SPATIAL MODELS IN SMALL AREA ESTIMATION IN THE CONTEXT OF OFFICIAL STATISTICS

Alessandra Petrucci[1]
Department of Statistics, Informatics, Applications, University of Florence, Italy

Monica Pratesi
Department of Economics and Management, University of Pisa, Italy

Received: 14 September 2013 / Accepted: 6 March 2014

Abstract. *Small area estimation (SAE) plays an important role in survey sampling due to growing demands for reliable small area statistics from both public and private sectors. This paper reviews some of the current techniques of small area estimation combined with spatial models available in the literature. Illustrative examples or applications are likewise presented in the context of official statistics where data sources, particularly surveys and censuses or surveys and administrative sources, have been combined using statistical models based on small area and domain estimation.*

Keywords: *Small area estimation, Spatial modelling, Spatial correlation.*

1. INTRODUCTION

Over the last decade there has been growing demand from both public and private sectors for producing estimates of population characteristics at disaggregated geographical levels, often referred to as small areas or small domains (Rao, 2003a).

Small area estimation (SAE) models are applied in many area of statistical research: environmental statistics, economics, demography, epidemiology, and so on. Every study shows that to use spatially referred data produces estimates more reliable than that obtained by traditional methods. This paper reviews some of the current techniques of small area estimation combined with spatial models available in the literature. The first studies that connect spatial relations 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 parameter, both increasing efficiency and

[1] Alessandra Petrucci, email: alessandra.petrucci@unifi.it

diminishing bias. We refer, among others, to Saei and Chambers (2005), Petrucci et al. (2005), Petrucci and Salvati (2006), Singh et al. (2005) and Pratesi and Salvati (2008). Area level models and unit level models are described with illustrative examples or applications in the context of official statistics. This limits the extension of our review to the models that had - as far as we know - a valuable application in the context of official statistics when producing estimates at subregional level.

The use of spatial information is likely to be most productive when the available model covariates are weak. In this case, spatial information can substantially strengthen prediction for non-sampled areas - provided there is significant spatial correlation. In particular, here we review the linear mixed models that include dependent random area effects to account for between area variation beyond that explained by auxiliary variables (see Section 2) and the geo-additive linear models that includes a non-linear spatial trend in the mean structure of the model (Section 3). These ideas extend to unit level random effect and M-quantile models, see Opsomer et al. (2008) and Salvati et al. (2012). Particularly here we illustrate an application of M-quantile approach to SAE when the non stationarity in the data is captured via geographically weighted regression (Section 4). Our final remarks and the lessons learned from the applications of the models are described in Section 5.

2. AREA LEVEL LINEAR MIXED MODELS

The most popular class of models for small area estimation 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)). Following mixed models methodology (Jiang and Lahiri, 2006), a best linear unbiased predictor (BLUP) is used to obtain the small area parameter of interest (usually the mean or total of the study variable). If, as usual, the variance components are unknown, the correspondent empirical best linear unbiased predictor (EBLUP) is used instead (see Rao (2003b), Chapters 6-7) for a detailed description). Under the classic SAE model we make the assumption of independence of the area-specific random effects. If the small domain of study are geographical areas, this assumption means that we don't take into account any possible spatial structure of the data.

Remembering again the first law of geography however, it is reasonable to

suppose that close areas are more likely to have similar values of the target parameter than areas which are far from each other, and that an adequate use of geographic information and geographical modeling can help in producing more accurate estimates for small area parameters Petrucci et al. (2005)). In addition, Pratesi and Salvati (2008) noted that geographical small area boundaries are generally defined according to administrative criteria without considering the eventual spatial interaction of the variable of interest. From all these considerations, it is reasonable to assume that the random effects between the neighboring areas (defined, for example, by a contiguity criterion) are correlated and that the correlation decays to zero as distance increases.

Consider a finite population partitioned into D small areas. The basic Fay and Herriot (FH) model relates linearly the quantity of inferential interest for d-th small area, θ_d to a vector of p area level auxiliary covariates $\mathbf{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dp})$, and includes a random effect v_d associated to the area; that is,

$$\theta_d = \mathbf{x}_d \boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D. \tag{1}$$

Here $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters and the random effects $\{v_d; d = 1, \ldots, D\}$ are independent and identically distributed, each with mean 0 and variance σ_v^2. Model (1) is called linking model since all small areas are linked by the common $\boldsymbol{\beta}$. Moreover, the FH model assumes that a design-unbiased direct estimator y_d of θ_d is available for each small area $d = 1, \ldots, D$, and that these direct estimators can be expressed as

$$y_d = \theta_d + e_d, \quad d = 1, \ldots, D, \tag{2}$$

where $\{e_d; d = 1, \ldots, D\}$ are independent sampling errors, independent of the random effects v_d, and where e_d has mean 0 and variance ψ_d assumed to be known, $d = 1, \ldots, D$. See Ghosh and Rao (1994). Model (2) is called sampling model. Combining both the linking model (1) and the sampling model (2) we obtain the linear mixed model

$$y_d = \mathbf{x}_d \boldsymbol{\beta} + v_d + e_d, \quad d = 1, \ldots, D. \tag{3}$$

Let us assume vectors $\mathbf{y} = (y_1, \ldots, y_D)'$, $\mathbf{v} = (v_1, \ldots, v_D)'$ and $\mathbf{e} = (e_1, \ldots, e_D)'$, and matrices $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_D')'$ and $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_D)$. Then the model in matrix notation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e}. \tag{4}$$

Model (4) can be extended to allow for spatially correlated area effects as follows. Let $\mathbf{v}$ be the result of a Simultaneously Autoregressive (SAR) process with unknown autoregression parameter ρ and proximity matrix $\mathbf{W}$ (see Anselin (1988b) and Cressie (1993), i.e.,

$$\mathbf{v} = \rho \mathbf{W}\mathbf{v} + \mathbf{u}. \tag{5}$$

We assume that the matrix $(\mathbf{I}_D - \rho\mathbf{W})$ is non-singular. Then $\mathbf{v}$ can be expressed as

$$\mathbf{v} = (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u}. \tag{6}$$

Here, $\mathbf{u} = (u_1, \ldots, u_D)'$ is a vector with mean $\mathbf{0}$ and covariance matrix $\sigma_u^2\mathbf{I}_D$, where $\mathbf{I}_D$ denotes the $D \times D$ identity matrix and σ_u^2 is an unknown parameter. We consider that the proximity matrix $\mathbf{W}$ is defined in row standardized form; that is, $\mathbf{W}$ is row stochastic. Then, $\rho \in (-1, 1)$ is called spatial autocorrelation parameter (Banerjee et al., 2004). Hereafter, the vector of variance components will be denoted $\omega = (\omega_1, \omega_2)' = (\sigma_u^2, \rho)'$. Equation (6) implies that $\mathbf{v}$ has mean vector $\mathbf{0}$ and covariance matrix equal to

$$\mathbf{G}(\omega) = \sigma_u^2[(\mathbf{I}_D - \rho\mathbf{W})'(\mathbf{I}_D - \rho\mathbf{W})]^{-1}. \tag{7}$$

Since $\mathbf{e}$ is independent of $\mathbf{v}$, the covariance matrix of $\mathbf{y}$ is equal to

$$\mathbf{V}(\omega) = \mathbf{G}(\omega) + \Psi.$$

Combining (4) and (6) the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e} \tag{8}$$

Under model (8), the Spatial BLUP of the quantity of interest $\theta_d = \mathbf{x}_d\boldsymbol{\beta} + v_d$ is

$$\tilde{\theta}_d(\omega) = \mathbf{x}_d\tilde{\boldsymbol{\beta}}(\omega) + \mathbf{b}_d'\mathbf{G}(\omega)\mathbf{V}^{-1}(\omega)[\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(\omega)], \tag{9}$$

where $\tilde{\boldsymbol{\beta}}(\omega) = [\mathbf{X}'\mathbf{V}^{-1}(\omega)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{V}^{-1}(\omega)\mathbf{y}$ is the generalised least squares estimator of the regression parameter $\boldsymbol{\beta}$ and $\mathbf{b}_d'$ is the $1 \times D$ vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the d-th position. The Spatial BLUP $\tilde{\theta}_d(\omega)$ depends on the unknown vector of variance components $\omega = (\sigma_u^2, \rho)'$. The two stage estimator $\tilde{\theta}_d(\hat{\omega})$ obtained by replacing ω in expression (9) by a consistent estimator $\hat{\omega} = (\hat{\sigma}_u^2, \hat{\rho})'$ is called Spatial EBLUP (see Singh et al. (2005), Petrucci and Salvati (2006)).

The classical hypothesis of independence of the random effects is overcome by considering correlated random area effects between neighbouring areas modeled through a SAR process with spatial autocorrelation coefficient and proximity matrix W (Anselin, 1988b). The corresponding estimators of the small area parameters are usually known as Spatial EBLUP (SEBLUP). In addition, the use of SAE models with spatially correlated random area effects gives a possible solution to the problem of estimating the parameter of interest for the areas in which no sample observations are available. With the traditional SAE model, the only prevision available for non-sampled areas is given by the "fixed term" of the mixed model, since the estimation of the random effect is not possible. On the contrary, the hypothesis of correlated random effects allows the estimation of the area-specific effects for all areas, both sampled and non-sampled. The addition of these estimated random effects to the fixed component of the model gives the prediction of the small area parameter in every area.

2.1 AN EXAMPLE IN OFFICIAL STATISTICS: THE AVERAGE PRODUCTION OF OLIVES PER FARM IN 53 ZONES OF THE TUSCANY REGION (ITALY)

The data are from the Farm Structure Survey (FSS, ISTAT 2003). The survey is carried out once every two years. The sample is selected by means of a stratified one-stage design with self-representation of larger farms (agricultural holdings). The sample size is 55,030 farms for Italy and 2,504 for Tuscany. The survey is carried out in order to produce accurate estimates of agricultural production at national and regional levels. In this case study, the target parameter is the farm production of olives in quintals at a subregional level in Tuscany. Tuscany is divided into 53 Agricultural Zones (AZs). They are defined on a geographical basis and are very useful small areas in economic studies. They are determined following the administrative boundaries of the 287 Municipalities of Tuscany. All the AZs are represented in the regional FSS sample. The area level sampling variances have been obtained by estimating the sampling variances of the small area direct estimators.

The exploratory analysis firstly tested the presence of the spatial dependence in the data. Essential to this are the definitions of the spatial location of the AZs and the spatial interaction matrix (W). The centroid of each AZ is considered to be the spatial reference for all the units (in the case of farms for AZs) residing in the same small area and it is defined to be the location of the small area. The Atlas of Coverage of the Tuscany Region maintained by the Geographical

Information System of the Regione Toscana provided all the information on coordinates, extensions and positions of the small areas of interest (UTM system). The Population Census and Agricultural Census databases provided all the auxiliary information related to the average farm production of olives (quintals per farm) and their covariates at small area level.

The spatial interaction matrix (W) for each location specifies which other locations in the system affect the value of the farm production of olives at that location. The elements of W are nonstochastic and exogenous to the SAE model. In our definition the elements of W take nonzero values (they are equal to 1) only for those pairs of AZs, that are contiguous to each other (first-order contiguity). This scheme is common to many real-life situations in the fields of geology, agriculture, and environmental science as well as in certain areas of health studies and epidemiology. Spatial autocorrelation in the target variables and in the auxiliary variables has been checked by means of the two best-known test statistics for spatial autocorrelation: Moran's I and Geary's C (Anselin, 1988a). The best explanatory variable for the target variable is the agricultural surface utilized for the production of olives (measured in hectares). For the covariate, the Moran's I statistics are significant at the 1% level, indicating that similar values are more spatially clustered than what might be purely by chance. This is consistent with the estimated values for Geary's C. Spatial dependence in the target variable is weaker, but still statistically significant.

The per farm production of olives was modelled by the Spatial Fay-Herriot model and by the more traditional Fay-Herriot model. For the spatial model the value of the estimated spatial autoregressive coefficient $\hat{\rho}$ was 0.686 (s.e. = 0.319) and the value of the estimated variance component $\hat{\sigma}_u^2$ was 0.792 (s.e. = 0.604) when we used the REML procedure. The accuracy of the estimates is measured by the coefficients of variation. The mean of the point estimates suggests a production of olives of about six quintals per farm with a slightly lower median value obtained in both the SEBLUP and EBLUP procedures. This is not a surprise as the distribution of the target variable in the population is skewed and concentrated on small production units. The average accuracy of the estimates is not appreciable: the CV is about 30% of the estimates. This can be mainly due to the high dispersion of the sample size in the areas and to the skewness of the distribution of the target variable. EBLUP on average is slightly more variable, even though its performance is in line with that of SEBLUP. The performances of EBLUP and SEBLUP are similar even though the spatial relationship appears to be of medium strength and significant. Indeed, there is not relevant difference between EBLUP

and SEBLUP estimates and also in their accuracy. This could be due to the low and nonsignificant value of the estimated variance component and the wide range of sampling variances.
This application discusses the spatial effects in data used for SAE on the performance of the BLUP obtained under the area level Fay-Herriot model. The performance of the BLUP was compared with that of the SBLUP via a simulation study in which the population was generated according to a spatial Fay-Herriot model and a wide range of values, ranging from -0.75 to 0.75, for the spatial correlation were used. The main finding is that the SBLUP outperforms the BLUP in terms of efficiency and relative bias in cases of both positive and negative spatial correlation, and this result does not depend on the size of the sampling variances in different area groups. In other words, the SBLUP is appropriate when spatial dependency is present in the data used for SAE. Obviously, in real-life situations the parameters of a spatial Fay-Herriot model are not known and must be estimated from survey data. In such a case, attention is devoted to Spatial Empirical BLUP (SEBLUP) and its mean squared errors.

More details of the results of this application can be found in Pratesi and Salvati (2009).

3. GEOADDITTIVE SAE MODELS

Until now, we have considered the spatial structure of the data at the area level: the only information used to build the proximity matrix of the SAR process is about the small area locations. However, if the spatial location is available for every unit, we can try to use it directly as a covariate of the SAE model. The application of bivariate smoothing methods, like kriging, produces a surface interpolation of the variable of interest. In particular, the geoadditive model analyses the spatial distribution of the study variable while accounting for possible covariate effects through a linear mixed model representation. Exploiting the common linear mixed model framework of both small area estimation models and geoadditive models, we can define the geoadditive SAE model. This model will have two random effect components: the area-specific effects and the spatial effects. The geoadditive SAE model belongs to a more general class of models introduced by Opsomer et al. (2008), called non-parametric SAE model, where the non-parametric component is a penalised spline model that accounts for a generic non-linear covariate.

Suppose that there are T small areas for which we want to estimate a quantity of interest and let y_{it} denote the value of the response variable for the ith unit, $i = 1,...,n$, in small area t, $t = 1,...,T$. Let $\mathbf{x}_{it}$ be a vector of p linear covariates associated with the same unit, then the *classic SAE model* is given by

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2), \quad u_t \sim N(0, \sigma_u^2), \tag{10}$$

where $\boldsymbol{\beta}$ is a vector of p unknown coefficients, u_t is the random area effect associated with small area t and ε_{it} is the individual level random error. The two error terms are assumed to be mutually independent, across individuals as well as across areas.

If we define the matrix $\mathbf{D} = [d_{it}]$ with

$$d_{it} = \begin{cases} 1 & \text{if observation } i \text{ is in small area } t, \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and $\mathbf{y} = [y_{it}]$, $\mathbf{X} = \left[\mathbf{x}_{it}^T\right]$, $\mathbf{u} = [u_t]$ and $\boldsymbol{\varepsilon} = [\varepsilon_{it}]$, then the matrix notation of (10) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{12}$$

with

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_u^2\mathbf{I}_T & 0\\ 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

The covariance matrix of $\mathbf{y}$ is

$$\mathrm{Var}(\mathbf{y}) \equiv \mathbf{V} = \sigma_u^2\mathbf{D}\mathbf{D}^T + \sigma_\varepsilon^2\mathbf{I}_n$$

and the BLUPs of the model coefficients are

$$\begin{aligned}\boldsymbol{\beta} &= \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},\\ \mathbf{u} &= \sigma_u^2\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

If the variance components σ_u^2 and σ_ε^2 are unknown, they are estimated by REML or ML methods and the model coefficients are obtained with the EBLUPs.

The formulation (12) is a linear mixed model, analogous to the geoadditive model (Kammann and Wand (2003)), thus it is straightforward to compose the geoadditive SAE model. Consider again the response y_{it} and the vector of p linear covariates $\mathbf{x}_{it}$, and suppose that both are measured at a spatial location $\mathbf{s}_{it}$, $\mathbf{s} \in$

$\Re^2$. The *geoadditive SAE model*[2] for such data is a linear mixed model with two random effects components:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{13}$$

with

$$\mathrm{E}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_\gamma^2\mathbf{I}_K & 0 & 0\\ 0 & \sigma_u^2\mathbf{I}_T & 0\\ 0 & 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

Now $\mathbf{X} = \left[\mathbf{x}_{it}^T, \mathbf{s}_{it}^T\right]_{1\le i\le n}$ has $p+2$ columns, $\boldsymbol{\beta}$ is a vector of $p+2$ unknown coefficients, $\mathbf{u}$ are the random small area effects, $\boldsymbol{\gamma}$ are the thin plate spline coefficients (seen as random effects) and $\boldsymbol{\varepsilon}$ are the individual level random errors. Matrix $\mathbf{D}$ is still defined by (11) and $\mathbf{Z}$ is the matrix of the thin plate spline basis functions

$$\mathbf{Z} = \left[C\left(\mathbf{s}_i - \boldsymbol{\kappa}_k\right)\right]_{1\le i\le n, 1\le k\le K}\left[C\left(\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k\right)\right]_{1\le h,k\le K}^{-1/2},$$

with K knots $\boldsymbol{\kappa}_k$ and $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log\|\mathbf{r}\|$.

Again, the unknown variance components are estimated via REML or ML estimators and are indicated with $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$. The estimated covariance matrix of $\mathbf{y}$ is

$$\hat{\mathbf{V}} = \hat{\sigma}_\gamma^2\mathbf{Z}\mathbf{Z}^T + \hat{\sigma}_u^2\mathbf{D}\mathbf{D}^T + \hat{\sigma}_\varepsilon^2\mathbf{I}_n \tag{14}$$

and the EBLUP estimators of the model coefficients are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{15}$$

$$\hat{\boldsymbol{\gamma}} = \hat{\sigma}_\gamma^2\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{16}$$

$$\hat{\mathbf{u}} = \hat{\sigma}_u^2\mathbf{D}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{17}$$

For a given small area t, we are interested in predicting the mean value of y

$$\bar{y}_t = \bar{\mathbf{x}}_t\boldsymbol{\beta} + \bar{\mathbf{z}}_t\boldsymbol{\gamma} + u_t$$

where $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{z}}_t$ are the true means over the small area t and are assumed to be known. The EBLUP for the quantity of interest is

$$\hat{\bar{y}}_t = \bar{\mathbf{x}}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_t\hat{\boldsymbol{\gamma}} + \mathbf{e}_t\hat{\mathbf{u}} \tag{18}$$

where $\mathbf{e}_t$ is a vector with 1 in the t-th position and zeros elsewhere.

[2] The same model formulation is in Opsomer et al. (2008), where is presented a model, called by the authors *non-parametric SAE model*, that accounts for a generic non-linear covariate.

3.1 AN EXAMPLE IN OFFICIAL STATISTICS: HOUSEHOLD PER-CAPITA CONSUMPTION EXPENDITURE IN ALBANIA

An application of a geoadditive model for official statistics is shown in (Bocci, 2009). In particular a geoadditive small area estimation model is applied in the field of poverty mapping at small area level in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania. The model parameters estimated using the dataset of the 2002 Living Standard Measurement Study (LSMS) is combined with the 2001 Population and Housing Census (PHC) covariate information. PHC and LSMS are both conducted by the INSTAT (Albanian Institute of Statistics).

The 2002 LSMS provides individual level and household level socio-economic data from 3,599 households drawn from urban and rural areas in Albania. Geographical referencing data on the longitude and latitude of each household were also recorded using portable GPS devices (World Bank and INSTAT, 2003). The sample was designed to be representative of Albania as a whole, Tirana, other urban/rural locations, and the three main agro-ecological areas (Coastal, Central, and Mountain). The survey was carried out by the Albanian Institute of Statistics (INSTAT) with the technical and financial assistance of the World Bank.

The Republic of Albania is divided in 3 geographical levels: prefectures, districts and communes. There are 12 prefectures, 36 districts and 374 communes, however the LSMS survey, which provides valuable information on a variety of issues related to living conditions in Albania, is stratified in 4 large strata (Costal Area, Central Area, Mountain Area and Tirana) and these strata are the smaller domain of direct estimation.

The covariates selected to fit the geoadditive SAE model are chosen following prior studies on poverty assessment in Albania (Betti et al., 2003; Neri et al., 2005). The selected household level covariates are: *size of the household* (in term of number of components); *information on the components of the household* (age of the householder, marital status of the householder, age of the spouse of the householder, number of children 0-5 years, age of the first child, number of components without work, highest level of education in the household); *information on the house* (building with 2-15 units, built with brick or stone, built before 1960, number of rooms per person, house surface < 40 m^2, house surface $40 - 69$ m^2, wc inside); *presence of facilities in the dwelling* (TV, parabolic, refrigerator, washing machine, air conditioning, computer, car); *ownership of agricultural land.*

All these variables are available both in LSMS and PHC surveys (see Neri et al. (2005) for comparability between the two sources); in addition, the geographical location of each household is available for the LSMS data.

The response variable is the logarithm of the household per-capita consumption expenditure. The use of the logarithmic transformation is typical for this type of data as it produces a more suitable response for the regression model.

Estimates of the log per-capita consumption expenditure in each of the 36 district area are derived using the geoadditive SAE model presented in (13).

After the preliminary analysis of various combination of parametric and non-parametric specifications for the selected covariates, the chosen model is composed by a bivariate thin plate spline on the universal transverse Mercator (UTM) coordinates, a linear term for all the other variables and a random intercept component for the area effect.

Almost all the parameters are highly significant at 5% level. The exceptions are the coefficients of 'marital status of the householder', 'number of children 0-5 years' and 'built with brick or stone' that are significant at 5% level, and the coefficient of 'building with 2-15 units' that is significant at 10% level.

The geoadditive SAE model (13) considers two random effects, once for the bivariate spline smoother and once for the small area effect, thus the estimated value of the log per-capita consumption expenditure in a specific location is obtained as sum of two components, once continuous over the space and once constant in each small area showing the presence of both a spatial dynamic and a district level effect in the Albanian consumption expenditure.

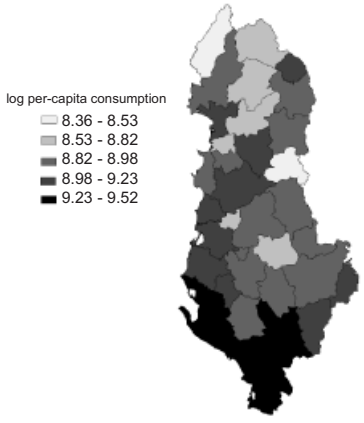

Figure 1: District level estimates of the mean of household log per-capita consumption expenditure.

The estimated parameters are then combined with the census mean values to obtain the district level estimates of the average household log per-capita consumption expenditure.

The mean square errors (MSEs), and consequently the coefficients of variations (CVs), are calculated using the robust MSE estimator of Salvati et al. (2010). All the CVs are less that 2%, with a mean value of 0.91%, thus the estimates have low variability. The higher values are registered in those districts where the sample size is quite low.

The results (Figure 1) show a clear geographical pattern, with the higher values in the south and south-west of the country and the lower value in the mountainous area (north and north-east). These results are consistent with previous applications on the same datasets presented in literature (Neri et al., 2005; Tzavidis et al., 2008). Refer to Bocci (2009) for more details on this application.

4. M-QUANTILE GWR SPATIAL MODELS

Following the M-quantile approach it is possible to specify a local M-quantile small area model via an M-quantile GWR model. Unlike SAR mixed models, M-quantile GWR models are local models that allow for a spatially non-stationary process in the mean structure of the model. This is obtained by assuming that the regression coefficients vary spatially across the geography of interest. The spatial extension to linear M-quantile regression is based on Geographically Weighted Regression (GWR) (see Brunsdon et al. (1996)) that extends the traditional regression model by allowing local rather than global parameters to be estimated. Here we report a brief description of the M-quantile GWR model following a recent paper by Salvati et al. (2012).

Given n observations at a set of L locations $\{u_l; l = 1,\ldots,L; L \leqslant n\}$ with n_l data values $\{(y_{jl}, \mathbf{x}_{jl}); i = 1,\ldots,n_l\}$ observed at location u_l, a linear GWR model is a special case of a locally linear approximation to a spatially non-linear regression model and is defined as follows

$$y_{jl} = \mathbf{x}_{jl}^T \boldsymbol{\beta}(u_l) + \varepsilon_{jl}, \tag{19}$$

where $\boldsymbol{\beta}(u_l)$ is a vector of p regression parameters that are specific to the location u_l and the ε_{il} are independently and identically distributed random errors with zero expected value and finite variance. The value of the regression parameter 'function' $\boldsymbol{\beta}(u)$ at an arbitrary location u is estimated using weighted least squares

$$\hat{\boldsymbol{\beta}}(u) = \Big\{ \sum_{l=1}^{L} w(u_l,u) \sum_{i=1}^{n_l} \mathbf{x}_{jl}\mathbf{x}_{jl}^T \Big\}^{-1} \Big\{ \sum_{l=1}^{L} w(u_l,u) \sum_{i=1}^{n_l} \mathbf{x}_{jl} y_{jl} \Big\},$$

where $w(u_l,u)$ is a spatial weighting function whose value depends on the distance from sample location u_l to u in the sense that sample observations with locations close to u receive more weight than those further away. In this paper we use a Gaussian specification for this weighting function

$$w(u_l,u) = \exp\Big\{ -d_{u_l,u}^2/2b^2 \Big\}, \tag{20}$$

where $d_{u_l,u}$ denotes the Euclidean distance between u_l and u and $b > 0$ is the bandwidth. As the distance between u_l and u increases the spatial weight decreases exponentially. For example, if $w(u_l,u) = 0.5$ and $w(u_m,u) = 0.25$ then observations at location u_l have twice the weight in determining the fit at location u compared with observations at location u_m. See Fotheringham et al. (2002) for a discussion of other weighting functions.

We can extend the M-quantile model

$$Q_q(\mathbf{x}_j;\psi) = \mathbf{x}_j^T \boldsymbol{\beta}_\psi(q). \tag{21}$$

to specify a linear model for the M-quantile of order q of the conditional distribution of y given $\mathbf{X}$ at location u, writing

$$Q_q(\mathbf{x}_{jl};\psi,u) = \mathbf{x}_{jl}^T \boldsymbol{\beta}_\psi(u;q), \tag{22}$$

where now $\boldsymbol{\beta}_\psi(u;q)$ varies with u as well as with q. Like (19), we can interpret (22) as a local linear approximation, in this case to the (typically) non-linear order q M-quantile regression function of y on $\mathbf{X}$, thus allowing the entire conditional distribution (not just the mean) of y given $\mathbf{X}$ to vary from location to location. The parameter $\boldsymbol{\beta}_\psi(u;q)$ in (22) at an arbitrary location u can be estimated by solving

$$\sum_{l=1}^{L} w(u_l,u) \sum_{j=1}^{n_l} \psi_q\{y_{jl} - \mathbf{x}_{jl}^T \boldsymbol{\beta}_\psi(u;q)\}\mathbf{x}_{jl} = \mathbf{0}, \tag{23}$$

where $\psi_q(\varepsilon) = 2\psi(s^{-1}\varepsilon)\{qI(\varepsilon > 0) + (1-q)I(\varepsilon \leqslant 0)\}$, s is a suitable robust estimate of the scale of the residuals $y_{jl} - \mathbf{x}_{jl}^T \boldsymbol{\beta}_\psi(u;q)$, e.g. $s = median|y_{jl} - \mathbf{x}_{jl}^T \boldsymbol{\beta}_\psi(u;q)|/0.6745$, and we will typically assume a Huber Proposal 2 influence function, $\psi(\varepsilon) = \varepsilon I(-c \leqslant \varepsilon \leqslant c) + sgn(\varepsilon) I(|\varepsilon| > c)$. Provided c is bounded away from zero, we can solve (23) by combining the iteratively re-weighted least

squares algorithm used to fit the 'spatially stationary' M-quantile model (21) and the weighted least squares algorithm used to fit a GWR model (Salvati et al., 2012).

Note that estimates of the local (GWR) M-quantile regression parameters are derived by solving the estimating equation (23) using iterative re-weighted least squares, without any assumption about the underlying conditional distribution of y_{jl} given $\mathbf{x}_{jl}$ at each location u_l. That is, the approach is distribution-free. For details see Salvati et al. (2012). SAR models allow for spatial correlation in the model error structure to be used to improve SAE. Alternatively, this spatial information can be incorporated directly into the M-quantile regression structure via an M-quantile GWR model for the same purpose.

We now assume that we have only one population value per location, allowing us to drop the index l. We also assume that the geographical coordinates of every unit in the population are known, which is the case with geo-coded data. The aim is to use these data to predict the area d mean of y using the M-quantile GWR model (22).

Following Chambers and Tzavidis (2006), and provided there are sample observations in area d, an area d specific M-quantile GWR coefficient, $\hat{\theta}_d$ can be defined as the average value of the sample M-quantile GWR coefficients in area d, otherwise we set $\hat{\theta}_d = 0.5$. Following Tzavidis et al. (2010), the bias-adjusted M-quantile GWR predictor of the mean $\bar{Y}_d$ in small area d is then

$$\hat{\bar{Y}}_d^{MQGWR/CD} = N_d^{-1}\Big[\sum_{j\in U_d}\hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j) + \frac{N_d}{n_d}\sum_{j\in s_d}\{y_j - \hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j)\}\Big], \quad (24)$$

where $\hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j)$ is defined via the MQGWR model (22).

4.1 AN EXAMPLE IN OFFICIAL STATISTICS: ESTIMATES OF THE INCOME AVERAGE AT MUNICIPALITY LEVEL IN TUSCANY

The aim of this application is to estimate the mean of equivalised household income at municipality level $\bar{Y}_d$ in Tuscany using $\hat{\bar{Y}}_d^{MQGWR/CD}$, the M-quantile GWR predictor of the mean. The maps of Figure 2 will contrast the results of the MQGWR model with the results obtained by the ordinary M-quantile linear model.

Available data to measure poverty and living conditions in Italy come mainly from sample surveys, such as the Survey on Income and Living Conditions (EU-

SILC). The data on the equivalised income in 2007 for 59 of the 287 Tuscany municipalities are available from the EU-SILC survey 2008. However, EU-SILC data can be used to produce accurate estimates only at the NUTS 2 level (that is, regional level). To satisfy the increasing demand from official and private institutions of statistical estimates on poverty and living conditions referring to smaller domains (LAU 1 and LAU 2 levels, that is provinces and municipalities), there is the need to resort to small area methodologies.

A set of explanatory variables is available for all the 287 municipalities from the Population Census 2001. We employ the M-Quantile GWR model for estimating the mean of household income in each of the 287 Municipalities (LAU 2) Note that with the spatial information included in the model we can obtain estimates for the 228 out of sample areas (areas with no sample units in it).

The selection of covariates to fit the small area models relies on prior studies of poverty assessment and on the availability of data.

More details can be found in the deliverables of the SAMPLE project (7FP Small Area Methods for Poverty and Living conditions Estimates www.sample-project.eu). In this example the MQGWR model uses spatial information to estimate the target statistics in the out-of-sample areas. Indeed synthetic estimates in the out-of-sample areas can be obtained also using the M-quantile linear model: this can be done letting the area representative quantile, θ_i, be equal to 0.5.

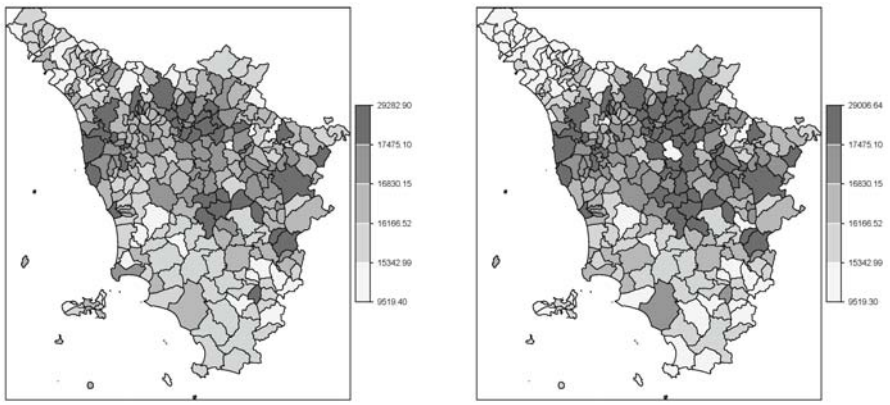

Figure 2: Estimates of the mean of equalized household income in the Municipalities of Tuscany

In Tuscany there is evidence of relevant (relative) poverty in the province of Massa-Carrara and Grosseto (Figure 2). However if analysed as stand alone region we can see dissimilarities between provinces and municipalities. Using spatial information we obtained estimates of the averages of the households equivalised income at LAU 2 level in Tuscany: looking at the estimates some dissimilarities between the provinces emerge. Above all under MQGWR model we capture more heterogeneity among municipalities. This results show the importance to "go deeper", i.e. obtain estimates at the lowest domain level and, at the same time, enphasise the importance of spatial information; see Giusti et al. (2012) for further details.

5. CONCLUSIONS

The increasing request of small area statistics is motivated by their growing use in formulating policies and programmes, in the allocation of government funds and in regional planning. Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on the local socio-economic, environmental and other conditions. Statistical surveys produce high quantities of data and estimates, but cost constraints in the design of sample surveys lead to small sample sizes within small areas. As a result, direct estimation using only the survey data is inappropriate as it yields estimates with unacceptable levels of precision. In such cases increasing the sample size can be a feasible alternative to small area estimation but may be too expensive even for national statistical institutes (see also the SMART system active on the Istat website http://smart.istat.it/smart/).

Small area estimation (SAE) is performed via models that "borrow strength" by using all the available data and not only the area specific data. Auxiliary information can also consist of geo-coded data about the spatial distribution of the domains and units of interest, obtained via geographic information systems. For example, the data can be obtained from digital maps that cover the domains of interest and so allow for the calculation of the centroids of the these domains, their borders, perimeter, areas and the distances between them. All these attributes are commonly available in official statistical agencies and they are helpful in the analysis of socio-economic data relating to these domains since these often display spatial structure, i.e. they are correlated with the so-called geography of the landscape.

In this context it is useful to recall the so called first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). The law is valid also for small geographical areas: close areas are more likely to have similar values of the target parameter than areas that are far from each other.

The applications presented here show the importance of using georeferenced information and it is evident that an adequate use of geographic information and geographical modelling can help in producing more accurate estimates for small area parameters.

The direct survey estimates based on small sample sizes can be considerably improved by using the area specific small area models. The spatial autocorrelation amongst the neighbouring areas may be exploited for improving the direct survey estimates. However, the model can be applied after studying the significant correlation amongst the small areas by virtue of their neighbourhood effects. In case of poor relation between the dependent and exogenous variables, the simple spatial model with intercept only, may equally improve the estimates. This model uses only the spatial autocorrelation to strengthen the small area estimates and do not require the use of exogenous variables. The spatial models, by using the appropriate weight matrix W, or a combination of weight matrices, can considerably improve the estimates. Weight matrix should be based on logical considerations and it may be used effectively for the cases, if for some reasons, reliable exogenous variables are not available. In addition one has to be careful about the increase in the MSE due to the variability caused by replacing the parameters by their estimates.

Spatial models in SAE help to exploit the information from the spatial distribution of individuals, groups and institutions, making it possible for the researchers to examine aspects that might not otherwise be evaluated, allowing gains of interpretation and knowledge of the phenomena under study and obtaining more accurate estimates.

References

Anselin, L. (1988a). *Spatial Econometrics*. Methods and Models, Boston.

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

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. New York: Chapman and Hall.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satelite data. In *Journal of the American Statistical Association,* 83: 28–36.

Betti, G., Ballini, F., and Neri, L. (2003). *Poverty and Inequality Mapping in Albania, Final Report to the World Bank*. World Bank.

Bocci, C. (2009). *Geoadditive Models for Data with Spatial Information*. Ph.D. thesis, Ph.D. in Applied Statistics, Department of Statistics, University of Firenze, Italy.

Brunsdon, C., Fotheringham, A.S., and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. In *Geographical Analysis,* 28: 281–298.

Chambers, R.L. and Tzavidis, N. (2006). M-quantile models for small area estimation. In *Biometrika,* 93: 255–268.

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

Cressie, N. (1993). *Statistics for Spatial Data (revised edition)*. Wiley, New York.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data. In *Journal of the American Statistical Association,* 74: 269–277.

Fotheringham, A.S., Brundson, C., and M., C. (2002). *Geographically Weighted Regression - The analysis of spatially varying relationship*. West Sussex, England: John Wiley & Sons Ltd.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). In *Statistical Science,* 9: 55–93.

Giusti, C., Marchetti, M., Pratesi, M., and Salvati, N. (2012). Robust small area estimation and oversampling in the estimation of poverty indicators. In *Survey Research Methods,* 6: 155–163.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). In *Test,* 15: 1–96.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. In *Applied Statistics,* 52: 1–18.

Neri, L., Ballini, F., and Betti, G. (2005). Poverty and inequality mapping in transition countries. In *Statistics in Transition,* 7: 135–157.

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

Petrucci, A., Pratesi, M., and Salvati, N. (2005). Geographic information in small area estimation: small area models ans spatially correlated random area effects. In *Statistics in Transition,* 7: 609–623.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. In *Journal of Agricultural, Biological and Environmental Statatistics,* 11: 169–182.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. In *International Statistical Review,* 70: 125–143.

Pratesi, M. and Salvati, N. (2008). Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects. In *Statistical Methods and Applications,* 17: 113–141.

Pratesi, M. and Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. In *Journal of Official Statistics,* 25: 37–53.

Rao, J.N.K. (2003a). *Small Area Estimation*. John Wiley & Sons, New York.

Rao, J.N.K. (2003b). *Small Area Estimation*. John Wiley.

Saei, A. and Chambers, R. (2005). *Small area estimation under linear and generalized linear mixed models with time and area effects*. Working Paper M03/15. Southampton Statistical Sciences Research Institute, University of Southampton.

Salvati, N., Chandra, H., Ranalli, M.G., and Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator. In *Computational Statistics and Data Analysis,* 54: 2159–2171.

Salvati, N., Pratesi, M., Tzavidis, N., and Chambers, R. (2012). Small area estimation via M-quantile geographically weighted regression. In *Test,* 21: 1–28.

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

Tobler, W.R. (1970). A computer movie simulating urban growth in the detroit region. In *Economic Geography,* 46: 234–240.

Tzavidis, N., Marchetti, S., and Chambers, R. (2010). *Robust estimation of small area means and quantiles*. Australian and New Zealand Journal of Statistics DOI 10.

Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R. (2008). M-quantile models with application to poverty mapping. In *Statistical Methods and Applications,* 17: 393–411.

World Bank and INSTAT (2003). Albania Living Standard Measurement Survey 2002. *Basic Information Document*. URL http://go.worldbank.org/IDTKJRT8Y0.