

Small area models for unemployment rate estimation at sub-provincial areas in Italy ^(*)

Michele D'Alò¹, Loredana di Consiglio¹, Stefano Falorsi¹, Monica Pratesi²,
M. Giovanna Ranalli³, Nicola Salvati², Fabrizio Solari¹

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

² Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa

³ Dipartimento di Economia, Finanza e Statistica, University of Perugia

e-mail: giovanna.ranalli@stat.unipg.it

Abstract: 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 EBLUP based on a linear mixed model with spatially correlated area effects and covariates given by sex, age and area level unemployment rate at previous census. In this work we investigate the use of nonparametric additive models based on penalized splines to prevent misspecifications of the functional relationship between the target variable and some covariates. Moreover, in the same way, nonparametric additive modelling can be used to include spatial information using low-rank thin plate splines (Opsomer et al, 2008). In addition, we analyze the possibility of reducing bias of EBLUP through the use of model based direct estimation (Chambers e Chandra, 2005). Finally, small area estimators based on logistic (mixed) models are explored to account for the binary nature of the response variable more appropriately. The performances of the aforementioned methods are studied via simulated experiments on 2001 Census data.

Keywords: Labour force survey; Mixed effects models; Additive Models.

1. Introduction

In Italy, the Labour Force Survey (LFS) is conducted quarterly by ISTAT according to a 2-2-2 rotation system to produce estimates of the labour force status of the population, at national, regional (NUTS2) and province (LAU1) levels. 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) defined as aggregations of municipalities. LLMAs are areas defined at every census in terms of daily working commuting flows. LLMAs are, in contrast with the NUTS3 and LAU1 levels, unplanned domains. In fact, the sampling design is as follows. Within a given province, the municipalities are classified as Self-Representing Areas (SRAs) - consisting of the larger municipalities - and Non Self-Representing Areas (NSRAs) - consisting of the

^(*) 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.*

smaller ones. In SRAs a stratified cluster sampling design is applied. Each municipality is a single stratum and the households are selected by means of systematic sampling. All members of each sampled household are interviewed. In NSRAs the sample is based on a stratified two stage sample design. The municipalities are the primary sampling units (PSUs), while the households are the Secondary Sampling Units (SSUs). The 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. The SSUs are selected by means of systematic sampling in each PSU. All members of each sample household are interviewed. The quarter sample size in terms of households is about 70,000 and about 1,350 municipalities are included in the sample. Note that some LLMA – generally the smaller ones – may have very small sample size; furthermore, usually more than 100 LLMA are not even 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 was 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 – and LLMA unemployment rate at previous census – area level – (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. The comparison will be conducted performing a simulation study on 2001 Census data that reproduces the estimation setting from the LFS. In particular, we will consider 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 and additive modeling. This allows to look at different ways of incorporating the individual level auxiliary information – and the effect of age in particular – and spatial area correlation structure. In addition, since bias has always been a major concern for National Statistical Institutes, we will also investigate bias reduction techniques as the Model Based Direct estimator recently introduced by Chandra and Chambers (2005). This approach uses sample weights derived from a population level version of the linear mixed model to define a direct estimator. Finally, we will also investigate the performance of EBLUP 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 will first introduce notation and then review the small area estimators investigated in the paper. We will first treat more classical small area techniques, and

then move to the Model based direct estimator of Chandra and Cambers (2005), to the Nonparametric regression based EBLUP of Opsomer et al. (2008), and finally to the logistic (mixed) model based estimators. 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;
- age_{ij} denotes the age of unit i in small area j as an integer value;
- clage_{hij} , for $h = 1, \dots, 14$ is an indicator variable that takes value 1 if unit i in small area j is in the h -th age class. Age classes are the following: 0 – 14, 15 – 19, 20 – 24, 25 – 29, 30 – 34, 35 – 39, 40 – 44, 45 – 49, 50 – 54, 55 – 59, 60 – 64, 65 – 69, 70 – 74, 75+;
- unem_j denotes the unemployment rate of small area j at the previous census;
- lat_j and lon_j denote the latitude and longitude of the centroid of small area j .

Note that the first three types of variables denote an individual level auxiliary information, while the last two denote an area level one.

2.1. Standard small area estimators

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

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

where $\hat{N}_j = \sum_{i \in s_j} w_{ij}$, s_j denotes the set of sampled units in area j and w_{ij} is the sampling weight after calibration adjustments (Deville and Särndal, 1992). It is well known that DIRECT_j does not use auxiliary information and cannot be computed for areas with zero sample size.

The GREG estimator 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, Chapter 2). Two different models have been employed to this end. The first model will be called ‘‘Model LFS’’ since it uses the same covariates as those used in the LFS. In fact, it considers sex by age classes interactions together with the unemployment rate effect. In particular,

$$y_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j + \epsilon_{ij}, \quad (2)$$

with $E(\epsilon_{ij}) = 0$ and $V(\epsilon_{ij}) = \sigma_\epsilon^2$. The GREG estimator takes the form

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

where

$$\hat{y}_{ij}^w = \sum_{h=1}^{14} \hat{\beta}_h^w \text{clage}_{hij} + \sum_{h=1}^{14} \hat{\beta}_{h+14}^w (\text{sex}_{ij} \times \text{clage}_{hij}) + \hat{\beta}_{29}^w \text{unem}_j, \quad (4)$$

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 LFS plus a linear effect on the geographical coordinates. Therefore in this case

$$y_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j + \beta_{30} \text{lon}_j + \beta_{31} \text{lat}_j + \epsilon_{ij}, \quad (5)$$

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

Under a plain model based approach, a synthetic estimator is considered under Model LFS in (2), that takes the form

$$\text{SYNTH-LFS}_j = \hat{N}_j^{-1} \sum_{i \in U_j} \hat{y}_{ij}, \quad (6)$$

with \hat{y}_{ij} given by least squares predictions under the linear model. In other words, it would be like predictions in (4) but with parameter estimates obtained without design weights.

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, the Linear Mixed Model LFS has the following form

$$y_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j + u_j + \epsilon_{ij}, \quad (7)$$

where u_j , for $j = 1, \dots, d$ denotes a set of random area effects independent of one another and independent of ϵ_{ij} , and such that $u_j \sim (0, \sigma_u^2)$. It is a mixed effects model in which the fixed part is as in Model LFS in (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

$$\text{EBLUP}_j = \frac{1}{\hat{N}_j} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{y}_{ij} \right\}, \quad (8)$$

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

$$\hat{y}_{ij} = \sum_{h=1}^{14} \hat{\beta}_h \text{clage}_{hij} + \sum_{h=1}^{14} \hat{\beta}_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \hat{\beta}_{29} \text{unem}_j + \hat{u}_j. \quad (9)$$

Under the aforementioned linear mixed model, the synthetic estimator in (6) has been computed in which

$$\hat{y}_{ij} = \sum_{h=1}^{14} \hat{\beta}_h \text{clage}_{hij} + \sum_{h=1}^{14} \hat{\beta}_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \hat{\beta}_{29} \text{unem}_j,$$

i.e. as in (9) but without the 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 as in (7). 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.

The estimator that is now used at ISTAT to actually compute the unemployment rates at LLMA is an EBLUP with spatially autocorrelated random area effects. In particular, the model can be written as in (7) but with random area effects such that

$$\mathbf{u} = \{u_1, \dots, u_d\} \sim \text{MN}(\mathbf{0}, \sigma_u^2 \mathbf{A}), \quad (10)$$

where the matrix \mathbf{A} depends on the distances among the areas and on an unknown parameter ρ connected to the spatial correlation coefficient among areas. In particular

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

with $\delta_{j,j'} = 0$ if $j = j'$ and $\delta_{j,j'} = 1$ otherwise. The estimator in this case will be denoted as SEBLUP-LFS (Saei and Chambers, 2003; D'Alò et al., 2004).

2.2. Model Based Direct estimator and EBLUP based on unit level nonparametric regression models

Recently Chandra and Chambers (2005) proposed an alternative approach to small area estimation under a linear mixed model. Model Based Direct estimation uses sample weights derived from a population level version of the linear mixed model to define a direct estimator for a small area of interest. In particular, under the Linear Mixed Model LFS in (7) it is computed as

$$\text{MBDE-LFS}_j = \frac{\sum_{i \in s_j} w_{ij}^m y_{ij}}{\sum_{i \in s_j} w_{ij}^m}, \quad (11)$$

where weights w_{ij}^m are such that $\sum_{j=1}^d \sum_{i \in s_j} w_{ij}^m y_{ij}$ is the EBLUP of the population total of y , $\sum_{j=1}^d \sum_{i \in U_j} y_{ij}$ (Royall, 1976). Note that the MBDE resembles a ratio type direct estimator, with weights that are not based on the sampling design, but on the linear mixed model. Note that this type of estimator cannot be computed for areas with no sample units. Simulation experiments in literature show that this estimator has a performance that reduces the bias of the corresponding EBLUP estimator, to the price of a higher variance.

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 extend 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 (e.g. Ruppert et al., 2003) as the representation for the non-parametric 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 (8) but with model predictions obtained using different models.

In particular, in our application we consider the following nonparametric models. The first one tries to investigate the structure of the relationship between the response and age, i.e.

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_{ij} + \beta_2 \text{unem}_j + m(\text{age}_{ij}) + u_j + \epsilon_{ij}, \quad (12)$$

where $m(\cdot)$ is an unknown smooth function of the variable age and u_j are independent random area effects. The nonparametric EBLUP under this model will be denoted by EBLUP-SplA. A second nonparametric model extends this model to the case of autocorrelated area effects. In particular, model is as in (12) but with random area effects as in (10). The estimator in this case will be denoted by SEBLUP-SplA.

Finally, a nonparametric trend in space is considered as an alternative way to model spatial correlation. In particular, the model is

$$y_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j + m(\text{lon}_j, \text{lat}_j) + u_j + \epsilon_{ij}, \quad (13)$$

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 estimator in this case will be denoted by EBLUP-LFS+SpIC.

2.3. Small area estimators based on logistic models

To treat appropriately the binary nature of the variable of interest, we also test the performance of some small area estimators based on logistic models (see e.g. Saei and Chambers, 2003). In general,

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) &= \log \frac{p_{ij}}{1 - p_{ij}} = \eta_{ij} \end{aligned}$$

and the small area estimator of the mean will have the following form

$$\text{LOGIT}_j = \frac{1}{\hat{N}_j} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{p}_{ij} \right\}, \quad (14)$$

where \hat{p}_{ij} is the estimated probability of being unemployed according to different models. We will consider the following forms for the linear predictor η_{ij} :

$$\eta_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j \quad (15)$$

$$\eta_{ij} = \beta_0 + \beta_1 \text{sex}_{ij} + \beta_2 \text{unem}_j + m(\text{age}_{ij}) \quad (16)$$

$$\eta_{ij} = \sum_{h=1}^{14} \beta_h \text{clage}_{hij} + \sum_{h=1}^{14} \beta_{h+14} (\text{sex}_{ij} \times \text{clage}_{hij}) + \beta_{29} \text{unem}_j + u_j. \quad (17)$$

The first two models do not account for the area effect and the corresponding estimators will be denoted by LOGIT-LFS and LOGIT-SplA, respectively. The third model inserts an uncorrelated random area effect and the resulting estimator will be denoted by MLOGIT-LFS.

3. Simulation results

The empirical study has been conducted selecting $R = 500$ two-stage LFS samples from the 2001 Census data set. The unemployment rate has been estimated in the 127 LLMA's belonging to the geographical area of "Center of Italy". Let \hat{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: $\text{RB}_j = \frac{1}{R} \left[\sum_{r=1}^R \frac{\hat{y}_j^r - \bar{y}_j}{\bar{y}_j} \right] 100.$
- Average Absolute RB: $\text{AARB} = \frac{1}{d} \sum_{j=1}^d |\text{RB}_j|.$
- Maximum Absolute RB: $\text{MARB} = \max_j |\text{RB}_j|.$
- % Relative Root Mean Squared Error: $\text{RRMSE}_j = \sqrt{\frac{1}{R} \left[\sum_{r=1}^R \left(\frac{\hat{y}_j^r - \bar{y}_j}{\bar{y}_j} \right)^2 \right]} 100.$
- Average RRMSE: $\text{ARRMSE} = \frac{1}{d} \sum_{j=1}^d \text{RRMSE}_j.$
- Maximum RRMSE: $\text{MRRMSE} = \max_j \text{RRMSE}_j.$

Table 1 reports such quantities for all the estimators computed. The names of the estimators are explained in Section 2. Note that the DIRECT estimator has been computed only on sampled areas, while MBDE has been replaced by the corresponding synthetic estimator for non-sampled areas.

Firstly we can note that the direct estimator has the lowest AARB and the highest ARRMSE as expected. The two GREG estimators increase bias and decrease variance as expected compared to DIRECT. Including the geographical coordinates of the area allows for a first simple way to account for spatial variability and provides a slight decrease in bias for the GREG. The corresponding EBLUP estimators – EBLUP-LFS and EBLUP-LFS+C – show a larger bias, but a much lower variance compared to the GREG ones. The MBDE-LFS estimator provides a performance that is a compromise between GREG-LFS and EBLUP-LFS both in terms of bias and variance.

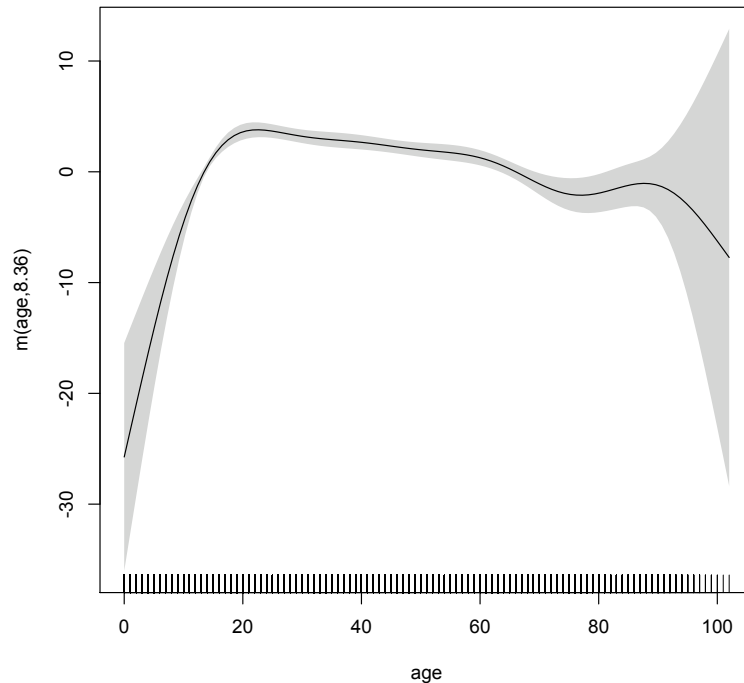
Table 1: *Simulation results: Average and Maximum Absolute Relative Bias, Average and Maximum Relative Root Mean Squared Error for all estimators.*

Estimator	AARB	MARB	ARRMSE	MRRMSE
DIRECT	2.9	20.4	51.7	90.7
GREG-LFS	7.2	83.3	40.2	93.8
GREG-LFS+C	6.9	71.5	40.0	82.8
SYNTH-LFS	18.2	91.1	19.8	91.8
SYNTH-EB-LFS	14.0	93.0	15.8	93.5
SYNTH-EB-LFS+C	12.4	79.7	16.4	81.0
EBLUP-LFS	13.2	92.5	16.2	93.1
EBLUP-LFS+C	11.9	79.5	16.7	80.7
SEBLUP-LFS	12.7	90.9	16.3	91.6
MBDE-LFS	8.8	86.3	35.3	92.6
EBLUP-SplA	13.2	89.8	16.5	90.5
SEBLUP-SplA	12.2	90.3	17.3	90.9
EBLUP-LFS+SplC	12.1	91.1	16.5	92.2
LOGIT-LFS	21.8	109.3	22.8	109.9
LOGIT-SplA	21.7	110.0	22.7	110.6
MLOGIT-LFS	11.4	82.3	21.8	83.9

The purely model based SYNTH-LFS shows both higher bias and variance compared to EBLUP-LFS, by this proving the existence of a significant area effect. In addition, SYNTH-EB-LFS shows a performance very close to the latter, by this proving that including the random area effect in the model is very important to reduce both the variability and the bias. We can compare the performance of EBLUP-LFS+C, SEBLUP-LFS and EBLUP-LFS+SplC to understand the spatial structure of the area effect. In fact, they all include the same covariates in the model they depend upon, and model the spatial structure in different ways. Their performance is comparable for both bias and variance, by this suggesting that the spatial structure of the area effect is not very complex. In addition, the simplicity of the model for EBLUP-LFS+C seems to keep bias more under control also in terms its maximum value. By comparing the performance of EBLUP-LFS and that of EBLUP-SplA we can note that there is little difference between the two. Recall that LFS covariates include the interaction of sex with 14 age classes, by this estimating 28 parameters. In EBLUP-SplA a common function for males and females is estimated using about 7 degrees of freedom. This may mean that the classes of age may be reduced or a more thorough functional relationship between age and the response may be studied.

The performance of the estimators based on logistic models is somehow disappointing. LOGIT-LFS is a fixed model that has a worse performance than the corresponding synthetic estimator – SYNTH-LFS. LOGIT-SplA shows a very similar performance to that of LOGIT-LFS by this again suggesting that a simpler and more parsimonious model in age should be detected. Figure 1 shows the estimated effect of age on the linear predictor scale, i.e. the estimate of $m(\text{age})$ in equation (15) for one of replicate sample. The probability of being unemployed increases until the age of 20, then it slowly

Figure 1: *The estimated effect of age on the linear predictor scale in the logistic additive model of equation (15) estimated on one replicate sample.*



decreases. This function is estimated from the data and uses approximately 8 degrees of freedom. In addition, LOGIT-SplA is a fixed model that shows a worse performance than EBLUP-SplA as far both bias and variance are concerned. Note that the former does not include the random area effects as the latter. Including such effect would imply the estimation of a generalized additive mixed model which is very computationally demanding. MLOGIT-LFS shows a significant reduction in bias as compared with LOGIT-LFS by this suggesting the presence of area effects.

4. Conclusions and future work

In this paper we have investigated the role of the type of model and auxiliary information used to estimate LLMA's unemployment rate from the Italian LFS. The performance of different small area estimators was compared via a simulation study in which the empirical sampling distribution of the estimators was obtained from the LFS sample. 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 MSE and this is true for all the considered estimators. This is true also for the estimators based on logistic models, although in this case the performance of the estimators is not satisfactory. More evidence and investigation is needed in this case to fully understand the role of the components of the model and the auxiliary information employed.

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

research. In our definition spatial interaction is explored using the spatial coordinates of the centroids of the small areas as covariates in the models and through the matrix A in model (10). Given the importance of the effect of the geography on the estimates, different models could be used in order to model the spatial interaction. There are geographical regression models under SAR and CAR specifications, which allow to estimate the strength of the spatial interaction using different tools to describe the contiguity of the small areas (Banerjee et al 2004).

MDB methods perform well but there are some problems to solve and some aspects to investigate. Negative weights impact on the utility of the MBD method and this remains unresolved and needs further attention. For example, negative weights, which occurred in some regions in the simulation study reported in Section 3, can lead to impossible (i.e. negative) estimates. Methods for dealing with negative weights under standard regression models have been discussed in the literature (e.g. Deville and Sarndal, 1992) but their application in the context of mixed models remains to be explored. In MBDE we assume that random area effects are independent between areas. However, we can extend the MBD approach under spatially correlated random area effect model similarly to the spatial-EBLUP (e.g. Pratesi and Salvati, 2008). Furthermore, we have not yet evidence of the effect of the introduction of nonparametric methods to MBD estimation.

References

- Banerjee S., Carlin B.P., Gelfand A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, New York.
- Battese G., Harter R., Fuller W. (1988) An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28–36.
- Chandra H., Chambers R.L. (2005) Comparing EBLUP and C-EBLUP for Small Area Estimation, *Statistics in Transition*, 7, 637–648.
- D’Alò M., Di Consiglio L., Falorsi S., Solari S., (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.
- Deville J.C., Särndal C.E., (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376–382.
- Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., Breidt F.J. (2008) Nonparametric small area estimation using penalized spline regression, *Journal of the Royal Statistical Society, Series B*, 70, 265–286
- Pratesi M., Salvati N. (2008) Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17-1, 114–131
- Rao J.N.K. (2003), *Small area estimation*, Wiley, New Jersey.
- Ruppert D., Wand M. P., Carroll, R. (2003) *Semiparametric Regression*, Cambridge University Press, Cambridge, New York.
- Saei A., 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