistics - Simulation and Computation* **41**, 632–643.

Fabrizi, E., and Trevisano, C. (2012). Bayesian estimation of log-normal means with finite quadratic expected loss, *Bayesian Analysis* **7**, to appear.

Fay, R.E., and Herriot, R.A. (1979). Estimation of Income from Small Places: An Application of James-Stein procedures o Census Data. *Journal of the American Statistical Association* **74**, 269–277.

Firth, D. (1988). Multiplicative errors: Log-normal or Gamma? *Journal of the Royal Statistical Society, Series B* **50**, 266–268.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

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

Gosh, P., and Albert, P.S. (2009) A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Computational Statistics and Data Analysis* **53**, 699–706.

Grunwald, G.K., and Jones, R.H. (2000). Markov models for time series with mixed distribution. *Environmetrics* **11**, 327–339.

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

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hyndman, R.J., and Grunwald, G.K. (2000). Generalized additive modeling of mixed distribution Markov models with application to Melbourne's rainfall. *Australian and New Zealand Journal of Statistics* **42**,145–158.

Jiang, J., and Lahiri, P. (2006). Estimation of Finite Population Domain Means - A Model Assisted Empirical Best Prediction Approach. *Journal of the American Statistical Association* **101**, 301–311.

Kammann, E.E., and Wand, M.P. (2003). Geoadditive Models. *Applied Statistics* **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.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, second edition, Chapman and Hall, London New York.

Olsen, M.K., and Schafer, J.L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.

Opsomer, J.D., and Claeskens, G., and Ranalli, M.G., and Kauermann, G., and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B* **70**, 265–286.

Pfeffermann, D., and Terryn, B., and Moura, F.A.S. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology* **34**, 235–249.

Ridout, M., and Hinde, J., and Demetrio, C.G.B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternative. *Biometrics* **57**, 219–223.

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

Spiegelhalter, D.J., and Best, N.G., and Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 583–639.

Tooze, J.A., and Grunwald, G.K., and Jones, R.H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical research* **11**, 341–355.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223-249.

Copyright © 2012
Emanuela Dreassi,
Alessandra Petrucci, Emila Rocco