Università di Pisa
Dipartimento di Statistica e Matematica
Applicata all'Economia

Report n.303

Estimation of Proportions for Small Areas Using.Unit Level Models With Spatially Correlated population – An Application to Poverty Mapping.

Ray Chambers, Hukum Chandra and Nicola Salvati

Pisa, FEBBRAIO 2008
- Stampato in proprio –

Via Cosimo Ridolfi, 10 - 56124 PISA - Tel. Segr. Amm. 050 2216231 Segr. Stud. 050 2216317 Fax 050 2216375
Cod. Fisc. 80003670504 - P. IVA 00286820501 - Web http://statmat.ec.unipi.it- E-mail: dipstat@ec.unipi.it

Estimation of Proportions for Small Areas Using Unit Level Models with Spatially Correlated Population - An Application to Poverty Mapping

Ray Chambers[1], Hukum Chandra[2] and Nicola Salvati[3]

Abstract

In this article investigates two model-based techniques of small area estimation (SAE) to estimate the small area proportions: the empirical best predictor (EBP) under a generalized linear mixed model (Rao, 2003, chapter 5; Saei and Chambers, 2003) and the model-based direct estimator (MBDE) under a linear mixed model (Chandra and Chambers, 2005). In order to define a unified method of SAE for both discrete and continuous data, we examine an application of linear assumption based MBDE to the binary data and we compare its performance with the EBP via empirical studies using real data. We also evaluate these methods of SAE based on small area models with spatially correlated area effects where the neighbourhood structure is described by a contiguity matrix. Our results show that both the MBDE and the EBP perform well. The EBP is a computation intensive method, in contrast, MBDE is easy to implement. In case of model misspecifications (e.g., data with less variability), the MBDE appears to be more robust. These results further show marginal gains due to spatial dependence between areas.

Key words: Small area proportions, empirical best predictor, generalised linear mixed model, spatial model, model-based direct estimation.

---

[1] Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. Email: ray@uow.edu.au

[2] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: hchandra@soton.ac.uk

[3] Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa, Pisa, Italy, Email: salvati@ec.unipi.it

1. Introduction

The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of statistical methods from both the theoretical and empirical point of view. The conventional estimates for small area quantities based on survey data alone are often unstable because of sample size limitations. In this perspective the model-based methodologies allow for the construction of efficient estimators and their confidence intervals by borrowing the strength through use of a suitable model. These small area estimators have several fields of application: from the production of social data to the production of environmental data. Small area models make use of explicit linking models based on random area-specific effects that take into account for between areas variation beyond that is explained by auxiliary variables included in the model. For continuous response variables, the empirical best linear unbiased predictor (EBLUP) approach under linear mixed model (LMM) is very common and proven to be efficient for small area estimation (SAE), see Rao (2003). However, for discrete response variables a generalized linear mixed model (GLMM), containing fixed and random effects, can be specified (McGilchrst, 1994). There is a growing need for current and reliable count data at small area level. For example, as in many countries, in Italy, this information need concerns Labour Force Survey (LFS) realized by the National Statistical Institute (ISTAT), which has been studied to warrant accuracy only for estimates at regional level. Further, in SAE under both LMM and GLMM, the random area effects are generally assumed to be independent. In practice, it should be more reasonable to assume that the random area effects between the neighbouring areas (for instance the neighbourhood could be defined by a distance criterion) are correlated and the correlation decays to zero as distance increases.

For the continuous response variables, Chandra and Chambers (2005) described the model-based direct estimation (MBDE) methods for SAE and Chandra *et al.* (2007) extended the MBDE for spatially correlated areas. The MBDE is a weighted linear estimator for small areas, defined by using the sample weights derived under population level LMM. Besides the ease of implementation,

the method is robust under model misspecifications. For the discrete response variables, an appropriate indirect model-based estimators for SAE, the empirical best predictor (EBP), is essentially based on GLMM. A major difficulty in use of GLMM for SAE is that the likelihood function often involves high dimensional integrals (computed by integrating a product of discrete and normal densities, which has no analytical solution) which are difficult to evaluate numerically. Although computationally attractive alternatives to the likelihood method are available, they can suffer of inconsistency (Jiang, 1998). In context of SAE besides the parameter estimation, the mean squared error (MSE) estimation for the EBP is another an outstanding problem because the analytical form of the MSE is not suitable to be calculated explicitly (Manteiga *at el.*, 2007), although an approximate MSE of the EBP can be derived under linear approximation (Saei and Chambers, 2003). Moreover, one can use resampling methods but these are computationally intensive. In other words, the MSE estimation for the EBP is not straightforward. An alternative is to ignore the deficiency of the LMM and proceed as if a linear model does hold. These options have the appeal that they are relatively simple and cheap to implement. However, these options sidestep the issues that the LMM is incorrect. Given the robustness of the estimation procedures, they can be expected to produce reasonable results. It is interesting to explore a less expensive (in term of loss of efficiency) and a unified method of SAE applicable to both discrete and continuous response variables.

In this paper we examine the application of MBDE of SAE to the binary respose variable and we compare its performance with the EBP. We also investigate MBDE and EBP under small area models with spatially correlated random area effects where the neighbourhood structure is described by a contiguity matrix. We evaluate the emiprical performance of different SAE methods via simulation studies using real data sets. That is, this paper is planned in two fold. The first aim is to evaluate the efficiency of the MBDE against EBP and the second to see the gains by incorporating the spatial dependence between the areas. Both of these issues have been investigated via empirical studies using real data. Finally, we do apply these methods to real data set from

Albania Living Standards Measurement Study (LSMS) to produce the poverty mapping for the district of Albania.

The rest of the paper is organised as follows. The next section defines the LMM and GLMM, associated small area estimators for the proportions and their respective estimators of mean squared errors. In the section 3 we then report the empirical results and their discussion. Section 4 illustrates the application of various methods using real data collected from the LSMS in Albania. Finally, section 5 concludes the paper with major findings and research prospects that need further attention.

2. The Methodology

In this section we introduce the GLMM and LMM. We then describe the related estimators for small area quantities based on these models and the MSE estimation. In particular, we focus on binary response variable with aim to estimate the population proportions for the variable of interest in small areas and estimates for the MSEs.

2.1 The Empirical Best Predictor for the Small Areas

It is well known that the GLMMs are suitable for the development of indirect estimates for small areas (Rao, 2003) when the response data is non-normal. The indirect estimators for small areas under the GLMM are the EBLUP-type estimators, often known as the empirical best predictors (EBP) for small areas. Let $U = \{1,...., N\}$ denotes the finite population of size N and assumed to partitioned into D non-overlapping sub-groups (or small areas), U_i each of sizes N_i with $i = 1,..., D$ such that $N = \sum_{i=1}^{m} N_i$. Let j and i respectively index the j unit within small area i, y_{ij} is the survey variable of interest (typically a binary variable) and known for sampled units, $\boldsymbol{x}_{ij}$ is the vector of auxiliary variables (including the intercept), known for the whole population. Let s_i and r_i respectively denotes the sample (of size n_i) and non-sample (of size $N_i - n_i$) in small area i. The objective is to make inference about the small area i population proportions, $p_i = N_i^{-1} \sum_{j \in U_i} y_j$

$= N_i^{-1}\left\{\sum_{j\in s_i} y_j + \sum_{j\in \bar{s}_i} y_j\right\}$. Let π_{ij} be the probability that a unit j in area i assumes value 1. Let u_i denotes is the random area effect for the small area i and assumed to be normally distributed with mean zero and variance φ. We assume that u_i's are independent and $y_{ij} \mid u_i \sim Bin(1,\pi_{ij})$ with $E(y_{ij} \mid u_i) = \mu_{ij} = \pi_{ij}$ and $Var(y_{ij} \mid u_i) = \sigma_{ij} = \pi_{ij}(1-\pi_{ij})$. A popular model for this type of data is the linear logistic mixed model of the form

$$\log it(\pi_{ij}) = \log\left\{\pi_{ij}/(1-\pi_{ij})\right\} = \eta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i, j = 1,...,N_i; i = 1,...,D \qquad (1)$$

where $\boldsymbol{\beta}$ $(p\times 1)$ is the vector of regression parameters. In small area literature for the estimation of unknown parameters, it is common practice to express the model (1) at the population level as below (Rao, 2003, chapter 6).

Let $\mathbf{y}_U$ be the $N\times 1$ vector of response variable with elements y_{ij}, $\mathbf{X}_U$ be the $N\times p$ known design matrix with rows x_{ij}, $\mathbf{G}_U = diag(\mathbf{1}_{N_i}; 1\le i\le D)$ is the known matrix of order $N\times q$, $\mathbf{1}_k$ is a column vector of ones of size k, $\mathbf{u} = (u_1,...,u_D)'$ and $\boldsymbol{\varsigma}_U$ denotes the $N\times 1$ vector of linear predictors η_{ij} given by (1). We define $\boldsymbol{\mu} = E(\mathbf{y}_U \mid \boldsymbol{u})$ the conditional mean function of the response vector $\mathbf{y}_U$ given $\mathbf{u}$ with elements μ_{ij} and $Var(\mathbf{y}_U \mid \boldsymbol{u}) = diag\{\sigma_{ij}\}$ the conditional covariance matrix. Let $g(.)$ be a monotonic function, the link (McCullagh and Nelder, 1989, page 27), such that $g(\boldsymbol{\mu})$ can be express as the linear model of form

$$g(\boldsymbol{\mu}) = \boldsymbol{\varsigma}_U = \mathbf{X}_U\boldsymbol{\beta} + \mathbf{G}_U\mathbf{u}. \qquad (2)$$

The equation (2) defines a GLMM, if $\mathbf{y}_U$ given $\boldsymbol{\mu}$ are independent and belong to the exponential family of distribution. The vector of random area effects $\mathbf{u}$ has mean $\mathbf{0}$ and variance $\Omega(\delta) = \varphi\mathbf{I}_D$, where $\mathbf{I}_D$ is the identity matrix of order D. For binomial response variable the link function $g(.)$ is a logit function, see equation (1). The relationship among $\mathbf{y}_U$ and $\boldsymbol{\eta}_U$ is represented through a known function $h()$, defined by $E(\mathbf{y}_U \mid \mathbf{u}) = h(\boldsymbol{\eta}_U)$. Suppose that our interest is to predict linear

parameters for small areas $\boldsymbol{\theta} = \mathbf{a}_U \mathbf{y}_U$, where $\mathbf{a}_U = diag\{\mathbf{a}_i', i = 1,..,D\}$ is a $D \times N$ matrix and $\mathbf{a}_i' = (a_{i1},...,a_{iN_i})$ is a vector of known elements. For estimation of the population proportion for small area i, $\mathbf{a}_i'$ denote the population vector with value N_i^{-1} for each population unit in area i and zero elsewhere.

Without loss of generality, we arrange the vector $\mathbf{y}_U$ so that its first n elements correspond to the sample units, and then partition $\mathbf{a}_U$, $\mathbf{y}_U$, $\boldsymbol{\eta}_U$, $\mathbf{X}_U$ and $\mathbf{G}_U$ according to sample and non-sample units as

$$\mathbf{a}_U = \begin{bmatrix} \mathbf{a}_s \\ \mathbf{a}_r \end{bmatrix}, \ \mathbf{y}_U = \begin{bmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{bmatrix}, \ \boldsymbol{\eta}_U = \begin{bmatrix} \boldsymbol{\eta}_s \\ \boldsymbol{\eta}_r \end{bmatrix}, \ \mathbf{X}_U = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \text{ and } \mathbf{G}_U = \begin{bmatrix} \mathbf{G}_s \\ \mathbf{G}_r \end{bmatrix}.$$

Here a subscript of s denotes components defined by the n sample units while a subscript of r is used to denote corresponding components defined by the remaining $N-n$ non-sample units. We then write $E(\mathbf{y}_s \mid \mathbf{u}) = h(\boldsymbol{\eta}_s)$ and $E(\mathbf{y}_r \mid \mathbf{u}) = h(\boldsymbol{\eta}_r)$. Typically, $h()$ is obtained as $g^{-1}()$. The parameter of interest $\boldsymbol{\theta} = \mathbf{a}_U \mathbf{y}_U$ can be expressed as

$$\boldsymbol{\theta} = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r \mathbf{y}_r = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r \boldsymbol{\beta} + \mathbf{G}_r \mathbf{u}) . \quad (3)$$

Here $\mathbf{y}_s$ the vector of sample values is known, whereas the second term of (3), which depends on the non-samples values $\mathbf{y}_r = h(\mathbf{X}_r \boldsymbol{\beta} + \mathbf{G}_r \mathbf{u})$, is unknown and can be predicted by fitting the model (3) for sample data. In this paper, $\mathbf{y}_s = \{y_{sij}\}$ denotes the vector of sample values of the binary survey variable y, where $y = 1$ e.g. if the consumption expenditure per household is less than a poverty line, 0 otherwise. Similarly, $\mathbf{y}_r = \{y_{rij}\}$ represents the vector of non-samples values of the survey variable. It is obvious that the parameter of interest p_i for each small area can be obtained by using as prediction of each element $\{y_{rij}\}$.

For known $\boldsymbol{\Omega}(\boldsymbol{\delta})$, the values of $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated by Penalized Quasi-Likelihood (PQL) under model (3) fitted for sample data (Breslow and Clayton, 1993). This gives the best linear unbiased estimate (BLUE) for $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) for $\mathbf{u}$. Hence, using

(3) we obtain the BLUP-type estimator of θ. In practice $\Omega(\delta)$ is unknown and the vector of variance components δ is estimated from the sample data. Using estimated value $\hat{\delta}$ of the δ leads to the empirical BLUE $\hat{\beta}$ for β and the empirical BLUP $\hat{u}$ for u and thus the empirical BLUP type estimator of θ is given by

$$\hat{\theta} = a_s y_s + a_r h(X_r \hat{\beta} + G_r \hat{u}) \tag{4}$$

The GLMM involves the likelihood function which does not have close form analytical expression. Several approximations to the likelihood function and approximately maximum likelihood estimators have been proposed in the literature. The PQL approach is most popular estimation procedure for the GLMM and it constructs a linear approximation of the distribution of non-normal response variable Y and assumes the linearised dependent variable is approximately normal. This approach is reliably convergent but it has been noticed that the PQL tends to underestimate variance components as well as fixed effect coefficients (Breslow and Clayton, 1993; Jang and Lim, 2006). McGilchrst (1994) introduced the idea to use BLUP to obtain approximate restricted maximum likelihood (REML) estimates for GLMMs. This link between BLUP and REML is described in Harville (1977) for the normal case. An iterative procedure that combines the PQL estimation of β and u with REML estimation of $\Omega(\delta)$ is described in Saei and Chambers (2003). In our empirical results reported in section 3, we adopted their algorithm for parameters estimation.

Turning now to estimation of mean squared error (MSE) of the EBLUP-type predictor (4) we define $H_r = H(\hat{\eta}_r) = \partial h(\eta_r)/\partial \eta_r \big|_{\eta_r = \hat{\eta}_r}$ and $\hat{B}_s = \partial^2 l_1 / \partial \eta_s \partial \eta_s' \big|_{\eta_s = \hat{\eta}_s}$, the matrix of second derivatives of l_1 (the log-likelihood function l_1 defined by the vector y_s given u) with respect to η_s at $\eta_s = \hat{\eta}_s$. Similarly, $\hat{B}_r = \partial^2 l_1 / \partial \eta_r \partial \eta_r' \big|_{\eta_r = \hat{\eta}_r}$. We put $X_r^* = a_r H_r X_r$ and $G_r^* = a_r H_r G_r$. Then an approximate estimate of the mean squared error for the EBLUP-type estimator (4) (see Saei and Chambers, 2003; Manteiga *et al.*, 2007) is

$$mse(\hat{\theta}) = m_1(\hat{\delta}) + m_2(\hat{\delta}) + 2m_3(\hat{\delta}) + m_4(\hat{\delta}) \tag{5}$$

where

$m_1(\hat{\boldsymbol{\delta}}) = \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s^* \boldsymbol{G}_r^{*\prime}$ with $\hat{\boldsymbol{T}}_s^* = (\hat{\mathbf{U}}^{-1} + \boldsymbol{G}_s' \hat{\boldsymbol{B}}_s \boldsymbol{G}_s)^{-1}$,

$m_2(\hat{\boldsymbol{\delta}}) = C_r \left(\boldsymbol{X}_s' \hat{\boldsymbol{B}}_s \boldsymbol{X}_s - \boldsymbol{X}_s' \hat{\boldsymbol{B}}_s \boldsymbol{G}_s \hat{\boldsymbol{T}}_s^* \boldsymbol{G}_s' \hat{\boldsymbol{B}}_s \boldsymbol{X}_s \right)^{-1} C_r'$ with $C_r = \left\{ \boldsymbol{X}_r^* - \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s^* \boldsymbol{G}_s' \hat{\boldsymbol{B}}_s \boldsymbol{X}_s \right\}$,

$m_3(\hat{\boldsymbol{\delta}}) = \left\{ tr(\hat{\nabla}_t \hat{\Sigma}_s^* \hat{\nabla}_k') v(\hat{\boldsymbol{\delta}}) \right\}$ and $m_4(\hat{\boldsymbol{\delta}}) = \mathbf{a}_r \hat{\boldsymbol{B}}_r \mathbf{a}_n'$. Let $\varsigma = \boldsymbol{G}_r^* \hat{\boldsymbol{T}}_s^*$ and $\boldsymbol{G}_{ij}^*$ be the j^{th} row of the matrix $\boldsymbol{G}_r^*$ and put $\hat{\nabla}_j = \partial(\varsigma_j)/\partial\delta\big|_{\delta=\hat{\delta}} = \partial(\boldsymbol{G}_{ij}^* \hat{\boldsymbol{T}}_s^*)/\partial\delta\big|_{\delta=\hat{\delta}}$ and $v(\hat{\boldsymbol{\delta}})$ is the asymptotic covariance matrix of estimates of variance components $\hat{\boldsymbol{\delta}}$ which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\boldsymbol{\delta}}$. This depends upon whether we are using ML or REML estimates $\hat{\boldsymbol{\delta}}$. In this paper we used REML estimates for $\hat{\boldsymbol{\delta}}$. See Saei and Chambers (2003) for these expressions for both ML and REML estimates for $\hat{\boldsymbol{\delta}}$.

As described earlier, for the estimation of small area proportion, $\mathbf{a}_i'$ is the population vector with value N_i^{-1} for each population unit in area i and zero everywhere else. In particular, using (4) the empirical best predictor (EBP) for the small area i population proportion p_i is

$$\hat{p}_{i,EBP} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\mu}_{ij} \right\} \tag{6}$$

where $\hat{\mu}_{ij} = \exp(\hat{\eta}_{ij}) \{1 + \exp(\hat{\eta}_{ij})\}^{-1} = \hat{\pi}_{ij}$ and $\hat{\eta}_{ij} = \mathbf{x}_{ij} \hat{\boldsymbol{\beta}} + \hat{u}_i$. Likewise, we obtain the MSE estimates for the EBP $\hat{p}_{i,EBP}$ from (5).

In many situations the physical location of the small areas is so relevant that the assumption of spatial independence of the small area models (as we considered earlier in this section) becomes questionable. That is, small area data exhibit a spatial structure and therefore use of spatial models becomes essential. Spatial dependency is the extent to which the value of an attribute in one location depends on the value of the attribute in nearby locations or small areas. Recently the problem has been addressed by introducing a common autocorrelation parameter among small areas extending the linear mixed model through the Simultaneously Autoregressive (SAR) process (Pratesi and Salvati 2007; Singh *et al.*, 2005; Petrucci and Salvati, 2006, Chandra *et al.*, 2007). The

focus here is on the introduction of the SAR process in the GLMMs where the vector of random area effects $\mathbf{v} = (v_i)$ satisfies

$$\mathbf{v} = \rho \mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_D - \rho \mathbf{W})^{-1}\mathbf{u} \tag{7}$$

where ρ is spatial autoregressive coefficient which determines the degree of spatial dependency of the model, $\mathbf{W}$ is proximity or contiguous matrix of order D. This matrix is symmetric and encapsulates the relative spatial arrangement (i.e. neighbourhood structure) of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. The simplest way to define such a matrix is as simple contiguity: the elements of $\mathbf{W}$ take non-zero values only for those pairs of areas that are contiguous to each other. Generally, for ease interpretation, the general spatial weight matrix is defined in row-standardized form; in this case ρ is called spatial autocorrelation parameter (Banerjee *et al.*, 2004). Here $E(\mathbf{u}) = \mathbf{0}$ and $Var(\mathbf{u}) = \varphi \mathbf{I}_D$ so $E(\mathbf{v}) = \mathbf{0}$ and $Var(\mathbf{v}) = \mathbf{\Omega}(\varphi, \rho) = \varphi \left[\left(\mathbf{I}_D - \rho \mathbf{W} \right) \left(\mathbf{I}_D - \rho \mathbf{W}^T \right) \right]^{-1}$, where $\mathbf{\Omega}(\varphi, \rho) = \mathbf{\Omega}(\boldsymbol{\delta})$ is the SAR dispersion matrix. Now to define the EBLUP-type estimator (or the EBP) under spatially correlated area effects, the linear predictor η_U is expressed as

$$\eta_U = \mathbf{X}_U \boldsymbol{\beta} + G_U \mathbf{v} \tag{8}$$

where the vector $\mathbf{v}$ is an D-vector of spatially correlated area effects that satisfies SAR model (7). The spatial EBLUP-type estimator or spatial empirical best predictor (denoted by SEBP) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r h(\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathbf{G}_r \hat{\mathbf{v}}). \tag{9}$$

Following (5) we can write the MSE of the SEBP using the variance components $\boldsymbol{\delta}=(\varphi, \rho)$ and $\hat{\mathbf{v}}$ in place of $\hat{\mathbf{u}}$.

2.2 The Model Based Direct Estimator for the Small Areas

The model-based direct (MBD) approach for SAE, investigated in Chandra and Chambers (2005) is effectively linear estimators and implicitly assumes that the variables of interest follow a LMM.

The empirical studies show in case of model misspecification, MBD estimation provides a robust set of small area estimates. Application of the MBD method of SAE to non-normal data uses a set of estimation procedure based on LMM and manipulates the data to make it fit a LMM. Following the notation of Chandra *et al.* (2007) a brief description the MBD method is given here. Let us assume that population values follow the linear mixed model

$$\mathbf{y}_U = \mathbf{X}_U\boldsymbol{\beta} + \boldsymbol{G}_U\mathbf{u} + \mathbf{e}_U . \tag{10}$$

Where $\mathbf{y}_U = (\mathbf{y}_1', \ldots, \mathbf{y}_D')'$, $\mathbf{X}_U = (\mathbf{X}_1', \ldots, \mathbf{X}_D')'$, $\boldsymbol{G}_U = diag(\boldsymbol{G}_i = \mathbf{1}_{N_i}; 1 \le i \le D)$, $\mathbf{u} = (u_1, \ldots, u_D)'$ and $\mathbf{e}_U = (\mathbf{e}_1, \ldots, \mathbf{e}_D)'$ partitioned to area components. The spatial independence between small areas indicates the covariance matrix of $\mathbf{y}_U$ has block diagonal structure, $\mathbf{V}_U = diag(\mathbf{V}_i; 1 \le i \le D)$ with $\mathbf{V}_i = \varphi\mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2\mathbf{I}_{N_i}$. In practice the variance components that define $\mathbf{V}_U$ are unknown and can be estimated from the sample data using methods described, for example, in Harville (1977). We denote these estimates by $\hat{\boldsymbol{\delta}} = (\hat{\varphi}, \hat{\sigma}_e^2)'$ and put a 'hat' on any quantity where these estimates are substituted for actual values, e.g. $\hat{\mathbf{V}}_U = diag(\hat{\mathbf{V}}_i; 1 \le i \le D)$ and $\hat{\mathbf{V}}_i = \hat{\varphi}\mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \hat{\sigma}_e^2\mathbf{I}_{N_i}$. Similar to as below (2) we again consider the decomposition of different terms into sample and non-sample components and from Royall (1976), under the population level linear mixed model (10), the sample weights that define the EBLUP for the population total of y are

$$\mathbf{w}_{EBLUP} = (w_{j,EBLUP}) = \mathbf{1}_s + \hat{\mathbf{H}}'\left(\mathbf{X}_U\mathbf{1}_N - \mathbf{X}_s'\mathbf{1}_s\right) + \left(\mathbf{I}_s - \hat{\mathbf{H}}'\mathbf{X}_s'\right)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r . \tag{11}$$

where $\hat{\mathbf{H}} = \left(\sum_i \mathbf{X}_{is}'\hat{\mathbf{V}}_{iss}^{-1}\mathbf{X}_{is}\right)^{-1}\left(\sum_i \mathbf{X}_{is}'\hat{\mathbf{V}}_{iss}^{-1}\right)$. The MBD estimator of the proportion for small area i (Chandra *et al.*, 2007) is then defined as

$$\hat{p}_{i,MBD} = \sum_{j\in s_i} w_{j,EBLUP}\, y_j \Big/ \sum_{j\in s_i} w_{j,EBLUP} . \tag{12}$$

A robust estimator (Chandra *et al.*, 2007; Royall and Cumberland, 1978) of the mean squared error of the MBDE (12) is

$$M(\hat{p}_{i,MBD}) = v(\hat{p}_{i,MBD}) + \left\{b(\hat{p}_{i,MBD})\right\}^2 . \tag{13}$$

where $v(\hat{p}_{i,MBD}) = N_i^{-2}\sum_{j\in s_i}\left\{a_j^2 + (N_i - n_i)n^{-1}\right\}\hat{\lambda}_j^{-1}(y_j - \hat{\mu}_j)^2$ with $a_j = \left(\sum_{s_i} w_k\right)^{-1}\left(N_i w_j - \sum_{s_i} w_k\right)$ is the estimate of the prediction variance of the MBDE (12), and $b(\hat{p}_{i,MBD}) = \left(\sum_{k\in s_i} w_{ik}\right)^{-1}\sum_{j\in s_i} w_{ij}\hat{\mu}_j - N_i^{-1}\sum_{j\in U}\hat{\mu}_j$ is the estimate of its prediction bias. Here $\hat{\mu}_j = \sum_{k\in s}\phi_{kj}y_k$ are unbiased estimators of the area specific individual expected values $\mu_j = E(y_j \mid x_j); j\in s_i$ and $\hat{\lambda}_j = (1-\phi_{jj})^2 + \sum_{k\in s(-j)}\phi_{kj}^2 \approx 1$. Under model (10), $\hat{\mu}_j = \mathbf{x}_j\hat{\beta} + G_j\hat{u}_i; j\in s_i$ is the estimators of the μ_j. The MSE estimator (13) is called a robust model-based estimator because it does not depend on the second order moments assumptions and thus robust to misspecification of the second order moment of the working model. See Chambers et al (2007).

In order to take into account the correlation between neighbouring areas we consider the use of spatial models for random area effects similar to we described for the GLMM in section 2.1. That is a SAR error process for the vector of random area effects. Then the underlying LMM is

$$\mathbf{y}_U = \mathbf{X}_U\boldsymbol{\beta} + \boldsymbol{G}_U\mathbf{v} + \mathbf{e}_U \tag{14}$$

where the vector $\mathbf{v}$ satisfy the SAR model (7).

The covariance matrix of $\mathbf{y}_U$ is $Var(\mathbf{y}_U) = \mathbf{V}_U = \sigma_e^2\mathbf{I}_N + \boldsymbol{G}_U\boldsymbol{\Omega}(\boldsymbol{\delta})\boldsymbol{G}_U'$ with $\boldsymbol{\delta} = (\varphi, \sigma_e^2, \rho)'$ and $\boldsymbol{\Omega}(\boldsymbol{\delta}) = \varphi[(\mathbf{I}_D - \rho\mathbf{W})(\mathbf{I}_D - \rho\mathbf{W}')]^{-1}$. Usually the vector of parameters $\boldsymbol{\delta} = (\varphi, \sigma_e^2, \rho)'$ is unknown and replaced by an asymptotically consistent estimator $\hat{\boldsymbol{\delta}} = (\hat{\varphi}, \hat{\sigma}_e^2, \hat{\rho})'$. When all random effects are normally distributed, the parameter vector $\boldsymbol{\delta}$ can be estimated via ML as well as REML (Pratesi and Salvati, 2007; Chandra *et al.*, 2007). Numerical approximations to either the ML or REML estimators $\hat{\varphi}, \hat{\sigma}_e^2$ and $\hat{\rho}$ can be obtained via a two-step procedure. At the first step, the Nelder-Mead algorithm (Nelder and Mead, 1965) is used to approximate these estimates. The second step then uses these approximations as starting values for a Fisher scoring algorithm. This is necessary because the log-likelihood function has multiple local maxima (Pratesi and Salvati, 2007). In

empirical studies, reported in Section 3, we carried out parameter estimation via REML using the *lme* function in the **R** environment.

For the MBD estimation under (14) we note that the EBLUP sample weights (11) depend on the structure of the random area effects in the LMM (10) only via the their sample and population covariance structure. Consequently, extension to more complex covariance structures requires only that $\hat{\mathbf{V}}_{ss}^{-1}$ and $\hat{\mathbf{V}}_{sr}$ be recomputed under these more complex models. When (14) holds, the corresponding spatial EBLUP weights $\mathbf{w}_{SEBLUP} = (w_{j,SEBLUP})$ are therefore still given by (11), but where now the variance-covariance matrix are $\hat{\mathbf{V}}_{ss}^{-1} = \left\{\hat{\sigma}_e^2 \mathbf{I}_s + \boldsymbol{G}_s \hat{\varphi}[(\mathbf{I}_D - \hat{\rho}\mathbf{W})(\mathbf{I}_D - \hat{\rho}\mathbf{W}')]^{-1}\boldsymbol{G}_s'\right\}^{-1}$ and $\hat{\mathbf{V}}_{sr} = \hat{\varphi}\boldsymbol{G}_s[(\mathbf{I}_D - \hat{\rho}\mathbf{W})(\mathbf{I}_D - \hat{\rho}\mathbf{W}')]^{-1}\boldsymbol{G}_r'$. The spatial-MBDE for small area i proportion p_i is $\hat{p}_{i,SMBDE}$ (denoted by SMBDE) and the corresponding estimator of its mean squared error are then given by (12) and (13) respectively, with the weights (11) used there replaced by the spatial EBLUP weights w_{SEBLUP} defined above.

3. Simulation Studies

In this section we present simulation studies to illustrate the performance of the four methods of SAE discussed in the previous section. These are described as below:

(i) the empirical best predictor under the GLMM (2) with spatially independent area effects, denoted by EBP,

(ii) the MBD estimation under the LMM (10) with spatially independent area effects, denoted by MBDE,

(iii) the empirical best predictor under the GLMM (8) with spatially dependent area effects, denoted by SEBP and

(iv) the MBD estimation under the LMM (14) with spatially dependent area effects, denoted by SMBDE.

We carried out design-based simulation studies using two real data sets. This evaluates the performance of these methods in the context of real population and realistic sampling methods. The two data sets used in the design-based simulations are:

i) *The ISTAT farm structure survey.* This is a sample of 529 farms from the farm structure survey in Tuscany (Italy) carried out by ISTAT. Here we used these sample farms to generate a population of $N = 22977$ farms by sampling with replacement from the original sample of 529 farms with probabilities proportional to their sample weights. We drew 1000 independent stratified random samples from this (fixed) population, with total sample size in each draw equal to the original sample size (529) and with the small areas of interest defined by the 23 Local Economy Systems (LESs) of the North Tuscany region. The small areas sample sizes varied from 4 to 48 and were fixed to be the same as in the original sample. The response variable y takes value 1 if olive production (quintals) of a farm is less than median production and 0 otherwise. Our aim is to estimate the proportion of farms with olive production below median in each LES using utilized olive surface (hectares) as the auxiliary variable. The results from this simulation are set out in Tables 1 and 2.

ii) *The Environmental Monitoring and Assessment Program (EMAP) survey.* The sample of 349 plots in the lakes from the North-eastern states of the U.S. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited, in different plots, several times during the study period (1991-1995). The total number of measurements is 551. The 349 plot are the result of their grouping by lake and by 6-digit Hydrologic Unit Codes (HUC). Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University supplied this data set, developed by EMAP. The HUCs are considered as regions of interest. In three areas sample sizes were only 1's. Therefore we decided to combine these regions with their similar regions. Consequently, we left with 23 small areas. Sample sizes in these 23 areas vary from 2 to 45. We generated a population of size $N = 21028$ by sampling N times with replacement from the above sample of 349 plots

(units) and with probability proportional to a unit's sample weight; and then 1000 independently stratified random samples of the same size as the original sample were selected from this (*fixed*) simulated population. HUC sample sizes were also fixed to be the same as in the original sample. The variable of interest y takes value 1 if Acid Neutralizing Capacity (ANC)-an indicator of the acidification risk of water bodies- in water resource surveys is less than 500 and 0 otherwise. The elevation of the lake is the auxiliary variable. We are interested in estimation of small area proportion of plots for which ANC less than 500. Results from this simulation experiment are set out in Table 3 and 4.

We computed three measures to compare the performance of the different estimators: the relative bias (RB) and the relative root mean squared error (RRMSE), both expressed as percentages, of estimates of the small area proportions and the coverage rate of nominal 95 per cent confidence intervals for these proportions. In the evaluation of coverage performances intervals are defined by the estimate of small area proportion plus or minus twice their standard error (Chandra and Chambers, 2005)..

In Table 1 we reported the relative bias and relative root mean squared error for small area proportions estimated using four different methods of small area estimation (EBP, MBDE, SEBP and SMBDE) based on repeated sampling from the simulated Northern Tuscany population. Corresponding coverage rates for nominal 95% intervals for small area proportions, true and estimated values of small area proportions generated by these methods are shown in Table 2. Analogous results for repeated sampling from the simulated EMAP population are presented in Tables 3 and 4.

In Table 1 the unstable performance of both EBP and SEBP in region 5 and 6 are noteworthy. These unstable results are due mainly to there being little or no variability in the data in these two regions. In contrast, the MBDE and SMBDE methods appear unaffected by such behaviour. Further in these cases both EBP and SEBP produces over estimates for small area proportions (Table 2). Although results are not presented here the empirical best linear unbiased predictor (EBLUP,

Prasad and Rao, 1990) under linear assumption is worst in these cases and leads to unexpected negative or greater than 1 estimates of small area proportions. Furthermore with same magnitude of average relative root mean squared error of EBP (and SEBP) and MBDE (and SMBDE), the average relative bias of MBDE (and SMBDE) is smaller than that of EBP (and SEBP). In regional estimation there is nothing to choose in terms of relative biases, however in terms of relative RMSE it seems advantageous to include spatial effects in EBP, with a marginal gain. Moreover, no significant difference in performance of MBDE is noticed due to spatial effects. The average relative RMSE of SEBP and SMBDE are marginally smaller than EBP and MBDE, respectively. Although on average estimate of area proportions is equally good for all methods, however, average coverage rates are over estimated if spatial effects are ignored in small area models (Table 2), which again show an advantage of including spatial structure.

[Table 1 about here.]

[Table 2 about here.]

In Table 3 we noticed that results for regions 5 and 9 are missing. In these regions true small area proportions (i.e. small area proportions for population) is zero (see Table 4). Consequently, we could not calculate the performance measures (i.e. relative bias and relative root mean square error) since these terms contain zero in their denominator. The average results in Table 3 are based on the average of remaining 21 areas. In terms of relative biases and relative RMSEs the conclusions from Table 3 are identical to results of Northern Tuscany population reported in Table 1.

[Table 3 about here.]

Further, in Table 4 we observed an over coverage rates for few regions. It seems clear that the MSE for the MBDE/SMBE is being significantly overestimated. This is particularly puzzling for regions 1, 2, 3,4,5,6,9,16 and 17. A critical examination of results revealed that in these regions true small area population proportion is either 1 (regions 1, 2, 3, 4,6,16 and 17) or 0 (regions 5 and 9). In addition, in these regions the estimated small area proportions via MBDE/SMBE are same as true values for all simulation runs so true MSEs turn out to be zero. However, the estimates of MSEs are

not zero. This resulted in an overestimation of MBDE/SMBE mean squared errors. Although true MSE is not exactly zero for the EBP/SEBP methods since they are indirect estimator but similar problem exist with MSE estimation these methods as well for such regions.

[Table 4 about here.]

Overall gain by incorporating spatial effects in small models for binary variable is marginal. However, an interesting point to note here is an application of linear assumption based MBD approach of SAE for the binary survey variable. In this context an obvious estimator is indirect method of SAE based on GLMM. That is the empirical best predictor method described in section 2. Empirical results based on two real populations clearly show MBD method performs well when applied to the binary variable and there is no significant efficiency loss. In general, when model holds correctly (i.e. under normal circumstance of the data) indirect method based GLMM are slightly more efficient than the MBD (Table 1 and 3). However, when data have less variability or less suitable for modelling MBDE provides more robust small area estimates. Note that linear assumption based indirect method (i.e. EBLUP) is not suitable in this case. The MBDE approach has ease of implementation. In contrast, the EBP is a computation intensive method and based on approximation methods for parameters estimation.

4. Application to Poverty Mapping for Albania

Poverty alleviation programs have constituted, in the recent years, an important issue in the policy agenda of most developing countries. This is the reason why poverty analysis has become a widely spread tool to support policy applications. Among poverty analysis techniques, poverty mapping is considered one of the most efficient methods that permit sufficient disaggregation of a poverty measure to small geographical units. In other words, it highlights poor areas that had previously been unnoticed or had been considered poor but with little support evidence. Poverty mapping has different uses ranging from targeting anti-poverty programs, to provide visual information on spatial distribution of poverty, consequently there are different methods that permit to consider the

spatial dimension in the poverty analysis. In particular, poverty maps based on sophisticated small area models, constitutes a relevant improvement in the ability to target the poor. This methodology starts estimating poverty measures (such as the Head Count Index) by the use of national representative household surveys. Then it goes one step further, with the integration of other sources than the household survey and the application of small area models, poverty mapping is able to produce a high resolution spatial distribution of poverty that would be not possible with the household survey alone. A new line of attack to deal with the poverty mapping is the inclusion in the process of estimation of information regarding the distance (geographical, spatial, etc.) among the domains. It seems logical to assume, for example, that contiguous areas are more similar than far away areas (Petrucci *et al.*, 2003).

In previous section we investigated the empirical performances of the two approach of SAE (MBDE and EBP) under the design-based simulation studies using two real data sets. We considered both spatial dependence and independence between the areas. Our results show both approach of SAE is performing well and only a marginal gain due to spatial structure. What follows next, we apply the MBDE and EBP methods of SAE without the spatial effects to estimate the proportion of households for which the per-capita consumption expenditure falls below a minimum level (poverty line) necessary to meet the basic food and non-food needs by district in Albania: the Head Count Ratio (HCR). The poverty line that we used to obtain the poverty estimates in Albania is set equal to 4,891 Leks per month (World Bank, 2003). There are twelve prefectures in Albania with a prefecture consisting of several districts. There are thirty-six districts in total (Figure 1).

[Figure 1 about here.]

The dataset used in the following analysis is constituted of information taken from two sources: the Living Standards Measurement Study (LSMS) conducted in Albania in 2002, and the Census data 2002. The survey provides valuable information on a variety of issues related to living conditions of the people in Albania. An attempt to obtain direct estimates of household per-capita consumption expenditure at district level reveals the lack of precision (increased variance) of the

direct estimates particularly for districts with small sample sizes. The selection of covariates to fit the small area models relies on prior studies of poverty assessment in Albania (Betti, 2003; Tzavidis *et al.*, 2007). We have selected the following household level variables: the household size, the presence of facilities in the dwelling (TV, parabolic dish antenna, refrigerator, air conditioning, personal computer), ownership of dwelling, ownership of land and ownership of car.

The results are shown in Table 5 and maps for the HCR at district level are reported in Figure 2. The maps are particularly informative: the location and the concentration of poor become immediately obvious from an examination of the Table 5 and the associated maps 1-2. The district of Bulqize (poverty head count ratio of 65% by EBP and 67% by MBD), Kurbin (44% and 48%) and Peqin (35% and 42%) are the poorest. The district of Vlore is the better off in terms of percentage of households below the poverty line (5%) by using the EBP, whereas from the MBD, the less poor district is Gjirokaster (0%). These districts are followed by Sarande (7% and 4%). We can note that the standard errors associated with the EBP are slightly less than those obtained for MBD estimators. According to the HCR the districts in the mountain region of Albania (north and north-east) are the worse off.

[Table 5 about here.]

[Figure 2 about here.]

We point out that our results are not directly comparable with results obtained from the World Bank by using the approach of Elbers, Lanjouw and Lanjouw (2003 - ELL method), because we didn't have access to the complete database employed for producing poverty estimates with the ELL method. Anyway, the results appear to be consistent with World Bank experts' opinions and with results obtained by applying the ELL method.

5. Concluding Remarks

In this paper we present an empirical investigation of use of spatial model in SAE for binary survey variable when normality assumption is not valid. We have shown the application of GLMM based

empirical best predictor and linear mixed model based MBD methods for small area proportion estimation. Our empirical results, based on real data indicate that the gains from inclusion of spatial structure in SAE do not appear to be large. This is especially true for MBD estimation based on this structure (SMBDE), where the extra spatial information seems to have very little impact on the distribution of the SEBLUP weights that characterise this method of estimation. However, MBDE method seems to working well for discrete survey (binary) variable. Our results show that the MBDE performs well and represents an alternative to the EBP for the discrete data. We also note that in case of model misspecification (e.g. data with less variability), the MBDE appears to provide a more robust set of small area estimates.

There are many issues that still need to be explored in the context of using unit level models with spatially distributed area effects in SAE. The most important of these is identification of situations where inclusion of spatial information does have an impact, and the most appropriate way of then including this spatial information in the small area modelling process. An important practical issue in this regard relates to the computational burden in fitting spatial models to survey data. With the large data sets common in survey applications it can be extremely difficult to fit spatial models without access to high-end computational facilities. Although spatial information is becoming increasingly available in environmental, epidemiological and economic applications, there has been comparatively little work carried out on how to efficiently use this information. A further issue relates to the link between the survey data and the spatial information. In this paper we have assumed that all areas have sample units. In many situations this is not true, with survey data available only from a sample of areas. However, we often have spatial information for all areas. Saei and Chambers (2005) have explored the use of this spatial information in order to efficiently estimate the characteristics of the so-called 'out of sample' areas for the continuous response variables. A similarly work needs to be extended for the discrete response variable. Finally, we note that the spatial models considered in this paper have been based on neighbourhoods defined by contiguous areas. It is easy to see that this is just one way of introducing spatial dependence

between area effects, and several other options remain to be investigated, e.g. geographical weighted regression etc.

References

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

Betti, G. (2003) Poverty and Inequality Mapping in Albania: Final Report, *World Bank and INSTAT (mimeo)*, Washington DC and Tirana.

Breslow, N.E. & Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Model. *Journal of the American Statistical Association*, **88**, pp. 9-25.

Chambers, R., Chandra, H. & Tzavidis, N. (2007) On Robust Mean Squared Error Estimation for Linear Predictors for Domains, IASS Conference: Small Area Estimation 2007, Pisa.

Chandra, H. & Chambers, R.L. (2005) Comparing EBLUP and C-EBLUP for Small Area Estimation, *Statistics in Transition*, **7**, pp. 637-648.

Chandra, H., Salvati, N. & Chambers, R. (2007) Small Area Estimation for Spatially Correlated Populations. A Comparison of Direct and Indirect Model-Based Methods, *Statistics in Transition* **8**, pp. 887-906.

Elbers C, Lanjouw J & Lanjouw P (2003) Micro-Level Estimation of Poverty and Inequality, *Econometrica*, **71** (1), pp. 355-64.

Manteiga, G. W., Lombardia, M.J., Molina, I., Morales, D. & Santamaria, L. (2007) Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model, *Computational Statistics & Data Analysis*, **51**, pp. 2720-2733.

Jiang, J. (1998) Consistent Estimators in Generalized Linear Mixed Models, *Journal of the American Statistical Association* **93**, pp. 720-729.

Harville, D.A. (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, **72**, pp. 320–338.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models* (New York: Chapman and Hall).

McGilchrst, C.A. (1994). Estimation in Generalized Mixed Models, *Journal of the Royal Statistical Society Series B*, **56**, pp. 61-69.

Nelder, J. & Mead, R. (1965) A Simplex Method for Function Minimization, *Computer Journal*, 7, pp. 308–313.

Petrucci, A., Salvati N. & Seghieri C. (2003) Spatial Regression Models for Poverty Analysis, *Environment and Natural Resources Series*, 7, FAO, Roma.

Pratesi, M. & Salvati, N. (2007) Small Area Estimation: the EBLUP Estimator with Autoregressive Random Area Effects, forthcoming in *Statistical Methods & Applications*.

Petrucci, A. & Salvati, N. (2006) Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, *Journal of Agricultural, Biological and Environmental Statistics*, **11**, pp. 169-182.

Rao, J.N.K. (2003) *Small Area Estimation* (New York: Wiley).

Royall, R.M. (1976) The Linear Least-Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association*, **71**, 657-664.

Royall, R.M. & Cumberland, W.G. (1978) Variance Estimation in Finite Population Sampling, *Journal of the American Statistical Association*, **73**, pp. 351-358.

Saei, A. & Chambers, R. (2003) Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects, *Methodology Working Paper- M03/15*, Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom.

Saei, A. & Chambers, R. (2005) Out of Sample Estimation for Small Areas using Area Level Data *Methodology Working Paper- M05/11*, Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom.

Singh, B.B., Shukla, G.K. & Kundu, D. (2005) Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, **31**, pp. 183-195.

Tzavidis, N., Salvati, N., Pratesi, M. & Chambers, R. (2007). M-quantile Models with Application to Poverty Mapping, forthcoming in *Statistical Methods and Applications*.

World Bank (2003) Albania Poverty Assessment, *World Bank Report* **No. 26213-AL**, Washington DC.

Table 1 Relative Bias and Relative Root Mean Squared Errors for the Northern Tuscany data.

Regions are arranged in order of increasing population size.

	Relative Bias, %				Relative Root Mean Squared Error, %			
Regions	EBP	SEBP	MBDE	SMBDE	EBP	SEBP	MBDE	SMBDE
1	30.05	21.64	0.01	0.02	54.35	47.66	56.83	56.83
2	-4.58	-1.46	-0.30	-0.30	13.03	12.39	17.60	17.60
3	-14.38	-12.06	0.00	0.00	14.53	12.34	0.00	0.00
4	-0.71	3.63	-1.18	-1.18	20.58	18.96	36.82	36.82
5	147.71	126.17	5.78	5.74	167.49	144.74	110.26	110.23
6	82.45	113.25	-9.99	-10.51	97.36	125.18	102.60	102.09
7	5.49	6.87	-0.36	-0.37	26.43	25.74	29.62	29.62
8	-1.22	-0.71	-0.03	-0.09	10.62	10.14	14.26	14.26
9	-9.65	-7.61	-0.16	-0.26	20.06	18.10	23.31	23.31
10	-9.01	-8.91	0.16	0.09	9.81	9.68	7.31	7.35
11	12.31	13.35	0.56	0.50	28.81	28.85	31.08	31.07
12	3.65	2.06	-0.72	-0.73	19.94	19.08	23.51	23.52
13	-0.44	7.60	-0.15	-0.20	22.72	24.24	29.12	29.11
14	-5.20	-5.38	-0.16	-0.16	6.40	6.63	4.33	4.33
15	1.45	-2.33	0.41	0.39	14.80	15.11	18.78	18.78
16	-4.02	-4.91	0.11	-0.15	11.22	11.43	15.21	15.24
17	-0.05	2.38	-1.03	-1.03	21.66	21.16	23.90	23.90
18	-3.11	1.47	-0.52	-0.64	14.52	14.35	17.72	17.72
19	-0.59	-4.02	0.64	0.61	9.84	10.93	12.93	12.94
20	2.49	2.64	0.59	0.57	14.11	13.82	15.69	15.69
21	0.12	-1.77	-1.04	-1.07	10.83	10.78	12.71	12.71
22	0.17	-0.52	-0.31	-0.36	15.08	14.63	18.62	18.62
23	11.42	9.01	-1.93	-1.94	27.91	26.29	28.47	28.47
Mean	10.62	11.32	-0.42	-0.48	28.35	27.92	28.29	28.27

Table 2 Small area proportions and coverage rates for the Northern Tuscany data. Intervals are defined by the small area mean estimate plus or minus twice their corresponding estimated root mean squared error. Regions are arranged in order of increasing population size.

	Small Area proportion					Coverage rates			
		Estimated							
Regions	True	EBP	SEBP	MBDE	SMBDE	EBP	SEBP	MBDE	SMBDE
1	0.27	0.36	0.33	0.27	0.27	1.00	0.96	0.96	0.97
2	0.79	0.75	0.78	0.79	0.79	1.00	0.92	0.99	0.99
3	1.00	0.86	0.88	1.00	1.00	1.00	1.00	1.00	1.00
4	0.63	0.63	0.65	0.62	0.62	1.00	0.98	0.97	0.98
5	0.10	0.24	0.22	0.10	0.10	1.00	0.96	1.00	1.00
6	0.07	0.14	0.16	0.07	0.07	1.00	0.98	1.00	1.00
7	0.44	0.47	0.47	0.44	0.44	0.99	0.94	0.98	0.96
8	0.64	0.63	0.63	0.64	0.64	1.00	0.98	0.95	0.96
9	0.50	0.45	0.46	0.50	0.50	0.99	0.90	0.95	0.95
10	0.89	0.81	0.81	0.89	0.89	1.00	0.93	1.00	1.00
11	0.25	0.29	0.29	0.26	0.26	1.00	0.97	0.98	0.97
12	0.48	0.50	0.49	0.47	0.47	0.99	0.96	0.97	0.96
13	0.40	0.40	0.43	0.40	0.40	0.98	0.91	0.96	0.96
14	0.95	0.90	0.90	0.95	0.95	1.00	0.98	1.00	1.00
15	0.57	0.58	0.56	0.58	0.58	0.99	0.97	0.96	0.95
16	0.51	0.49	0.48	0.51	0.51	1.00	0.93	0.96	0.97
17	0.36	0.36	0.37	0.35	0.35	0.99	0.95	0.99	0.98
18	0.43	0.42	0.44	0.43	0.43	0.99	0.97	0.98	0.98
19	0.66	0.66	0.64	0.67	0.67	1.00	0.97	0.96	0.96
20	0.48	0.50	0.50	0.49	0.49	0.99	0.98	0.97	0.97
21	0.55	0.55	0.54	0.54	0.54	0.99	0.98	0.96	0.96
22	0.44	0.44	0.43	0.43	0.43	1.00	0.97	0.97	0.96
23	0.24	0.27	0.26	0.24	0.24	0.99	0.97	0.99	0.98
Mean	0.51	0.51	0.51	0.51	0.51	1.00	0.96	0.98	0.97

Table 3 Relative Bias and Relative Root Mean Squared Errors for the EMAP data. Regions are arranged in order of increasing population size.

	Relative Bias, %				Relative Root Mean Squared Error, %			
Regions	EBP	SEBP	MBDE	SMBDE	EBP	SEBP	MBDE	SMBDE
1	-8.13	-9.16	0.00	0.00	8.27	9.47	0.00	0.00
2	-1.72	-0.66	0.00	0.00	1.82	0.79	0.00	0.00
3	-14.08	-18.18	0.00	0.00	14.15	18.65	0.00	0.00
4	-4.23	-3.86	0.00	0.00	4.28	3.95	0.00	0.00
5	*	*	*	*	*	*	*	*
6	-1.06	-2.06	0.00	0.00	1.10	2.20	0.00	0.00
7	2.41	2.25	-0.86	-0.92	15.83	15.42	21.62	21.64
8	6.43	0.29	-0.68	-0.67	75.18	71.60	93.71	93.71
9	*	*	*	*	*	*	*	*
10	0.50	1.10	0.47	0.37	18.06	17.85	21.33	21.33
11	-2.40	-0.81	-0.18	-0.21	6.16	5.71	7.12	7.15
12	10.84	15.66	0.85	0.68	28.92	32.03	38.78	38.74
13	36.37	28.01	-1.48	-1.53	73.68	68.63	76.01	75.99
14	-0.35	-0.62	-0.16	-0.26	6.45	6.42	7.40	7.43
15	4.53	2.96	-0.91	-1.12	23.48	22.97	26.21	26.20
16	-4.65	-5.03	0.00	0.00	4.71	5.12	0.00	0.00
17	-2.64	-2.60	0.00	0.00	2.69	2.66	0.00	0.00
18	3.48	8.45	-0.89	-0.91	24.27	26.52	25.27	25.26
19	0.44	0.14	-0.97	-1.16	5.91	5.87	7.92	7.98
20	2.21	3.69	0.94	1.01	27.50	27.66	25.31	25.33
21	-0.72	-0.55	-0.45	-0.51	5.20	5.10	6.44	6.47
22	-2.17	-1.39	0.20	0.21	11.35	11.08	11.06	11.06
23	0.52	-0.45	-1.04	-1.26	8.43	8.39	10.98	11.03
Mean	1.22	0.82	-0.25	-0.30	17.50	17.53	18.05	18.06

* In these regions, true population proportion (dominators in calculating the relative measures) is zero so the relative bias or relative RRMSE cannot be calculated.

Table 4 Small area proportions and coverage rates for the EMAP data. Intervals are defined by the small area mean estimate plus or minus twice their corresponding estimated root mean squared error. Regions are arranged in order of increasing population size.

	Small Area proportion					Coverage rates			
		Estimated							
Regions	True	EBP	SEBP	MBDE	SMBDE	EBP	SEBP	MBDE	SMBDE
1	1.00	0.92	0.91	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	0.86	0.82	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	0.96	0.96	1.00	1.00	1.00	1.00	1.00	1.00
5	0.00	0.25	0.25	0.00	0.00	1.00	1.00	1.00	1.00
6	1.00	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00
7	0.76	0.77	0.77	0.75	0.75	0.87	0.87	0.85	0.93
8	0.28	0.30	0.28	0.28	0.28	0.91	0.87	0.98	0.80
9	0.00	0.11	0.10	0.00	0.00	1.00	1.00	1.00	1.00
10	0.63	0.64	0.64	0.64	0.64	0.94	0.94	0.94	0.96
11	0.93	0.91	0.92	0.93	0.93	1.00	0.94	1.00	1.00
12	0.43	0.49	0.51	0.44	0.44	0.98	0.96	0.94	0.94
13	0.22	0.30	0.28	0.22	0.22	0.97	0.97	0.97	0.97
14	0.86	0.86	0.86	0.87	0.87	0.93	0.94	0.97	0.99
15	0.36	0.37	0.37	0.36	0.35	0.95	0.96	0.98	0.97
16	1.00	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00
17	1.00	0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00
18	0.55	0.57	0.60	0.55	0.55	0.90	0.86	0.96	0.96
19	0.79	0.80	0.79	0.79	0.78	0.97	0.97	0.94	0.94
20	0.46	0.47	0.48	0.47	0.47	0.87	0.87	0.97	0.98
21	0.88	0.88	0.88	0.88	0.88	0.96	0.96	0.99	1.00
22	0.82	0.81	0.81	0.83	0.83	0.93	0.92	0.99	1.00
23	0.65	0.65	0.64	0.64	0.64	0.97	0.97	0.95	0.95
Mean	0.68	0.69	0.69	0.68	0.68	0.96	0.96	0.98	0.97

Table 5 Estimates and standard errors (in bracket) of the Head Count Ratio (HCR) index by district of Albanian for the LSMS data.

District	EBP	MBDE
BERAT	0.207 (0.037)	0.175 (0.047)
BULQIZE	0.651 (0.038)	0.679 (0.066)
DELVINE	0.197 (0.069)	0.286 (0.105)
DEVOLL	0.139 (0.065)	0.117 (0.086)
DIBER	0.206 (0.025)	0.242 (0.036)
DURRES	0.291 (0.033)	0.298 (0.033)
ELBASAN	0.206 (0.028)	0.239 (0.047)
FIER	0.114 (0.019)	0.122 (0.026)
GRAMSH	0.322 (0.038)	0.368 (0.065)
GJIROKASTER	0.077 (0.040)	0.000 (0.081)
HAS	0.155 (0.042)	0.073 (0.063)
KAVAJE	0.094 (0.028)	0.082 (0.036)
KOLONJE	0.208 (0.095)	0.293 (0.139)
KORCE	0.130 (0.028)	0.111 (0.039)
KRUJE	0.317 (0.064)	0.389 (0.081)
KUCOVE	0.241 (0.064)	0.237 (0.070)
KUKES	0.247 (0.026)	0.402 (0.128)
KURBIN	0.441 (0.053)	0.484 (0.071)
LEZHE	0.099 (0.031)	0.054 (0.044)
LIBRAZHD	0.286 (0.029)	0.351 (0.057)
LUSHNJE	0.132 (0.025)	0.134 (0.031)
MALESI E MADHE	0.221 (0.063)	0.306 (0.100)
MALLAKASTER	0.184 (0.057)	0.143 (0.064)
MAT	0.135 (0.044)	0.118 (0.082)
MIRDITE	0.149 (0.059)	0.122 (0.089)
PEQIN	0.352 (0.077)	0.421 (0.105)
PERMET	0.137 (0.060)	0.163 (0.087)
POGRADEC	0.258 (0.055)	0.259 (0.057)
PUKE	0.288 (0.071)	0.227 (0.086)
SARANDE	0.067 (0.028)	0.036 (0.036)
SKRAPAR	0.240 (0.087)	0.223 (0.097)
SHKODER	0.161 (0.027)	0.179 (0.034)
TEPELENE	0.207 (0.062)	0.237 (0.077)
TIRANE	0.150 (0.016)	0.117 (0.039)
TROPOJE	0.238 (0.038)	0.324 (0.088)
VLORE	0.051 (0.016)	0.028 (0.024)

Figure 1 Albanian district map.

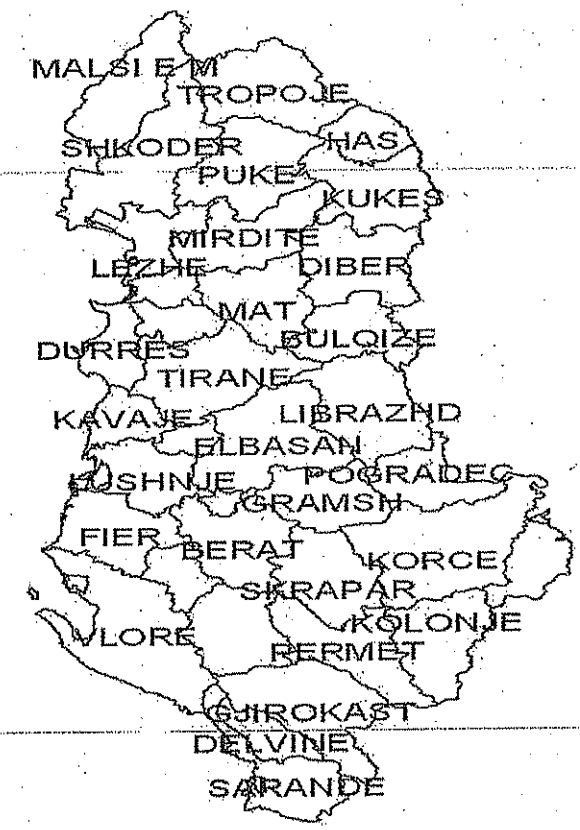

Figure 2 District level estimates of Head Count Ratio (HCR) under (i) EBP and (ii) MBDE.