Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 66(1) 2012 43-53

Use of Spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-Provincial Areas in Italy

Michele D'Alò[1], Loredana Di Consiglio[1], Stefano Falorsi[1], M. Giovanna Ranalli[2*] and Fabrizio Solari[1]

[1]*Direzione Centrale per le Tecnologie e il Supporto Metodologico, ISTAT*

[2]*Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, Italy*

Received 29 April 2011; Revised 31 August 2011; Accepted 01 September 2011

SUMMARY

The goal of this paper is to analyze the possibility to improve the performance of the estimation at sub-regional level from ISTAT Labour Force Survey. In particular, we refer to estimation of unemployment rates for small domains cutting across survey strata, i.e. Local Labour Market Areas, defined as aggregation of municipalities. Currently, such quantities are estimated by means of an EBLUP based on a linear mixed model with spatially correlated area effects and covariates given by the area level unemployment rate at previous census and sex by age classes. In this work we explore the use of alternative models to incorporate spatial information at different levels. In particular, we investigate the use of different distance measures in the correlation structure among small areas. In addition, additive models are employed to include spatial information at the municipality level using low-rank thin plate splines. Finally, small area estimators based on logistic (mixed) models are explored to account more properly for the binary nature of the response variable. Spatial information is included in this type of models too. The performance of the aforementioned methods is studied via simulation experiments on 2001 Census data.

Keywords : Labour force survey, Mixed effects models, Generalized additive models, Thin plate splines, Distance measures.

1. INTRODUCTION

In Italy, the Labour Force Survey (LFS) is conducted quarterly by the National Statistical Institute, ISTAT, according to a 2-2-2 rotation system to produce estimates of the labour force status of the population at a national, regional (NUTS2) and provincial (LAU1) level. In addition, since the year 1996 ISTAT also disseminates yearly LFS estimates of employed and unemployed counts at a finer level given by 686 Local Labour Market Areas (LLMAs). LLMAs are aggregations of municipalities and are defined at every census in terms of daily working commuting flows. LLMAs are, in contrast with NUTS2 and LAU1 areas, unplanned domains. In fact, the sampling design is as follows. Within a given Province (LAU1), municipalities are classified as Self-Representing Areas (SRAs) – larger municipalities – and Non Self-Representing Areas (NSRAs) – smaller ones. In SRAs a stratified cluster sampling design is applied. Each municipality is a single stratum and households are selected by means of systematic sampling. In NSRAs, the sample is based on a stratified two stage sample design. Municipalities are primary sampling units (PSUs), while households are Secondary Sampling Units (SSUs). PSUs are divided into strata of the same dimension in terms of population size. One PSU is drawn from each stratum without replacement and with probability proportional to the PSU population size. SSUs are selected by means of systematic sampling in each PSU. All members of each sample household, both

[*] *Corresponding author* : M. Giovanna Ranalli
E-mail address : giovanna.ranalli@stat.unipg.it

in SRAs and in NSRAs, are interviewed. Estimates of variables related to employement are based on the economically active population (people aged 15-74).

In each quarter, about 70,000 households and 1,350 municipalities are included in the sample. Note that some LLMAs – generally the smaller ones – may have a very small sample size. Furthermore, usually more than 100 LLMAs are not included in the sample at all (i.e. they have a zero sample size). Direct estimates may therefore have very large errors or they may not even be computable, thereby requiring resort to small area estimation techniques. Until 2003 a design based composite type estimator has been adopted. Since 2004, together with the redesign of the LFS sampling strategy, ISTAT estimates such quantities using an EBLUP based on a unit level linear mixed model with spatially autocorrelated random area effects. The following covariates are inserted in the fixed part of the model: sex by age classes – individual level covariates – and LLMA unemployment rate at previous census – area level covariates (D'Alò *et al.* 2004).

In this work we wish to investigate the performance of alternative small area estimators of the unemployment rate at LLMA level based on different ways of incorporating spatial information. The comparison will be conducted performing a simulation study on 2001 Census data that reproduces the estimation setting from the LFS. We will first consider distance measures alternative to the Euclidean distance currently employed for the unit level small area estimator. In particular, road distance and commuting times are considered in describing similarity among small areas. In addition, spatial information is also available at a finer definition, i.e. at a municipality level. We will first consider simple ways of introducing such information as covariates in linear mixed models. We will then move to the recently introduced EBLUP based on nonparametric regression that allows to combine small area random effects with a smooth, nonparametrically specified trend (Opsomer *et al.* 2008). By using penalized splines as the representation for the nonparametric trend, Opsomer *et al.* (2008) express the nonparametric small area estimation problem as a mixed effect model regression. The latter can be easily extended to handle bivariate smoothing. This allows to look at the spatial correlation structure in a different way. Finally, we will also investigate the performance of predictors based on logistic models and logistic mixed models to more properly account for the binary nature of the variable of interest. Different sets of covariates in the fixed part of the model will be considered. The paper is organized as follows. Section 2 provides a more detailed overview of the methods considered in the simulation, while Section 3 reports the simulation results. Final remarks and directions for future work are sketched in Section 4.

2. SMALL AREA TECHNIQUES FOR UNEMPLOYMENT RATE ESTIMATION

In this section we first introduce notation and then review the small area estimators considered and compared in the paper. We first treat more classical small area techniques, then move to the nonparametric regression based EBLUP of Opsomer *et al.* (2008) and finally to the logistic (mixed) model based estimators (e.g. Saei and Chambers 2003). Let a finite population U of dimension N be partitioned into d small domains (areas) of interest such that $\bigcup_{j=1}^{d} U_j = U$ and $\sum_{j=1}^{d} N_j = N$. The characteristic of interest y is observed on a sample s; in particular, let y_{ij} take value 1 if unit i in small area j is unemployed and 0 otherwise. We are interested in estimating the small area mean of y given by

$$\bar{y}_j = N_j^{-1} \sum_{i \in U_j} y_{ij}.$$

The following auxiliary variables are known for each unit in the population:

- sex_{ij} is the indicator variable that takes value 1 if unit i in small area j is a female and 0 otherwise;
- clage_{hij}, for $h = 1, ..., 7$ is an indicator variable that takes value 1 if unit i in small area j is in the h-th age class. The following age classes are considered: 15–19, 20–24, 25–29, 30–39, 40–49, 50–59, 60-74;
- unem_j denotes the unemployment rate of small area j at the previous census;
- lat_{ij} and lon_{ij} denote the latitude and longitude of the centroid of the municipality whom unit i in small area j belongs to.

2.1 Standard Small Area Estimators

The direct estimator for $\bar{y}_j$ is denoted by

$$\text{DIRECT}_j = \hat{N}_j^{-1} \sum_{i \in s_j} w_{ij} y_{ij}, \qquad (1)$$

where $\hat{N}_j = \sum_{i \in s_j} w_{ij}$, s_j is the set of sampled units in area j and w_{ij} is the sampling weight. The latter is given by the basic design weight – the inverse of the inclusion probability from the two stage sampling design – multiplied by an adjustment factor given by calibration adjustments on the known population distribution of socio-demographic variables like sex, age classes, nationality. Note that DIRECT$_j$ cannot be computed for areas with zero sample size.

A generalized regression type estimator for $\bar{y}_j$ is based on a standard linear model and can be expressed as an adjustment of the direct estimator for differences between the sample and population area means of the covariates inserted in the model (see e.g. Rao 2003, Chapt. 2). Two different models have been considered in this paper to this end. The first model uses the same covariates as those used in the Italian LFS. In fact, it considers sex by age classes interactions together with the unemployment rate effect. In particular,

$$y_{ij} = \sum_{h=1}^{7} \beta_h \text{clage}_{hij} + \sum_{h=1}^{7} \beta_{h+7} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{15} \text{unem}_j + \varepsilon_{ij}, \qquad (2)$$

with E(ε_{ij}) = 0 and V(ε_{ij}) = σ_ε^2. The generalized regression estimator takes form

$$\text{GREG}_j = \text{DIRECT}_j + \hat{N}_j^{-1} \Big(\sum_{i \in U_j} \hat{y}_{ij}^w - \sum_{i \in s_j} w_{ij} \hat{y}_{ij}^w \Big), \qquad (3)$$

where

$$\hat{y}_{ij}^w = \sum_{h=1}^{7} \hat{\beta}_h^w \text{clage}_{hij} + \sum_{h=1}^{7} \hat{\beta}_{h+7}^w (\text{sex}_{ij} \times \text{clage}_{hij}) + \hat{\beta}_{15}^w \text{unem}_j, \qquad (4)$$

and $\hat{\beta}^w$ denotes weighted least squares parameter estimates, with weights given by w_{ij}.

The second model considers the same set of covariates as model (2) plus a linear effect on the geographical coordinates. Therefore in this case

$$y_{ij} = \sum_{h=1}^{7} \beta_h \text{clage}_{hij} + \sum_{h=1}^{7} \beta_{h+7} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{15} \text{unem}_j + \beta_{16} \text{lon}_{ij} + \beta_{17} \text{lat}_{ij} + \varepsilon_{ij}, \qquad (5)$$

and the generalized regression estimator takes form (3) but with predictions that reflect the extra covariates included. The generalized regression estimator will be denoted as GREG-LFS under the first model and as GREG-LFS+C under the second one.

Under a model based framework, after the seminal work of Battese *et al.* (1988), it is now very common to use a linear mixed model to account for within area variation. In particular, in our application we will consider

$$y_{ij} = \sum_{h=1}^{7} \beta_h \text{clage}_{hij} + \sum_{h=1}^{7} \beta_{h+7} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{15} \text{unem}_j + u_j + \varepsilon_{ij}, \qquad (6)$$

where u_j, for j = 1,..., d denotes a set of random area effects independent of one another and independent of ε_{ij}, and such that u_j : (0, σ_u^2). It is a mixed effects model in which the fixed part is as in model (2). The role of the random effects in the model is to characterize differences in the conditional distribution of y given the covariates between the small areas. Restricted Maximum Likelihood estimates of the unknown parameters and predictions $\hat{u}_j$ are obtained under the assumption that all random effects are normally distributed. The small area estimator of the mean is then

$$\text{EBLUP}_j = \frac{1}{\hat{N}_j} \left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{y}_{ij} \right), \qquad (7)$$

where r_j denotes the non sampled set of units in area j such that U_j = $s_j \cup r_j$ and the unobserved value for population unit $i \in r_j$ is predicted using

$$\hat{y}_{ij} = \sum_{h=1}^{7} \hat{\beta}_h \text{clage}_{hij} + \sum_{h=1}^{7} \hat{\beta}_{h+7} (\text{sex}_{ij} \times \text{clage}_{hij}) + \hat{\beta}_{15} \text{unem}_j + \hat{u}_j. \qquad (8)$$

Under the aforementioned linear mixed model, a synthetic estimator is also considered, that takes form SYNTH$_j$ = $\hat{N}_j^{-1}\sum_{i\in U_j}\hat{y}_{ij}$, with $\hat{y}_{ij}$ given by

$$\hat{y}_{ij} = \sum_{h=1}^{7}\hat{\beta}_h \text{clage}_{hij} + \sum_{h=1}^{7}\hat{\beta}_{h+7}(\text{sex}_{ij}\times\text{clage}_{hij}) + \hat{\beta}_{15}\text{unem}_j,$$

i.e. as in (8) but without the predicted area effect $\hat{u}_j$. The EBLUP and this synthetic estimator have been computed also under another linear mixed model for which the fixed effects are those given in (5) and random components are as in (6). The estimator under the first model will be denoted as EBLUP-LFS, while the one under the second one as EBLUP-LFS+C. Similarly, the two synthetic estimators will be denoted by SYNTH-EB-LFS and SYNTH-EB-LFS+C, respectively.

2.2 Spatial EBLUP

The estimator that is now used at ISTAT to actually compute the unemployment rates for LLMAs is an EBLUP with spatially autocorrelated random area effects. In particular, the model can be written as in (6) but with random area effects such that

$$\mathbf{u} = (u_1,\ldots,u_d)\sim MN(0,\sigma_{\text{u}}^2\mathbf{A}), \tag{9}$$

where the matrix **A** depends on the distances among the areas and on an unknown parameter ρ connected to the spatial correlation coefficient among areas. In particular,

$$\mathbf{A} = [a_{j,j'}] = \left\{\left[1+\delta_{j,j'}\exp\left(\frac{\text{dist}(j,j')}{\rho}\right)\right]^{-1}\right\}, \tag{10}$$

with $\delta_{j,j'}=0$ if $j=j'$ and $\delta_{j,j'}=1$ otherwise, and $\text{dist}(j,j')$ denotes the Euclidean distance between area *j* and *j*′. The estimator in this case will be denoted as SEBLUP-LFS-Eu (Saei and Chambers 2003; D'Alò *et al*. 2004).

In this work we will also consider two alternative distance measures to describe the similarity among areas. In particular, the EBLUP with spatially autocorrelated random area effects will be computed also when dist(*j*, *j*′) in (10) is measured using a road distance – SEBLUP-LFS-Ro – and using commuting times – SEBLUP-LFS-Ti .

2.3 Small Area Estimators Based on Unit Level Thin Plate Splines Regression Models

Although very useful in many estimation contexts, in linear mixed models the fixed part of the model may not be flexible enough to handle estimation contexts in which the relationship between the variable of interest and some covariates is more complex than a linear model. Opsomer *et al*. (2008) usefully extended Battese *et al*. (1988) approach to the case in which the small area random effects can be combined with a smooth, nonparametrically specified trend. By using penalized splines (see e.g. Ruppert *et al*. 2003, Chapt. 4) as the representation for the nonparametric trend, the nonparametric small area estimation problem is expressed as a mixed effect model regression (see Opsomer *et al*. 2008, for more details on this). Once parameter estimates are obtained using REML, the small area estimator of the mean is as in (7) but with model predictions obtained using a different model.

In our application we consider the following nonparametric model in which a nonparametric trend in space is considered as an alternative way to model spatial correlation. In particular, the model is given by

$$y_{ij} = \sum_{h=1}^{7}\beta_h \text{clage}_{hij} + \sum_{h=1}^{7}\beta_{h+7}(\text{sex}_{ij}\times\text{clage}_{hij}) + \beta_{15}\text{unem}_j + m(\text{lon}_{ij},\text{lat}_{ij}) + u_j + \varepsilon_{ij}, \tag{11}$$

where $m(\cdot,\cdot)$ is a bivariate unknown smooth function to be learnt from the data and the area random effects are in this case uncorrelated. The bivariate function is approximated using low-rank thin plate splines, that are considered in Opsomer *et al*. (2008), too. In particular,

$$m(\text{lon}_{ij},\ \text{lat}_{ij}) = \beta_{16}\text{lon}_{ij} + \beta_{17}\text{lat}_{ij} + \sum_{k=1}^{K}\alpha_k z_{kij},$$

where z_{kij} is the $(ij)\times k$ entry of the matrix

$$\mathbf{Z} = [C(\mathbf{x}_{ij}-\boldsymbol{\kappa}_k)]_{\substack{i\in s_j,\, j=1,..d\\ 1\le k\le K}}[C(\boldsymbol{\kappa}_k-\boldsymbol{\kappa}_{k'})]^{-1/2}_{1\le k,k'\le K}\ , \tag{12}$$

in which $C(u)=\|u\|^2\log\|u\|$ are radial basis functions, $\mathbf{x}_{ij}$ = (lon$_{ij}$, lat$_{ij}$) denotes the geographical coordinates for observation *ij* and κ_k are fixed spline *knots*, for $k = 1, ..., K$. The second matrix on the right-hand side of expression (12) applies a linear transformation to the radial basis functions in the first matrix and has only the cosmetic role of making the radial spline look approximately like a thin plate spline when the number of knots approaches n. In fact, the number of knots K is usually much lower than the sample dimension. Ruppert *et al*. (2003, Chapt. 13) suggest to have one knot every four observations, with a minimum of 20 and a maximum of 150. The choice of the location of the knots is more important than that of their number and, in our experience, it is also more relevant in two than in one dimension. Knots should be sufficiently spread out over the spatial domain. In one dimension this is accomplished by placing knots at equally spaced sample quantiles. In two dimensions there is no easy extension of this approach, but the *clara* algorithm that is based on clustering and selects K representative objects out of n can be used to this end; it is implemented in the package cluster of the software R. The coefficients α_k for the knots are not considered as parameters, but as another random-effect vector in the linear mixed-model specification. Therefore, the complete specification of the random components for model (11) is given by

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\alpha}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} \text{ and } \mathrm{V}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\alpha}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{pmatrix}\sigma_u^2\,\mathbf{I} & 0 & 0\\ 0 & \sigma_\alpha^2\,\mathbf{I} & 0\\ 0 & 0 & \sigma_\varepsilon^2\,\mathbf{I}\end{pmatrix}.$$

Such specification allows to fit model (11) using standard linear mixed model software once the knots are selected and matrix **Z** is computed. Note also that the packages SemiPar and gcmv in R allow to fit model (11) very easily and possibly using different sets of basis functions. The estimator in this case will be denoted by EBLUP-LFS+SplC .

It could be argued that model (11) uses spatial information twice: firstly, through the function $m(\cdot,\cdot)$ in latitude and longitude and, secondly, with the area random effect **u** since small areas are geographical domains in this case. To verify whether such duplication is indeed useful, a low-rank thin plate spline model has also been used without area random effects. The small area estimates, therefore, are essentially synthetic estimates based on model

$$y_{ij} = \sum_{h=1}^{7}\beta_h \,\mathrm{clage}_{hij} + \sum_{h=1}^{7}\beta_{h+7}(\mathrm{sex}_{ij}\times \mathrm{clage}_{hij})$$
$$+\beta_{15}\,\mathrm{unem}_j + m(\mathrm{lon}_{ij},\mathrm{lat}_{ij}) + \varepsilon_{ij},$$

and will be denoted by SYNTH-LFS+SplC .

2.4 Small Area Estimators Based on Logistic Models

All the methods described so far assume normality of the response variable. It seems therefore natural to account more properly for the binary nature of the variable of interest, by using small area estimators based on logistic (mixed) models (see e.g. Saei and Chambers 2003). Note, however, that if in classical statistics using normal models for a binary variable instead of the appropriate logistic ones is usually deprecated, for small area statistics there is not a clear-cut evidence of the superiority in terms of performance of the latter over the former. The estimation problem we are treating in this paper is an example of this. An invited session on small area estimates in official statistics discussed estimation of labour force status in different countries at the Small Area Estimation Conference held in Elche in June 2009. Elazar *et al*. (2009) discuss superiority of binomial and multinomial logistic mixed models in simulations for estimates of labour force status at Local Government Areas in Australia, while Boonstra *et al*. (2009a) do not find evidence for the superiority of logistic mixed models over their normal counterparts in estimation of unemployed counts in Dutch municipalities. It seems therefore important to investigate and compare the performance of these different models for the Italian data, too.

To this end we consider the following general model

$$\begin{cases} y_{ij} \sim Bernoulli(p_{ij}) \\ \mathrm{logit}(p_{ij}) = \log\dfrac{p_{ij}}{1-p_{ij}} = \eta_{ij}' \end{cases}$$

under which the small area estimator of the mean takes form

$$\mathrm{LOGIT}_j = \frac{1}{\hat{N}_j}\left(\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\hat{p}_{ij}\right), \qquad (13)$$

where $\hat{p}_{ij}$ is the estimated probability of being unemployed according to different models. Note that, once model parameters are estimated, the prediction of population proportions can be obtained also when individual covariate information is not available by taking into account the collection of all demographic classes in which the population can be divided by covariate cross-classification. The hypothesis is that all individuals *j* in area *i* belonging to the same class have the same probability p_{ij} (see Malec *et al.* 1997, for more details on this). We will consider different forms for the linear predictor η_{ij}. In particular,

$$\eta_{ij} = \sum_{h=1}^{7}\beta_h \,\text{clage}_{hij} + \sum_{h=1}^{7}\beta_{h+7}(\text{sex}_{ij}\times\text{clage}_{hij}) + \beta_{15}\,\text{unem}_j,$$

$$\eta_{ij} = \sum_{h=1}^{7}\beta_h \,\text{clage}_{hij} + \sum_{h=1}^{7}\beta_{h+7}(\text{sex}_{ij}\times\text{clage}_{hij}) + \beta_{15}\,\text{unem}_j + \beta_{16}\,\text{lon}_{ij} + \beta_{17}\,\text{lat}_{ij},$$

$$\eta_{ij} = \sum_{h=1}^{7}\beta_h \,\text{clage}_{hij} + \sum_{h=1}^{7}\beta_{h+7}(\text{sex}_{ij}\times\text{clage}_{hij}) + \beta_{15}\,\text{unem}_j + m(\text{lon}_{ij},\text{lat}_{ij}).$$

The first model includes the standard covariates, while in the second one geographical coordinates are included linearly. These will produce small area estimators that will be denoted by LOGIT-LFS and LOGIT-LFS+C, respectively. The third model includes the unknown bivariate function in latitude and longitude discussed in Section 2.3 and estimated again by including a random effect for the splines coefficients. The resulting small area estimator will be denoted by LOGIT-LFS+SplC.

A twin triple of logistic models is also considered that includes an uncorrelated normal random area effect in the linear predictor. The resulting estimators will be denoted in clear notation by MLOGIT-LFS , MLOGIT-LFS+C and MLOGIT-LFS+SplC. Available software for generalized mixed models can be used to estimate these logistic models when we introduce one – or two as in this latter case – random effects. In fact, for the bivariate spline models once knots are selected and the **Z** matrix is built as illustrated in Section 2.3, an additional uncorrelated random effect is only needed to accomplish estimation. Note that, however, differently from the normal case, maximum likelihood estimates of the parameters and of the variance components in a logistic-normal mixed model is hindered by the presence of a ***J*** dimensional integral, so that direct calculation is intractable and well known computational issues arise. Therefore, more attention should be paid when computing these estimators by, possibly, comparing results coming from different softwares that use different algorithms.

3. SIMULATION RESULTS

The empirical study has been conducted selecting *R* = 500 samples from the 2001 Census data set. Each sample has been selected using the same two-stage scheme adopted for the LFS, and described in the Introduction, in order to reproduce the same estimation setting. The unemployment rate has been estimated in the 120 LLMAs belonging to the geographical area of "North-East of Italy". Let $\hat{\bar{y}}_j^r$ be the value of a small area estimator of the mean at replicate *r*, then the following evaluation criteria have been computed.

- % Relative Bias: $RB_j = \frac{1}{R}\left[\sum_{r=1}^{R}\frac{\hat{\bar{y}}_j^r - \bar{y}_j}{\bar{y}_j}\right]100$.
- Average Absolute RB: $AARB = \frac{1}{d}\sum_{j=1}^{d}|RB_j|$.
- Maximum Absolute RB: $MARB = \max_j |RB_j|$.
- % Relative Root Mean Squared Error: $RRMSE_j =$ $\sqrt{\frac{1}{R}\left[\sum_{r=1}^{R}\left(\frac{\hat{\bar{y}}_j^r - \bar{y}_j}{\bar{y}_j}\right)^2\right]}100$.
- Average RRMSE: $ARRMSE = \frac{1}{d}\sum_{j=1}^{d}RRMSE_j$.
- Maximum RRMSE: $MRRMSE = \max_j \ RRMSE_j$.

Table 1 reports such quantities for all the estimators computed. The software R has been used to compute all the estimators. Description of the estimators is given in the previous Section. Recall that the DIRECT estimator can be computed only on sampled areas. The unknown bivariate function $m(\text{lat}_{ij}, \text{lon}_{ij})$ is estimated by low-rank thin plate splines

Table 1. Simulation results: Average and Maximum Absolute Relative Bias, Average and Maximum Relative Root Mean Squared Error for all estimators.

ESTIMATOR	AARB	MARB	ARRMSE	MRRMSE
DIRECT*	3.01	15.22	60.24	119.51
GREG-LFS	8.40	94.05	51.85	157.22
GREG-LFS+C	7.39	67.96	51.47	124.26
SYNTH-EB-LFS	22.90	205.74	24.93	206.75
SYNTH-EB-LFS+C	18.68	144.48	22.64	147.68
SYNTH-LFS+SplC	14.35	112.09	22.63	122.69
EBLUP-LFS	20.41	193.79	24.98	195.95
EBLUP-LFS+C	17.26	138.79	22.70	142.64
EBLUP-SplC	13.91	115.17	22.80	125.86
SEBLUP-LFS-Eu	18.50	178.74	24.11	181.86
SEBLUP-LFS-Ro	17.26	167.48	23.27	171.26
SEBLUP-LFS-Ti	17.44	167.96	23.43	171.90
LOGIT-LFS	22.65	164.50	24.99	165.14
LOGIT-LFS+C	21.63	109.67	25.42	111.12
LOGIT-LFS+SplC	16.61	93.81	21.95	96.39
MLOGIT-LFS	19.88	152.65	25.76	154.79
MLOGIT-LFS+C	19.58	103.31	25.63	105.15
MLOGIT-LFS+SplC	16.29	93.95	22.47	96.53

*Performance is evaluated only on replicates for which small areas were not empty.

as illustrated in Section 2.3 using $K = 45$ knots selected out of the 1483 municipalities in the population. The knots have been kept fixed over repeated sampling. The number of municipalities in each sample is on average 250.

First it can be observed that the two GREG estimators have very low bias and high variance compared to the other model based estimators, as expected. The inclusion of the geographical coordinates of the municipality is a first simple way to account for spatial variability and provides a decrease in bias for the GREG. The bias reduction is confirmed also for all the other estimators, in fact, if we compare the LFS and the LFS+C versions, inserting the spatial information at the municipality level allows a decrease in bias in all cases and a corresponding decrease in the overall error. Such improvement is evident also from the corresponding values on the second column of Table 1, showing significant reductions of the MARB. The synthetic estimators show higher bias and a similar variance compared to the corresponding EBLUP based estimators, by this ascertaining the significance of an area effect, also when using the spatial information.

To find out the best way to model the spatial structure of the data, we can compare the performance of EBLUP-LFS+C , SEBLUP-LFS-Eu and EBLUP-LFS+SplC. In fact, they all include the same covariates in the model, but model different spatial information in different ways. Since SEBLUP-LFS-Eu has a better performance than EBLUP-LFS both in terms of bias and variance, spatial autocorrelation of the small areas is significant. Moreover, it seems to be better modelled using different distance measures than the Euclidean, given that both SEBLUP-LFS-Ro and SEBLUP-LFS-Ti show both a smaller bias and error. This may likely happen for the peculiar shape and geo-morphological structure of the small areas of interest. On the other side, EBLUP-LFS+C and EBLUP-LFS+SplC use the geographical coordinates of the municipalities: the former uses a plane, while the latter uses a smooth bivariate function. EBLUP-LFS+SplC shows a significant reduction in bias and a comparable ARRMSE, by this proving the existence of a fairly complicated and fine spatial structure. In addition, note that also SYNTH-LFS+SplC includes spatial information via p-splines and shows a performance very close to that of EBLUP-LFS+SplC, by this suggesting that, once spatial *p*-splines are employed, area effects are less significant. Note also that these two latter estimators have the lowest value for MARB among those based on normal models.

Logistic model based estimators have a performance comparable to that of the corresponding linear model based estimators, both in terms of bias and variance. It should be noted, however, the significant reduction in MARB for the logistic models when the spatial information is used. Even the simple inclusion of the geographical coordinates as in LOGIT-LFS+C and MLOGIT-LFS+C plays an important role. In addition, inclusion of the spatial component is important here by providing the lowest values of AARB and MARB. Fig. 1 shows the spatial structure $\hat{m}(\text{lat}_{ij},\text{lon}_{ij})$ as approximated in one of the random samples on the linear predictor scale for the logistic model. Note that there is an increase in the probability of being unemployed following a North-South and a West-East trajectory. Note also that we have investigated the performance of a different set of basis functions for logistic (mixed) models. In particular, we used a tensor product of B-splines as suggested in Wood (2006) but the results are very similar to those based on radial basis functions and are not reported here.

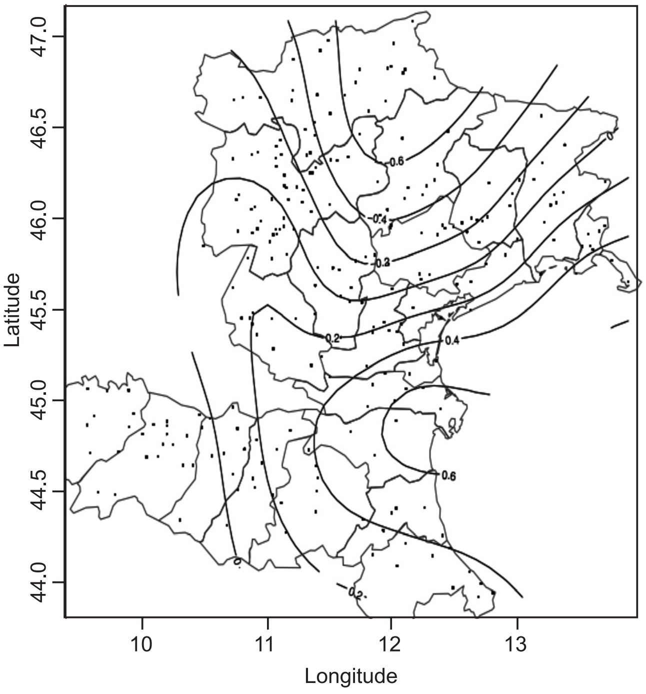

Fig. 1. The estimated effect of the bivariate structure on the linear predictor scale for the logistic model on one replicate sample (LAU1 provinces superimposed).

Fig. 2 compares the ARB for each area for pairs of selected estimators, by plotting the difference in ARB. Areas are arranged by increasing population size. Therefore, the first areas are those that are most likely and more frequently out of sample. The reduction in ARB is particularly significant for relatively smaller areas for the estimators that use spatial information and for those based on a logistic mixed model. The second panel shows that there is little difference when moving from a fixed to a mixed logistic model when using the same covariates. It makes a big difference, on the other hand, to use the spatial information (see first and second plot of the second row). The comparison between linear and logistic models, on the other hand, is not clear cut (see the third plot on the first and the second row).

Finally, Fig. 3 shows the Monte Carlo expected value of selected estimators with respect to their true value for each area. Logistic model based estimators show a tendency to over-shrink estimates. In particular, there is a clear overestimation in LLMAs with a very small unemployment rate. This tendency, however, is mitigated by the use of spatial information. EBLUP-SplC, on the other hand, seems to have, overall, a very good behavior in tracking both small and large true values. By looking at the plots of Fig. 3, it seems that there is still room to improve the specification of the models. With this regards auxiliary information plays a central role. In this experiment we have used all the unit level available information. It could be interesting to incorporate other possibly available area level information and/or move to spatial area level models.

In general, EBLUP-LFS+SplC and MLOGIT-LFS+SplC seem to be the overall best performing estimators. Therefore, on one side, the inclusion of spatial information at a finer (municipality) level than that currently used (area level) using bivariate smoothing techniques seems to be a useful tool to both decrease bias and improve efficiency. On the other side, however, given the same structure for the linear predictor, there is not a clearcut evidence of the superiority of logistic vs normal models: the former shows a tendency to overestimate small unemployment rates, while the latter has a larger MARB, but considerable computation advantages.

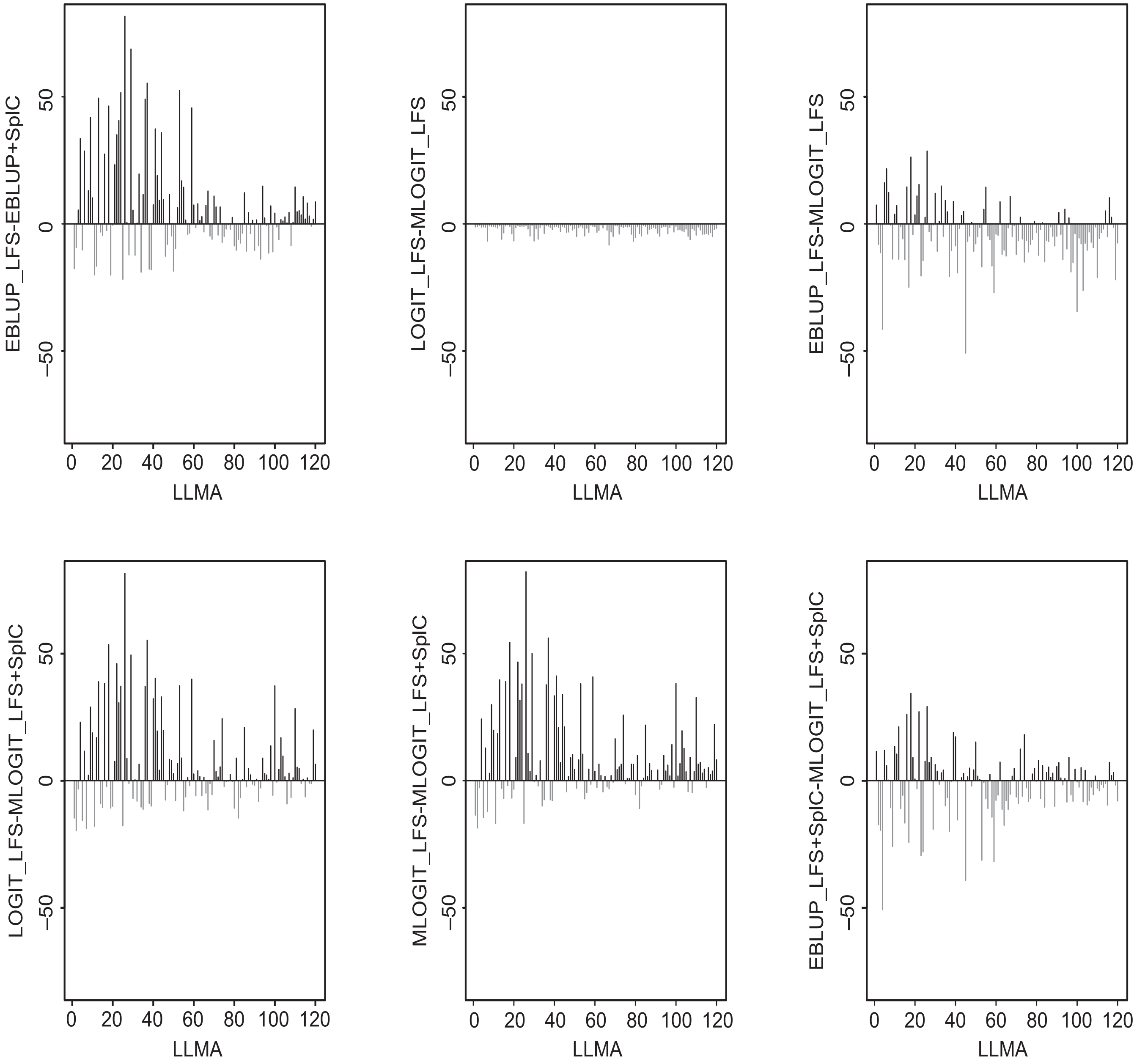

Fig. 2. Difference of ARB between pairs of selected estimators for each small area. Areas are arranged by increasing population size.

4. CONCLUSIONS AND FUTURE WORK

In this paper we investigate the role of the type of model and spatial auxiliary information used to estimate LLMAs unemployment rate from the Italian LFS. The performance of different small area estimators is compared via a simulation study in which the empirical sampling distribution of the estimators is obtained from using the LFS sampling design on the 2001 Census data. To summarize, our main finding here is that the spatial structure of the data helps to improve the accuracy of the estimates as measured by the empirical Bias and MSE, and this is true for all the considered estimators. In particular, reductions in bias are displayed when modeling the spatial structure via 2-dimensional *p*-splines. This is true for estimators based on both normal and logistic models. In addition, we have noted that, once spatial information is included in the model via *p*-splines, use of the area effect plays a very little role in our estimation problem.

The work done has some limitations, from which we envision new directions for future research. In

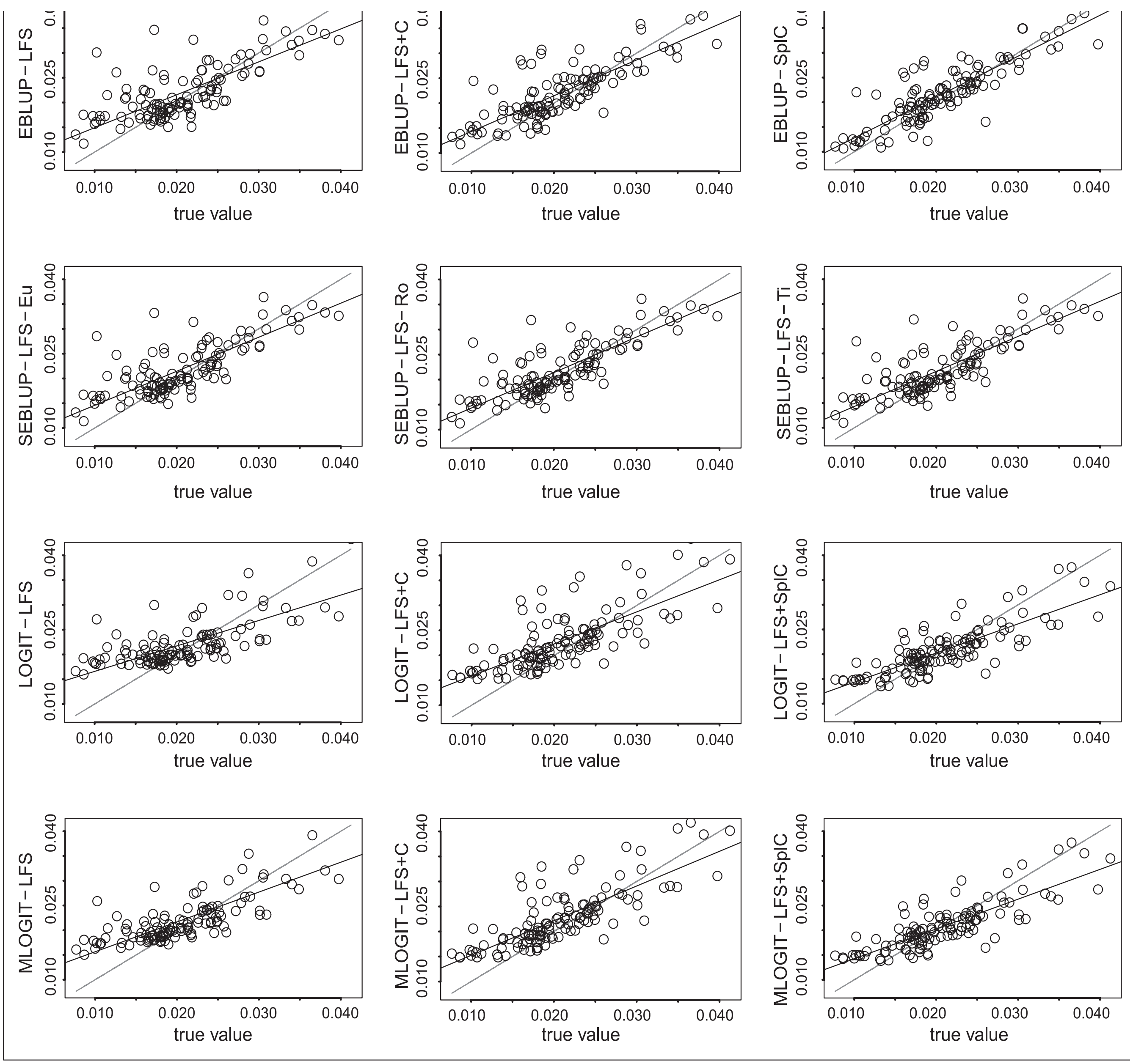

Fig. 3. Comparison of Monte-Carlo expectation and true value for each area for selected estimators. Grey is the 45° line, black is the regression line of Monte-Carlo expectations on true values.

particular, it is of great interest to explore the performance of estimators based on the models considered in this work that account to some extent of the complex sampling scheme employed in the Italian Labour Force Survey. Some preliminary work on pseudo-EBLUP estimators and on estimators based on linear mixed models with an extra random effect for municipalities (D'Alò *et al.* 2009) shows encouraging results and it would be interesting to extend it to the more complex models employed here. This is important not only for bias reduction, but also as a first step towards estimates that are consistent with respect to direct estimates produced at LAU1 level. In other words, it is also important to try to address the issue of benchmarking from within the definition of the model and not only as an *a-posteriori* adjustment.

In addition, we note that this work may be also considered as a model selection exercise conducted via Monte Carlo simulation. It would be very interesting to compare the results obtained here with those coming from model selection conducted on a subset of samples using Information Criteria or other methods developed for small area estimation problems like Fence methods (Jiang *et al.* 2008). The effectiveness of model selection measures could be evaluated both on a subset of samples with different levels of balancing with respect to the covariates and via a simulation study (see e.g. Boonstra *et al.* 2009b).

Finally, we have not mentioned variance estimation for the estimators considered in the simulation study. Note that variance estimators based on the classical Prasad and Rao (1990) approach have been proposed for the *p*-spline estimator for normal responses in Opsomer *et al*. (2008) and for logistic mixed models in Gonzales-Manteiga *et al.* (2007). It will be interesting, using these results, to derive and evaluate a variance estimator also for the *p*-splines logistic mixed model used in this paper.

ACKNOWLEDGEMENTS

Work supported by the project PRIN 2007 Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics.

REFERENCES

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

Boonstra, H., Buelens, B. and Smeets, M. (2009a). Estimation of labour force status for Dutch municipalities. *Proceedings of SAE2009, Conference on Small Area Estimation, Elche, Spain,* available at http://cio.umh.es/data2/M1A%20hbta@cbs.nl.pdf

Boonstra, H.J., Buelens, B. and Smeets, M. (2009b). Model selection for small area estimation. *Technical Report* DMH-2009-01-22-HBTA, Statistics Netherlands, Heerlen.

D'Alò, M., Di Consiglio, L., Falorsi, S. and Solari, F. (2004). The impact of the auxiliary information in the estimation of unemployment rate at sub-regional level: Further investigation on the Italian results in the Eurarea project. In *Proceeding of the European Conference on Quality and Methodology in Official Statistics*. Mainz, Germany.

D'Alò, M., Di Consiglio, L., Falorsi, S. and Solari, F. (2009). The use of sample design features in small area estimation. *Proceedings of the 57th Session of the International Statistical Institute*, *Durban, South Africa,* available at http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1449.pdf

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.,* **87**, 376-382.

Elazar, D., Hansen, N., Scealy, J., Lei, X., Khoo, J. and Davies, C. (2009). Small area estimates of labour force status in Australia. *Proceedings of SAE2009*, *Conference on Small Area Estimation, Elche, Spain.* available at http://cio.umh.es/data2 M1A%20 daniel.elazar@abs.gov.au.pdf

Gonzalez-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and Santamaria, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Comput. Statist. Data Anal.*, **51(5)**, 2720-2733.

Jiang, J., Rao, J.S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Ann. Statist.*, **36(4)**, 1669-1692.

Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.,* **92**, 815-826.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *J. Roy. Statist. Soc., Series B: Statistical Methodology*, **70(1)**, 265-286.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.,* **85**, 163-171.

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

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

Saei, A. and Chambers, R. (2003). Small area estimation under linear and generalized linear model with time and area effects. *Working Paper* M03/15, Southampton Statistical Sciences Research Institute, University of Southampton.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics,* **62(4),** 1025-1036.