Dipartimento di Statistica
"Giuseppe Parenti"

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

WORKING PAPER 2010/14

Geoadditive Small Area Model for the Estimation of Consumption Expenditure in Albania

Chiara Bocci

Università degli Studi
di Firenze

Geoadditive Small Area Model for the Estimation of Consumption Expenditure in Albania

Chiara Bocci
Department of Statistics "G. Parenti", University of Florence

Abstract

In the last few years the demand of spatially detailed statistical data is considerably increased due also to the development of statistical methods for small area. In the past, the high degree of spatial detail of such information was not so useful for practical purposes as firms and local authorities were interested in information aggregated at some pre-specified level. However, the area definition and the assignment of the data to appropriate areas can pose problems in the estimation process. In particular, in small area estimation the importance of this matter is represented by the fact that some parameters of the model can be related to the between-area relationships. Geoadditive models can face this problem analyzing directly the spatial distribution of the study variable while accounting for possible covariate effects. This paper presents the implementation of a geoadditive model to small area estimation. The geoadditive SAE model is apply in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania.

Keywords: spatial statistics, semiparametric methods, socio-economic data.

1 Introduction

The analysis of the regional spatial pattern of socio-economic processes has become a relevant area of statistics and economics. Since the early seventies, regional economics has been defined as the field concerned with the role of space, distance and regional differentiation in economics (Richardson, 1970).

Several reasons support this subject: first, the spatial clustering of economic activities is a product of the regional differences and could reflect individual inequalities that are object of policies; second, the geographical pattern can have great influence on the results of economic policies; and third, exploring spatial clustering of economic activities is a relevant input to model economic theories at a regional scale.

Research in this area focuses on the specification and estimation of spatial effects in a theoretical economic model, and on the use of such estimates to obtain spatial interpolations and predictions of the study variables. The set of methodologies concerned with this target belongs to the field of spatial econometrics

(Anselin, 1988; Arbia, 2006), that is defined by Anselin (1988) as the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models.

The explosive growth of spatial data and widespread use of spatial databases have emphasized the need for the discovery of spatial knowledge. Moreover, very rich databases of spatially referenced socio-economic data are available from local statistical offices and in the last few years the demand of spatially detailed statistical data is dramatically increased.

Nowadays, the fields of spatial statistics is broadly understood. In general, spatial statistics is concerned with statistical and mathematical descriptors of spatial structure and it focuses on the nature of space and spatial data. In this way it can face with problems which are characterized by the difficulties associated with assessing the importance of spatial dependence and spatial heterogeneity, the so-called "spatial effects" mentioned before.

Extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional techniques for extracting spatial patterns. Therefore, the area definition and the assignment of the data to appropriate areas can pose problems in the estimation process.

It is worth to stress the usefulness of the geographical location for the analysis of non stationary spatial phenomena. The "global" dependence models, such as the classical regression model, assume the independence of the data from the spatial location, generate spatially autocorrelated residuals and bring often to wrong conclusions. Thus, statistical models which take into account the spatial variability can help to understand the underlying phenomenon.

Nonparametric and semiparametric models are attractive alternatives to parametric models because they admit at the start that the true model structure is unknown. However, nonparametric estimation suffers from the rapidly increase of the variance of the estimates with the number of variables. In this situation, semiparametric models become an effective alternative to full nonparametric estimation. The advantage of the semiparametric approach is that it imposes parametric structure where it may be reasonable, while leaving the structure of the model unrestricted for another set of variables. Thus, the semiparametric approach is a particularly easy and flexible approach for modeling broad spatial trends while also permitting the effects of other explanatory variables to vary by location.

In social sciences, maps are useful tools to describe the spatial distribution of poverty in a country, especially when they represent small geographic units, such as municipalities or districts. This information is extremely useful to policymakers and researchers in order to formulate efficient policies and programs.

As pointed out in Neri et al. (2005), in order to produce poverty maps, large data sets are required which include reasonable measures of income or consumption expenditure and which are representative and of sufficient size at low levels of aggregation to yield statistically reliable estimates. Household budget surveys or living standard surveys covering income and consumption usually used to calculate distributional measures are rarely of such a sufficient size; whereas census or other large sample surveys large enough to allow disaggregation have little or no information regarding monetary variables. Then, the required small area estimates are usually based on a combination of sample surveys and administrative data.

This study discusses a new approach to identify and include the spatial pattern in small area estimation, using recent advances in semiparametric models that allow incorporation of spatial location as an additional component. Thus, the estimated spatial patterns reflect the propensity of the considered characteristic in a region, after controlling for other unit-level effects. In particular, the work focuses on socio-economic data collected by the World Bank program on Living Standard Measurement Study (LSMS) (Grosh and Glewwe, 2000). The program is designed to assist policy makers in their efforts to identify how policies could be designed and improved to positively affect outcomes in health, education, economic activities, housing and utilities, etc.

We apply a geoadditive small area estimation model in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania. We combine the model parameters estimated using the dataset of the 2002 Living Standard Measurement Study with the 2001 Population and Housing Census covariate information.

The paper is structured as follows. In the next section, the methodology is extensively discussed. Section 3 describes the datasets used in the analysis and presents the empirical results. In section 4 we discuss the use of two possible MSE estimators through a desing-based simulation study. The final section summarizes the main findings and discusses possible future works.

2 Methodology

Geostatistical methodologies are concerned with the problem of producing a map of a quantity of interest over a particular geographical region based on measurement taken at a set of locations in the region. The aim of such a map is to describe and analyze the geographical pattern of the phenomenon of interest.

These methodologies are born and apply in areas such as environmental studies and epidemiology, where the spatial information is traditionally recorded and available. However, in the last years the diffusion of spatially detailed statistical data is considerably increased and these kind of procedures - possibly with appropriate modifications - can be used as well in any statistical fields of application.

Basically, to obtain a surface estimate we can exploit the exact knowledge of the spatial coordinates (latitude and longitude) of the studied phenomenon by using bivariate smoothing techniques, such as kernel estimate or kriging (Cressie, 1993; Ruppert et al., 2003). However, usually the spatial information alone does not properly explain the pattern of the response variable and we need to introduce some covariates in a more complex model.

Geoadditive models, introduced by Kammann and Wand (2003), answer this problem as they analyze the spatial distribution of the study variable while accounting for possible linear or non-linear covariate effects. Under the additivity assumption they can handle such covariate effects by merging an additive model (Hastie and Tibshirani, 1990) - that accounts for the relationship between the variables - and a kriging model - that accounts for the spatial correlation - and by expressing both as a linear mixed model. The linear mixed model representation is a useful instrument because it allows estimation using mixed model methodology and software.

Let r_i, $1 \leq i \leq n$, be a continuous predictor of y_i at spatial location $\mathbf{s}_i$, $\mathbf{s} \in \Re^2$. A geoadditive model for such data can be formulated as

$$\boxed{y_i = f(r_i) + h(\mathbf{s}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2),} \tag{1}$$

where f is an unspecified smooth function of one variable and h is an unspecified bivariate smooth functions.

Considering a low-rank truncated linear spline for f and a low-rank thin plate spline for h, the model (1) can be written as a mixed model (Kammann and Wand, 2003)

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},} \tag{2}$$

with

$$\mathrm{E}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_r^2\mathbf{I}_{K_r} & 0 & 0\\ 0 & \sigma_s^2\mathbf{I}_{K_s} & 0\\ 0 & 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

where

$$\begin{aligned}\mathbf{X} &= \left[1, r_i, \mathbf{s}_i^T\right]_{1\leq i\leq n},\\ \boldsymbol{\beta} &= \left[\beta_0, \beta_r, \boldsymbol{\beta}_s^T\right],\\ \boldsymbol{\gamma} &= \left[\gamma_1^r, \ldots, \gamma_{K_r}^r, \gamma_1^s, \ldots, \gamma_{K_s}^s\right],\end{aligned}$$

and $\mathbf{Z}$ is obtained by concatenating the matrices containing spline basis functions to handle f and h, respectively

$$\boxed{\mathbf{Z} = \left[\mathbf{Z}_r | \mathbf{Z}_s\right],}$$

$$\begin{aligned}\mathbf{Z}_r &= \left[(r_i - \kappa_1^r)_+, \ldots, (r_i - \kappa_{K_r}^r)_+\right]_{1\leq i\leq n},\\ \mathbf{Z}_s &= \left[C\left(\mathbf{s}_i - \boldsymbol{\kappa}_k^s\right)\right]_{1\leq i\leq n, 1\leq k\leq K_s} \cdot \left[C\left(\boldsymbol{\kappa}_h^s - \boldsymbol{\kappa}_k^s\right)\right]_{1\leq h,k\leq K_s}^{-1/2},\end{aligned}$$

where $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log \|\mathbf{v}\|$ and $\kappa_1^r, ..., \kappa_{K_r}^r$ and $\boldsymbol{\kappa}_1^s, ..., \boldsymbol{\kappa}_{K_s}^s$ are the knots locations for the two functions.

The amount of smoothing for both the additive component and the geostatistical component of the model can be quantified through the variance components ratios $\sigma_\varepsilon^2/\sigma_r^2$ and $\sigma_\varepsilon^2/\sigma_s^2$.

The addition of others explicative variables is straightforward: smoothing components are added in the random effects term $\mathbf{Z}\boldsymbol{\gamma}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}\boldsymbol{\beta}$ term. Moreover, the mixed model structure provides a unified and modular framework that allows to easily extend the model to include various kind of generalization and evolution (Ruppert et al., 2009). In particular, under this framework the geaoddtive model and the classic small area estimation (SAE) model can be easily combined (Opsomer et al., 2008).

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 (Rao, 2003) is given by

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2), \quad u_t \sim N(0, \sigma_u^2), \tag{3}$$

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, both across individuals as well as across areas.

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

$$d_{it} = \begin{cases} 1 & \text{if observation } i \text{ is in small area } t, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

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 (3) is

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon},} \tag{5}$$

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 (5) is a linear mixed model, analogous to the geoadditive model (2), thus it is straightforward to compose the geoadditive SAE model, which is a particular specification of the non-parametric SAE model introduced by Opsomer et al. (2008). 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 for such data is a linear mixed model with two random effects components:

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon},} \tag{6}$$

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 (4) 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 $C(\mathbf{v}) = \|\mathbf{v}\|^2\log\|\mathbf{v}\|$.

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

$$\hat{\mathbf{V}} = \hat\sigma_\gamma^2\mathbf{Z}\mathbf{Z}^T + \hat\sigma_u^2\mathbf{D}\mathbf{D}^T + \hat\sigma_\varepsilon^2\mathbf{I}_n \tag{7}$$

and the EBLUP estimators of the model coefficients are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{8}$$

$$\hat{\boldsymbol{\gamma}} = \hat\sigma_\gamma^2\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{9}$$

$$\hat{\mathbf{u}} = \hat\sigma_u^2\mathbf{D}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{10}$$

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

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

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

$$\boxed{\hat{\bar y}_t = \bar{\mathbf{x}}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_t\hat{\boldsymbol{\gamma}} + \mathbf{e}_t\hat{\mathbf{u}}} \tag{11}$$

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

3 Estimation of the Household Per-capita Consumption Expenditure in Albania

3.1 Data

The Republic of Albania is divided in 3 geographical levels: there are 12 prefectures, 36 districts and 374 communes. The two main sources of statistical information available in Albania are the 2001 Population and Housing Census (PHC) and the 2002 Living Standard Measurement Study (LSMS), both conducted in Albania by the INSTAT (Albanian Institute of Statistics).

The 2002 LSMS provides individual level and household level socio-econo-mic data from 3,599 households drawn from urban and rural areas in Albania. 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).

Four survey instruments were used to collect information for the 2002 Albania LSMS: a household questionnaire, a diary for recording household food consumption, a community questionnaire, and a price questionnaire. The household questionnaire included all the core LSMS modules as defined in Grosh and Glewwe (2000), plus additional modules on migration, fertility, subjective poverty, agriculture, and nonfarm enterprises. Geographical referencing data on the longitude and latitude of each household were also recorded using portable GPS devices (World Bank and INSTAT, 2003).

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). We selected the following household level covariates:

- *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 produce a more suitable response for the regression model (see the distributions presented in Figure 1).

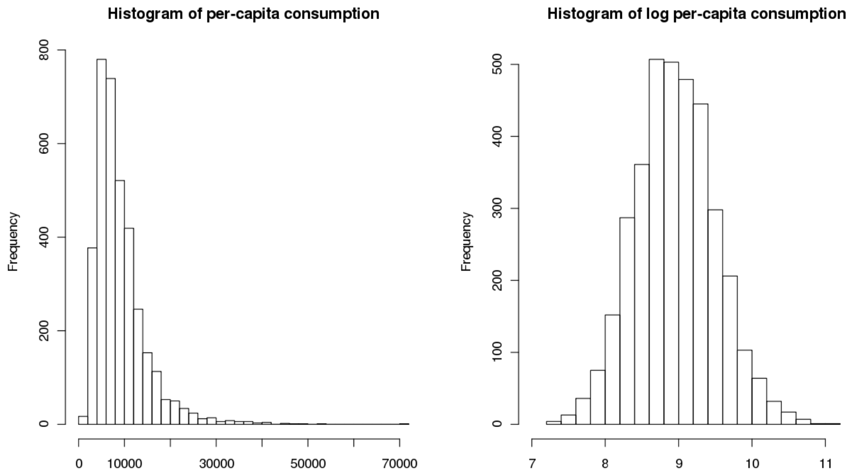

Figure 1: Distribution of the household per-capita consumption expenditure, both in original scale and in logarithmic scale.

3.2 Results

Estimates of the log per-capita consumption expenditure in each of the 36 district area are derived using the geoadditive SAE model presented in (6).

After the preliminary analysis of various combination of parametric and nonparametric 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. The spline knots are selected setting $K = 100$ and using the *clara* space filling algorithm of Kaufman and Rousseeuw (1990) that is available in the `R` package `cluster`. The model is then fitted by REML using the `lme` function in the `R` package `nlme`.

The estimated parameters are presented in Table 1, along with their confidence interval at 95% and the p-values. Excluding the intercept and the coordinates coefficients (that are required by the model structure), almost all the parameters are highly significant. 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 resulting spatial smoothing of the log per-capita consumption expenditure is presented in Figure 2. The geoadditive SAE model (6) considers two random

Table 1: Estimated parameters of the geoadditive SAE model for the household log per-capita consumption expenditure at district level.

Parameter	Estimate	Confidence Interval	p-value
Fixed Effects			
Intercept	7.11	(-34.32;48.55)	0.736
X coordinate	-0.0594	(-0.7807;0.6618)	0.872
Y coordinate	0.0393	(-0.8700;0.9487)	0.932
household size	-0.0775	(-0.0913;-0.0638)	< 0.001
age of the householder	0.0029	(0.0014;0.0044)	< 0.001
marital status of the householder	0.0745	(0.0004;0.1485)	0.049
age of the spouse or husband	-0.0021	(-0.0035;-0.0008)	0.001
number of children 0-5 years	-0.0202	(-0.0382;-0.0023)	0.027
age of the first child	-0.0023	(-0.0037;-0.0009)	0.001
number of components without work	-0.0661	(-0.0784;-0.0537)	< 0.001
high level of education	0.0913	(0.0648;0.1178)	< 0.001
medium level of education	0.2397	(0.2007;0.2788)	< 0.001
building with 2-15 units	0.0261	(-0.0034;0.0557)	0.083
built with brick or stone	0.0342	(0.0001;0.0684)	0.049
built before 1960	-0.0442	(-0.0734;-0.0151)	0.003
number of rooms per person	0.1364	(0.1037;0.1690)	< 0.001
house surface < 40 m^2	-0.0518	(-0.0932;-0.0105)	0.014
house surface $40 - 69^2$	-0.0365	(-0.0625;-0.0105)	0.006
wc inside	0.0511	(0.0190;0.0833)	0.002
TV	0.1066	(0.0510;0.1623)	< 0.001
parabolic	0.0768	(0.0473;0.1062)	< 0.001
refrigerator	0.1183	(0.0827;0.1539)	< 0.001
washing machine	0.1140	(0.0843;0.1438)	< 0.001
air conditioning	0.2434	(0.1593;0.3275)	< 0.001
computer	0.2403	(0.1668;0.3138)	< 0.001
car	0.3233	(0.2846;0.3621)	< 0.001
ownership of agricultural land	0.0484	(0.0153;0.0815)	0.004
Random Effects			
σ_γ	0.4096	(0.2700;0.6214)	< 0.001
σ_u	0.1756	(0.1290;0.2389)	< 0.001
σ_e	0.3285	(0.3208;0.3363)	< 0.001

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 (showed in the second map) and once constant in each small area (showed in the third map). From these maps, it is evident the presence of both a spatial dynamic and a district level effect in the Albanian consumption expenditure.

The estimated parameters (presented in Table 1) are then combined with the census mean values as in (11) to obtain the district level estimates of the average household log per-capita consumption expenditure. Due to the unavailability of

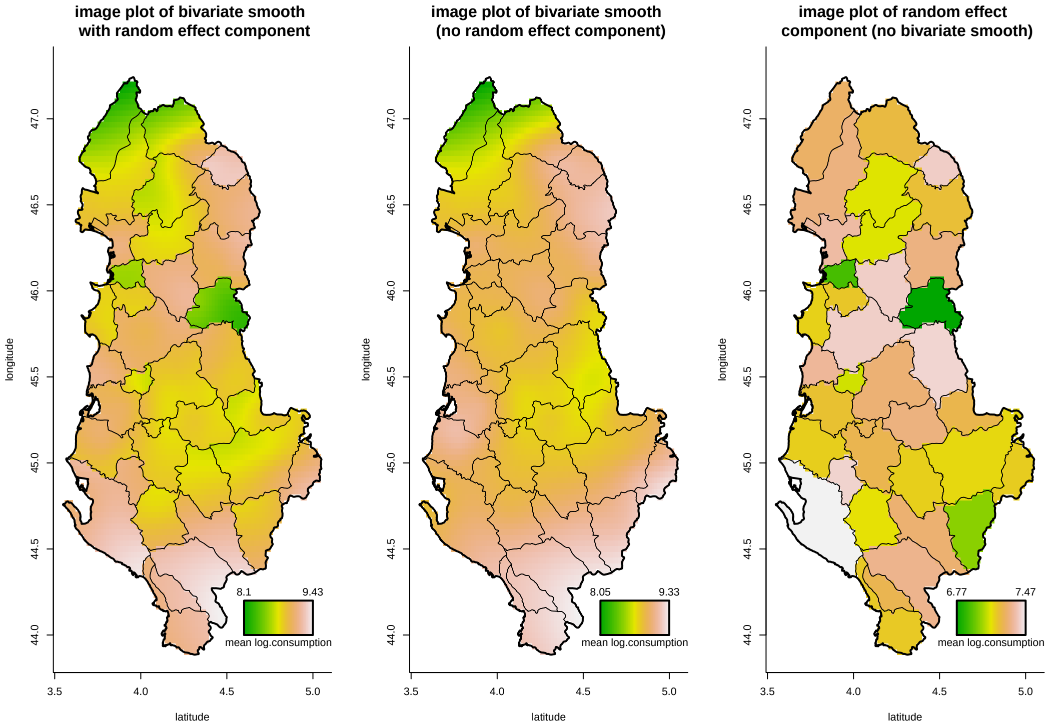
image plot of bivariate smooth with random effect component
image plot of bivariate smooth (no random effect component)
image plot of random effect component (no bivariate smooth)
longitude
47.0
46.5
46.0
45.5
45.0
44.5
44.0
latitude
3.5
4.0
4.5
5.0
8.1
9.43
8.05
9.33
6.77
7.47
mean log.consumption

Figure 2: Spatial smoothing and district random effects of the household log per-capita consumption expenditure. Map (a) shows the smoothing obtained with the geoadditive sae model as sum of two components: the bivariate smoothing, map (b), and the small area random effects, map (c).

the geographical coordinates for the PHC dataset, the true $\overline{\mathbf{z}}_t$ cannot be calculated. We approximate the missing information by using the centroids of each small area to locate all the units belonging to the same area.

The district level estimates are showed in Figure 3 and in Table 2. The mean square errors (MSEs), and consequently the coefficients of variations (CVs), presented in Table 2 are calculated using the robust MSE estimator of Salvati et al. (2010). Further discussion about MSE estimation is presented in Section 4. 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 (see Table 3). In addition, district 22 suffers particularly from the centroid approximation due to its geographical morphology: it is mostly mountainous and the urban area is mainly in the south.

The map presents 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).

Table 2: District level estimates of the mean of household log per-capita consumption expenditure. The root mean squared error (RMSE) and the coefficient of variation (CV%) are obtained with the robust MSE estimator of Salvati et al. (2010).

Code	District Name	Estimate	RMSE	CV%
1	Berat	8.91	0.0472	0.53
2	Bulqize	8.35	0.0514	0.62
3	Delvine	9.46	0.1552	1.64
4	Devoll	9.17	0.1529	1.67
5	Diber	8.96	0.0542	0.60
6	Durres	8.98	0.0601	0.67
7	Elbasan	8.93	0.0368	0.41
8	Fier	9.13	0.0441	0.48
9	Gramsh	8.82	0.0426	0.48
10	Gjirokast	9.52	0.1130	1.19
11	Has	9.15	0.1046	1.14
12	Kavaje	9.22	0.0535	0.58
13	Kolonje	9.05	0.1608	1.78
14	Korce	8.92	0.0630	0.71
15	Kruje	8.91	0.0758	0.85
16	Kucove	8.96	0.0449	0.50
17	Kukes	8.97	0.0753	0.84
18	Kurbin	8.67	0.0549	0.63
19	Lezhe	9.21	0.0773	0.84
20	Librazhd	8.88	0.0450	0.56
21	Lushnje	9.10	0.0576	0.63
22	Malesi e Madhe	8.53	0.1661	1.95
23	Mallakaster	9.11	0.0654	0.72
24	Mat	9.15	0.0969	1.06
25	Mirdite	8.79	0.1049	1.19
26	Peqin	8.74	0.0864	0.99
27	Permet	9.34	0.1365	1.46
28	Pogradec	8.88	0.0626	0.70
29	Puke	8.70	0.1388	1.60
30	Sarande	9.34	0.0809	0.87
31	Skrapar	8.93	0.0999	1.12
32	Shkoder	8.90	0.0640	0.72
33	Tepelene	8.95	0.0871	0.97
34	Tirane	9.23	0.0441	0.48
35	Tropoje	8.78	0.0679	0.77
36	Vlore	9.37	0.0635	0.68

4 MSE Estimation

Along with the definition of the non-parametric SAE model, Opsomer et al. (2008) study the theoretical properties of the mean squared error (MSE) of the small area mean estimator and propose both an analytic and a bootstrap estimator for the MSE quantity. Alternatively, Salvati et al. (2010) propose a robust estimator of the conditional MSE of the same non-parametric SAE model, based on the pseudo-linearization approach to MSE estimation described in Chambers et al. (2007).

We decided to apply both the analytic estimator of Opsomer et al. (2008) and the robust estimator of Salvati et al. (2010) and, in order to evaluate their performance, a desing-based simulation study is implemented.

We build a fixed pseudo-population of $N = 689733$ households by sampling N times with replacement and with probability proportional to the unit sample weights from the LSMS dataset. A total of 500 independent stratified random samples of the same size as the original sample is then selected from this pseudo-population, with districts sample sizes fixed to be the same as in the original sample. For each sample we apply the geoadditive SAE model of the previous section and we calculate the EBLUP (11) for the mean household log per-capita consumption expenditure of each district and the two relative MSE estimates.

The behaviour of the empirical true root MSE and its estimators for each district is shown in Figure 4. It can be seen that there isn't a substantial differ-

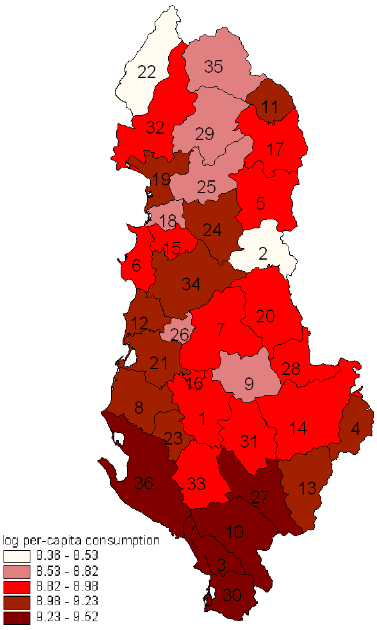

Figure 3: District level estimates of the mean of household log per-capita consumption expenditure.

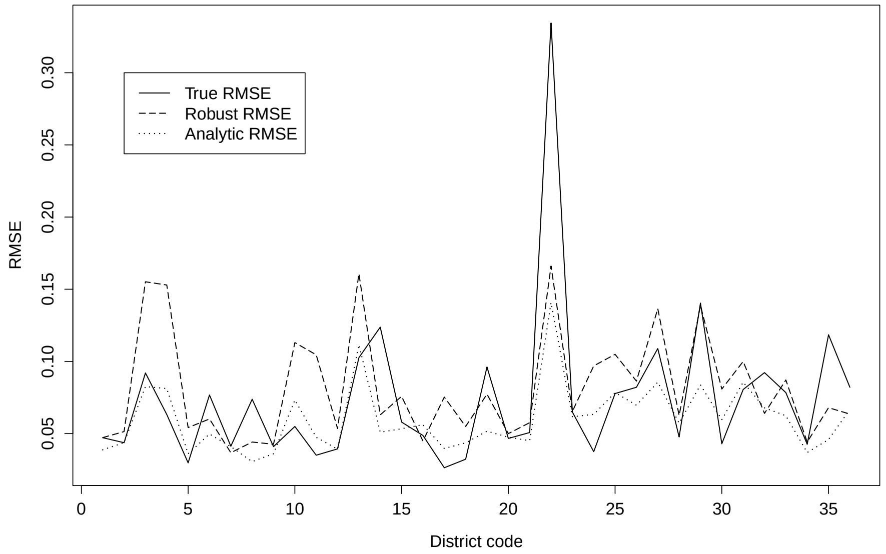

Figure 4: District level of actual design-base RMSE (solid line) and average estimated RMSE. The dashed line indicates the robust estimator of Salvati et al. (2010) and the dotted line indicates the analytic estimator of Opsomer et al. (2008).

ence in the performance of the two estimators, even if the analytic estimator of Opsomer et al. (2008) is always lower that the robust estimator of Salvati et al. (2010). However, the robust estimator seems to better track the irregular profile of the empirical RMSE, while the analytic estimator is slightly over-smoothed. The anomalous value of district 22 is due to the high value of the bias component (see Table 3) and both the estimators undervalue it. Following these considerations, we prefer to present the MSE estimated with the robust estimator of Salvati et al. (2010) (see Table 2).

The simulation study permits also to evaluate the performance of the geoadditive SAE EBLUP. For each district we compute the *Relative Bias* (RB) and the *Relative Root MSE* (RRMSE) defined as

$$RB = \frac{1}{M} \frac{\sum_{m=1}^{M} (\hat{\bar{y}}_{tm} - \bar{y}_t)}{\bar{y}_t}$$

and

$$RRMSE = \frac{\sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\bar{y}}_{tm} - \bar{y}_t)^2}}{\bar{y}_t},$$

where $\bar{y}_t$ denotes the actual district mean t and $\hat{\bar{y}}_{tm}$ is the predicted value at simulation m, $m = 1, ..., M$.

Table 3: Relative bias (RB) and relative RMSE (RRMSE) of the geoadditive SAE EBLUP for the mean household log per-capita consumption expenditure of each district.

Code	n_t	N_t	RB%	RRMSE%
1	120	25422	0.0415	0.5253
2	128	8499	-0.2331	0.5315
3	16	3211	-0.1014	0.9636
4	16	6229	0.4183	0.6849
5	232	16529	-0.2085	0.3340
6	160	42332	-0.6805	0.8689
7	152	47709	-0.2717	0.4713
8	224	45729	0.7459	0.8192
9	120	7538	0.2225	0.4748
10	32	10948	0.0481	0.5735
11	48	3450	-0.0556	0.3822
12	88	18294	-0.0752	0.4343
13	8	2291	0.8891	1.1532
14	136	34914	-1.3420	1.3859
15	39	13477	0.2356	0.6724
16	32	11019	-0.3109	0.5480
17	184	12183	-0.0256	0.3023
18	64	12938	0.0352	0.3769
19	64	13538	0.9217	1.0422
20	200	14345	-0.0589	0.5352
21	152	31953	0.3629	0.5639
22	24	9294	-3.6434	3.8363
23	32	7067	-0.2909	0.7124
24	32	11803	-0.1271	0.4165
25	16	5468	0.1970	0.8839
26	24	8814	-0.4073	0.9564
27	16	5377	0.3386	1.1810
28	48	17418	0.1240	0.5389
29	24	8633	-1.1829	1.6128
30	48	9874	-0.0584	0.4638
31	16	5453	0.4741	0.9140
32	140	43578	-0.9057	1.0329
33	32	11202	0.1509	0.8862
34	684	121020	0.1318	0.4668
35	88	5876	-1.2654	1.3579
36	152	36308	0.6493	0.8853

The values of RB and RRMSE are shown in Table 3: all the values are small and indicate that the geoadditive SAE EBLUP is quite stable. Once again, we note the anomalous value of district 22, that presents a relative bias of -3.64%.

5 Concluding remarks and open questions

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, especially in areas of social and economical interventions. 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.

The empirical evidence suggests that, despite being overlooked in the previous studies, the spatial location is an important component to understand the distribution of the consumption expenditure. In particular, the results of our analysis show that the consumption expenditure presents both spatial dynamics and area specific effects. Thus the region morphology can explain, to some degree, the spatial patterns of the household per-capita consumption expenditure that remain after controlling for all the descriptive household level covariates effect. The map of the estimated district means presents an evident 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), confirming the results of previous applications on the same datasets presented in literature.

Differently from other methods of analysis that exploit some spatial information, the geoadditive SAE model produces not only the map of estimated mean values, but also a spatial interpolation of all the observation. Thus, with this model we can produce an estimated value in any point of the country.

When we produce estimates of a parameter of interest over some pre-specified area, we should always consider the modifiable area unit problem (MAUP). With the geoadditive model we obtain a continuous surface estimation over the entire area, without define the area a priori, thus the MAUP can't occur. In our application the geoadditive model is associated with a SAE model, so in this case we need to define the areas before estimate the model, however the possible MAUP - if occurs - will be only related to the definition of the small area and not to the spatial interpolation of the studied phenomenon.

Finally, the results of the design-based simulation study show that the geoadditive SAE EBLUP for the mean is quite stable, and that the performance of the robust MSE estimator of Salvati et al. (2010) is slightly better than the performance of the analytic MSE estimator of Opsomer et al. (2008). However, the two estimator are quite comparable.

Concluding, we recall that the condition under which the geostatistics methodologies can be applied is the knowledge of the location of all population units at the point level. As find out in our study, this requirement is not so easy to be accomplished, especially if we work with socio-economic data. Usually it is much more easy to know the areas to which the population units belong to (i.e. census districts, blocks, municipalities, enumeration areas, etc.) and the classic approach is to refer the data with respect to the area centroids. An aspect to be explored is the use of a more precise spatial location data: an imputation approach which considers a more realistic hypothesis on spatial distribution. Further investigations will be done in this direction.

Acknowledgements

This research was supported by the Italian Ministry for Education, University and Research (PRIN 2007RHFBB3_004). The author would like to acknowledge Dr. Gero Carletto, Senior Economist in the Development Research Group of the World Bank and member of the Living Standards Measurement Study team, for providing the datasets of the 2002 Living Standard Measurement Study and the 2001 Population and Housing Census of Albania.

References

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

Arbia, G. (2006), *Spatial Econometrics: Statistical Foundation and Applications to Regional Convergence*, Springer, Berlin.

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

Chambers, R., Chandra, H. and Tzavidis, N. (2007), *On robust mean squared error estimation for linear predictors fro domains*, CCSR Working paper 2007-10, Cathie Marsh Centre for Census ans Survey Research, University of Manchester.

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

Grosh, M. and Glewwe, P. (2000), A Guide to Living Standards Measurement Study Surveys, *in* 'Household Accounting: Experience and Concepts in Compilation', Statistics Division. Department of Economic and Social Affairs. The United Nations.

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

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.

Neri, L., Ballini, F. and Betti, G. (2005), 'Poverty and inequality mapping in transition countries', *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', *Journal of the Royal Statistical Society, Series B* **70**, 265–286.

Rao, J. N. K. (2003), *Small area estimation*, John Wiley & Sons, New York.

Richardson, H. W. (1970), *Regional Economics*, MacMillan, London.

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2009), 'Semiparametric regression during 2003–2007', *Electronic Journal of Statistics* **3**, 1193–1256.

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

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008), 'M-quantile models with application to poverty mapping', *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*