

Estimation of Regional Income Indicators Using Small Area Methods for Panel Data¹

La Stima di Indicatori di Reddito Regionali: Impiego di Metodi per Piccole Aree in Presenza di Dati di Panel

M.R. Ferrante, S. Pacei, E. Fabrizi

Dipartimento di Scienze Statistiche - Università di Bologna

{ferrante, pacei, fabrizi}@stat.unibo.it

Riassunto: La European Community Households Panel è un'indagine che consente di stimare parametri relativi al reddito, disegnata in modo da fornire stime affidabili per macroaree (NUTS1). Obiettivo del presente lavoro è di individuare un opportuno stimatore del reddito medio equivalente, che produca stime affidabili ad un livello territoriale più dettagliato. A tale scopo si considerano alcuni stimatori EBLUP per piccole aree, basati su specificazioni alternative di modelli lineari misti per dati di panel. La capacità di tali stimatori, sinora poco esplorati in letteratura, di ridurre la variabilità delle stime rispetto ai tradizionali stimatori per dati sezionali viene valutata mediante uno studio di simulazione. I risultati sono incoraggianti e consentono di selezionare un appropriato stimatore per piccole aree del reddito medio equivalente.

Keywords: regional estimation, European Household Community Panel, average equivalised income, linear mixed models, EBLUP.

1. Introduction

In recent years a renewed interest in the measure of the regional economies has produced an increased demand of regional statistical information. With reference to income, various European Union policies have been implemented in order to reduce territorial disparities and to support, in this way, social cohesion.

For European countries the estimation of income parameters can be obtained using information collected in the European Community Households Panel (ECHP), an annual survey coordinated by Eurostat, designed to provide estimates for large areas within countries called NUTS1 (for Italy, they are North-East, North-West, Centre, South and Island; Eurostat, 2003). In order to obtain reliable estimates on a smaller scale, "indirect" estimators, that try to increase efficiency by using values of the variables under study in related domains (small areas), could be employed.

The main purpose of this paper is to identify a suitable small area estimator for the average equivalised income (EQI), producing reliable estimates at the regional or provincial level (NUTS2 or NUTS3) by using longitudinal ECHP information.

Because of the nature of the ECHP, we suggest some alternative ways to specify Linear Mixed Models (LMM) for longitudinal data, in the small area estimation context.

¹ The present paper is financially supported by Murst-PRIN "Combining data from different sources: some methodological and practical issues - Coordinator Prof. N. Torelli, 2002" and "Statistical analysis on changes of the Italian productive sectors and their territorial structure - Coordinator Prof. C. Filippucci, 2003".

Estimates obtained in this way are compared to the cross-sectional ones by means of a Monte Carlo simulation experiment, allowing the evaluation of the gain in efficiency realized by the proposed solutions.

Over the last decade, in Small Area Estimation (SAE) theory, the longitudinal dimension has been considered in the field of time series techniques (for a review see Pfeiffermann, 2002) and thus in aggregated area level models. Individual longitudinal models, which consider repeated observations on the same unit, have not been extensively explored yet. Moreover, for both individual and aggregated longitudinal models, the advantages of “borrowing strength over time” as well as over areas needs further investigations.

The outline of the paper is as follows. The next Section gives an account of the relevance of the regional economies measures. The European Union interest for territorial disparities on income and the need for statistical information at a detailed geographic scale on this phenomenon are discussed in Section 3. In Section 4 a short introduction to SAE is provided focusing on the application of Linear Mixed Models in this context. In Section 5 cross-sectional and longitudinal models associated to the EBLUPs considered in this work are discussed. The simulation study implemented to evaluate the empirical properties of these estimators is described in Section 6. Section 7 provides the results arising from the comparison between the performance of cross-sectional estimators and the longitudinal ones considered in this paper. Finally, Section 8 offers some concluding remarks and possible further developments.

2. Regional economic issues and the need for statistical information at a detailed geographic scale

In recent years, regional development issue has received much attention from the scientific community and play a more and more important role in the political agenda of many countries. There are several reasons for this. Institutional integration (e.g. the European Union, the World Trade Organization) and international economic trends such as globalisation are somewhat eroding national borders and creating competition among regions of different countries, thus making the study and governance of regional economies particularly relevant. Moreover, concentration is probably the most striking feature of economic geography: in most countries production tends to be concentrated around few urban areas, industries are localized in highly specialized districts or poles, labour productivity and unemployment rates show large spatial variation. The persistence of economic disparities represents therefore a serious challenge to the ability of countries to promote economic growth while ensuring social cohesion. These issues are particularly relevant in the case of Italy, whose sub-national disparities have been a distinguishing feature of the economic landscape throughout its recent history. In fact, Italy's sub-national disparities are larger than those experienced by other EU member countries. For instance, let's consider the unemployment rate for people aged 15-24 at the level of Eurostat's NUTS2 as an indicator of the economic condition of sub-national regions: we have that, in 2003, the worst figure for Italy (Campania: 59.5%) was twelve times larger than the best (Trentino Alto Adige: 5%). In the UK, maximum and minimum for the same aggregate were respectively 18.4% and 7.8%; in France, without considering DOMs and TOMs, 31.3% and 13.6%; in Germany, a country which is notoriously experiencing wide regional inequalities, 24.7% and 3.9%. Because of this

situation, many initiatives, projects and debates about the analysis of “regional disparities” are going on.

Moreover, in programming actions related to the 2000-2006 EU’s Structural Funds, a number of economic/structural indicators (such as level of investments, unemployment rate, availability and quality of services for business and individuals) have been introduced to describe the state of the regions. The effectiveness of the interventions is then to be evaluated against their impact on these economic and structural indicators (Fantozzi and Prisco, 2004).

The interest on the study and the analysis of regional disparities highlights the need for statistical information at the smallest possible geographic scale. Unfortunately, sample surveys, including those conducted by the National Institutes of Statistics, due to budget constraints, are designed to produce reliable estimates for very large regions (such as Eurostat’s NUTS1) which are not very useful for the analysis of regional concentration and disparities. Censuses provide an alternative, but they are characterized by high non-sampling errors and, moreover, are held once every five or ten years.

To tackle the problem, in recent years several studies on SAE have been conducted from both the National Institutes of Statistics and Universities. At the European level we mention the EU funded Eurarea project that aims at assessing the effectiveness of the currently employed SAE techniques and to propose enhancements of techniques using data on unemployment, income and household composition from a representative group of six European countries (Heady and Hennell, 2001).

3. Estimation of household income indicators at the regional level

The dramatic economic, social and political changes that Europe has experienced in recent years have stimulated research and surveys on income distribution, poverty and social exclusion at a detailed territorial level. As noted above, the assessment and reduction of the large inequalities in the distribution of income and wealth among member countries and regions represents, for the EU, a priority in order to stimulate an equal participation of all regions and members states to the economic life of the Union. Currently, the measurement of income and of poverty by Union’s institutions is based on the ECHP (Betti and Verma, 2002; Eurostat, 2002).

In Italy official estimates about poverty are based on data from the Consumer Expenditure Survey, which is conducted by the Italian National Institute of Statistics (ISTAT). A second source of information on household income is provided by the Bank of Italy with Survey on Income and Wealth, conducted every two years.

The mentioned surveys provide reliable estimates for large geographic areas (NUTS1), which are scarcely useful for measuring territorial economies or for providing information to region’s governance.

The Bank of Italy has carried out some research on regional comparisons of EQI (see Cannari and D’Alessio, 2003) using data from the Survey on Income and Wealth. In that work the efficiency of estimators is improved by combining successive repetitions of the survey and using auxiliary information from other sources.

ISTAT has replied to this growing demand on geographically detailed information both by increasing the sample size of the Consumer Expenditure Survey and by stimulating the research on SAE techniques in order to improve the efficiency of currently employed estimators (Coccia *et al.*, 2002).

Since increasing sample sizes is costly and it risks to have bad effects on nonresponse and nonsampling errors, the approach based on the development of SAE methods seems more promising. In this spirit, we have taken into account the opportunity to use the information collected by the ECHP to obtain reliable estimates about income at a detailed geographic scale (such as NUTS2 or even NUTS3). At the moment, as far as Italy is concerned, data from the ECHP have not been used intensively (Peracchi, 2002) and, however, only to obtain estimates at the national level (Gallo and Siciliani, 2004).

Moreover, ECHP will be probably replaced by Statistics on Income and Living Conditions in the Community (EU-SILC), a survey with many features in common with the ECHP. The EU-SILC is expected to have a sample size large enough to provide reliable estimates at least for the NUTS2 level (Ghellini *et al.*, 2001), allowing for the possibility of testing methods to obtain reliable estimates also for NUTS3.

In summary, considering its territorial coverage, its richness of information and its panel design, we think that the ECHP may be a very interesting data source for providing reliable estimates in small areas and, thus, in obtaining regional information for the differential territorial analysis.

In the economy of this study, we restrict our attention to Italy and to the five available repetition of the survey, from 1994 to 1998. For details about the design and the implementation of the ECHP see Peracchi (2002).

We conclude with some consideration on the sampling unit and on the outcome variable we focus on. In this study we consider the household as the unit and the EQI, that is household income standardized with respect to the household composition, as the outcome variable. Based on the modified-OECD scale, EQI is obtained dividing the household disposable income by the number k of “equivalent adults”, defined as $k = 1 + 0.5a + 0.3c$, where a is the number of “other adults” (apart from the one defined as the head of the household) and c the number of children aged 13 or less. In general, the EQI can be read as the amount of income that an individual, living alone, should dispose of to reach the same level of economic welfare he enjoys in his household.

4. Linear Mixed Models in Small Area Estimation

In many sample surveys we are interested not only in estimates for the whole population but also for some of its subsets. A “small area” is any subset of a given, finite population defined according to some geographic criterion, whose specific sample size is not large enough to allow for direct estimators of suitable precision. An estimator is called “direct” if it is based only on the area-specific portion of the sample.

To improve the estimators’ precision it is often appropriate to use “indirect” estimators that borrow strength by using values of the variable under study in related areas, thus increasing the effective sample size. The study of indirect estimators, called “Small Area Estimation Theory”, has developed fast in recent years. For a thorough, up to date introduction to the subject, see Rao (2003).

To “borrow strength” we need a model linking the areas, often through the use of auxiliary information, available from external sources such as censuses or administrative data. The linking model can be either implicit or explicit: the first case is typical of design based estimators, the second of model based ones. The use of explicit models has the merit that the underlying assumptions can be tested and validated. In the

estimation of small areas means or totals of continuous variables, the class of LMM is very popular. The general LMM can be described as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the n -vector of sample observations, $\mathbf{X}$ and $\mathbf{Z}$ are known $n \times p$ and $n \times h$ matrixes, $\boldsymbol{\beta}$ is an unknown parameter of fixed effects, $\mathbf{v}$ and $\mathbf{e}$ are vector of random variables such that:

$$E\begin{pmatrix} \mathbf{v} \\ \mathbf{e} \end{pmatrix} = 0 \quad V\begin{pmatrix} \mathbf{v} \\ \mathbf{e} \end{pmatrix} = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \quad (2)$$

It is often assumed that $\mathbf{e}$ and $\mathbf{v}$ are normally distributed and that $\mathbf{G}$ and $\mathbf{R}$ are known up to a vector of “variance components” $\boldsymbol{\psi} = (\psi_1, \dots, \psi_q)$. As a consequence:

$$Var(\mathbf{y}) = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T = V(\boldsymbol{\psi}) \quad (3)$$

In SAE, we aim at predicting a linear combination $\boldsymbol{\eta} = \mathbf{m}^T\boldsymbol{\beta} + \mathbf{k}^T\mathbf{v}$ of the fixed and the random effects ($\mathbf{v}$). The best linear unbiased predictor (BLUP) of $\boldsymbol{\eta}$ can be obtained minimizing the model MSE among all linear estimators, assuming $\boldsymbol{\psi}$ known (see Robinson, 1991 for details). The BLUP is based on:

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \tilde{\mathbf{v}}(\boldsymbol{\psi}) &= \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \end{aligned} \quad (4)$$

If the variance components are not known they can be estimated according to a number of methods, leading to a two-stage estimator named empirical best linear unbiased predictor (EBLUP). In this paper we focus on the Restricted Maximum Likelihood (REML) method of estimation of variance components. We note that no distributional assumption is needed in the derivation of the BLUP, but normality is required for REML estimation of variance components, and for obtaining the approximated MSE of the EBLUP. For an introduction to the theory at the basis of the approximated formula for the MSE of the EBLUP, see Rao (2003), section 6.2.5. As regards the robustness of the EBLUP with respect to departure from normality, we remark that REML estimators are asymptotically normal when the number of fixed effects is bounded (see Jiang, 2001).

5. Linear Mixed Models “borrowing strength” over both time and areas

The Linear Mixed Models used in SAE can be classified into two broad types: disaggregated “unit level” models, in which unit values of the study variable are modeled, and aggregated “area level” models, which relates small area means or totals to area-specific covariates.

Since household level observations are available in the ECHP data set, we start considering unit level models. About EQI covariates, we consider the opportunity to include in the models only those for which small area means could be resumed from the Census results. This because to obtain the EBLUP of a small area mean, the mean of the covariates for that small area population has to be known (see Rao, section 5.3). Unfortunately, for the considered data set, Census data are not of much use. In fact we are considering the waves of the ECHP for the period 1994-1998 that is quite far from the 1991 Census. More in general, the use of Census data may not be very useful in modeling longitudinal data, since it doesn't allow for time-dependent covariates.

By the way, none of the candidate auxiliary variables (percentage of employee in the household, level of education of the household's head, etc.) shows a significant correlation with income at the household level. Then, at first, we concentrate in some random error variance linear models (see Rao, 1999; section 5.5.2).

In the attempt to improve the efficiency of estimates, we consider some alternative ways to specify models that take into account the longitudinal dimension. For brevity, we list the most interesting ones with respect to the results that are going to be discussed in Section 7:

$$\begin{aligned}
DM1 \quad y_{jdt} &= \mu + v_d + \alpha_t + e_{jdt} \\
DM2 \quad y_{jdt} &= \mu + v_d + \alpha_t + \gamma_j + e_{jdt} \\
DM3 \quad y_{jdt} &= \mu + v_d + \alpha_t + e_{jdt}^* \\
DM4 \quad y_{jdt} &= \mu + \alpha_{dt} + e_{jdt}
\end{aligned} \tag{5}$$

where y_{jdt} is the generic individual level observation, $j=1, \dots, n_{dt}$, $d=1, \dots, m$, $t=1, \dots, T$ indexing respectively, the individual observation, the area, the time. The fixed effect μ is the overall mean for all areas and all times; the area effect v_d , the time effect α_t , the individual effect γ_j , the area specific time effect α_{dt} are all assumed to be normal distributed, each with a common variance:

$$v_d \stackrel{IID}{\sim} N(0, \sigma_v^2) \quad \alpha_t \stackrel{IID}{\sim} N(0, \sigma_\alpha^2) \quad \gamma_j \stackrel{IID}{\sim} N(0, \sigma_\gamma^2) \quad \alpha_{dt} \stackrel{IID}{\sim} N(0, \sigma_{dt}^2) \tag{6}$$

Moreover we assume that all the effects are independent of each other. As regards the residuals, we assume that:

$$e_{jdt} \stackrel{IID}{\sim} N(0, \sigma_e^2) \quad e_{jdt}^* = \rho e_{jdt-1}^* + u_{jdt} \quad u_{jdt} \stackrel{IID}{\sim} N(0, \sigma_u^2) \tag{7}$$

To evaluate the impact of using longitudinal data on efficiency of estimators, a random effect sectional model has been considered as well:

$$DSM : \quad y_{jdT} = \mu_T + v_d + e_{jdT} \quad v_d \stackrel{IID}{\sim} N(0, \sigma_v^2) \quad e_{jdT} \stackrel{IID}{\sim} N(0, \sigma_e^2) \tag{8}$$

Note that, since the focus is on the estimation for the last available time T , the sectional model is specified only for time T .

Let's introduce some considerations on each model proposed. Respect to the sectional model (8), having just an area specific effect, in *DM1* a time effect is introduced and, then, parameters are estimated based on data referred to all waves. Considering the availability of observations collected on the same unit for many waves, *DM2* introduces an individual effect that can hopefully improve the fit of the model and than the efficiency of EBLUP. The same rationale has driven us in specifying in *DM3* an *AR*(1) structure for the residuals. Moreover, note that models *DM1-DM3* assume that area effects do not evolve over time. A more flexible assumption is specified in model *DM4*, where an interaction effect between area and time is introduced, that can be interpreted either as a time-specific area effect, or an area-specific time effect.

In the absence of individual level covariates, we also consider the use of area-level covariates. Possible sources for these covariates could be the Census or administrative data. We already mentioned the problems involved by the use of Census data. As regards administrative data, information on personal income at the regional level published by the Italian Ministry of Economy and Finance could have been useful, but they are not available for two of the considered years ('96 and '97) because of the change from paper to electronic form.

An alternative source of area level information are sample surveys. The use of sample estimates as covariates would require some modeling of the associated sampling error, but in the economy of the simulation exercise we are performing we overlooked this problem, that may represent the object of future work (within a Bayesian framework this problem is considered in Fabrizi, 2002). In particular, as an area-level covariate for EQI we use the unemployment rate estimated from the Italian Labour Force Survey (annual average), a time-dependent variable.

We can then revisit all models listed above introducing a fixed effect: *DM1**, *DM2**, *DM3**, *DM4**, *DSM**, each defined replacing the overall mean μ with $\mathbf{x}_{dt}\boldsymbol{\beta}$, where $\mathbf{x}_{dt}$ is the Unemployment Rate for area d at the time t . All the assumptions about the random effects and the residuals remain unchanged.

In most of the introduced models, we do not make use of any individual level covariate or random effect. Therefore it is sensible to compare these models against area level models, with the same covariate (if any) and similar structure for the random effects. In detail we consider the following models:

$$\begin{aligned} AM1 \quad \hat{y}_{dt} &= \mu + v_d + \alpha_t + e_{dt} \\ AM4 \quad \hat{y}_{dt} &= \mu + \alpha_{dt} + e_{dt} \end{aligned} \tag{9}$$

In this case $\hat{y}_{dt}$ is the direct estimate (that is, the sample mean) of EQI for area d and time t , $d = 1, \dots, m$, $t = 1, \dots, T$; μ is defined as before, while for the random effects the following assumptions are introduced :

$$v_d \stackrel{IID}{\sim} N(0, \sigma_v^2) \quad \alpha_t \stackrel{IID}{\sim} N(0, \sigma_\alpha^2) \quad \alpha_{dt} \stackrel{IID}{\sim} N(0, \sigma_{dt}^2) \tag{10}$$

As regards the residuals it is assumed that $e_{dt} \sim N(0, \sigma_{dt}^2)$, where σ_{dt}^2 is the variance associated to the direct estimate $\hat{y}_{dt}$.

For the purpose of comparison a sectional area-level model is considered:

$$ASM \quad \hat{y}_{dT} = \mu_T + v_d + e_{dT} \quad v_d \overset{iid}{\sim} N(0, \sigma_v^2) \quad e_{dT} \sim N(0, \sigma_{dT}^2) \quad (11)$$

Introducing the annual average of the unemployment rate as a covariate we can substitute μ with $\mathbf{x}_d \boldsymbol{\beta}$, leave all the other assumption unchanged and obtain the models $AM1^*$, $AM4^*$, ASM^* .

6. The simulation study

To compare the properties of EBLUP estimators associated to models described in the previous section, a Monte Carlo simulation study is carried out. Moreover, we note that approximated formulae for the MSE of EBLUP estimators are difficult to obtain for many longitudinal LMMs. This approximation depends on the assumption of normality and is of order $o(m^{-1})$, m being the number of areas. A number of works exists in the literature using Monte Carlo experiments for comparing different small area estimators: in most cases they are based on Censuses (Falorsi, Falorsi and Russo, 1994; Ghosh, Nanja and Kim, 1996). We cannot rely on Census data, as we study household income, a variable not measured by the Italian Census. Moreover, we focus on longitudinal models and, then, we need to generate data for several times.

Therefore, we have generated a synthetic population on the basis of the ECHP panel sample (Falorsi, Falorsi and Russo, 1999; Singh, Mantel e Thomas, 1994). For simplicity we restricted our attention to Italy and to the five available waves of the ECHP survey, from 1994 to 1998. To avoid the problems involved by attrition and nonresponse, we consider only the households present in all the five repetitions as a basis for data generation. The size of the ECHP sample in the considered period ranges from 7132 (in 1996) to 6571 (in 1998), while the number of households present in all repetitions is 5018. The synthetic population has been generated by replications on the basis of the sectional weights of the 1998 occasions of the survey. The size of the generated population is about 22 millions, approximately the number of households in Italy.

We consider as small areas of interest the eleven regions singled out in the ECHP files. These regions are somewhat intermediate between NUTS1 and NUTS2 (North-West, Lombardia, North-East, Emilia-Romagna, Center, Lazio, Abruzzo-Molise, Campania, South, Sicilia, Sardegna). Since the release of ECHP data is subject to strong confidentiality constraints, the information on NUTS2 or NUTS3 is not available to us. The regions considered are rather large, hence we base our Monte Carlo experiment on a sample of size 1,000, roughly about 15% of the actual average size of the ECHP survey. In this way we try to mimic the problems of estimation at the NUTS2, or at an even more detailed territorial level.

Samples are drawn from the synthetic population by stratified random sampling without replacement, with strata given by the considered regions. Thus regions are treated as planned domains (as in ECHP), that is the sample sizes in the small areas are fixed in advance and is proportional to the size of the region in terms of the number of household. Area-specific sample sizes range from 26 to 167.

7. Results

To compare the performances of the EBLUP estimators associated to the models discussed in the Section 5, we use average measures, that is indicators describing the performances of the estimators “on average” with respect to the set of the small areas under study. This is the usual approach in the literature (see Rao, 2003; section 7.2.6). We selected at first two indicators, the Average Relative Mean Square Error (ARMSE) and the Average Relative Bias (ARB):

$$ARB = m^{-1} \sum_{d=1}^m \left\{ R^{-1} \sum_{r=1}^R \left| \frac{\tilde{y}_{dtr}^{EBLUP}}{\bar{Y}_{dt}} - 1 \right| \right\}$$

$$ARMSE = m^{-1} \sum_{d=1}^m \left\{ R^{-1} \sum_{r=1}^R \left(\frac{\tilde{y}_{dtr}^{EBLUP}}{\bar{Y}_{dt}} - 1 \right)^2 \right\} \quad (12)$$

where $\tilde{y}_{dt}^{EBLUP}$ is the EBLUP estimate for area d , area t and replicated sample r , $\bar{Y}_{dt}$ is the known population mean for the same area and time. $R = 500$ is the number of replicated samples used in the experiment, since the MSE of all the estimators under study becomes stable after about 400 replicates.

Moreover, the longitudinal estimators' capability in increasing the efficiency respect to direct and sectional estimators is evaluated by the Average Relative Efficiency obtained as ratio between *i*) the *AMSEs* (Average Mean Squared Error) of the various EBLUPs considered $\tilde{y}_{dt}^{EBLUP}$ to that one of the direct estimator $\hat{\hat{y}}_{dt}$ (*AEFF_d*) and *ii*) the *AMSE* of longitudinal EBLUPs ${}_l\tilde{y}_{dt}^{EBLUP}$ to the *AMSE* of the sectional EBLUP ${}_s\tilde{y}_{dt}^{EBLUP}$ (*AEFF_s*):

$$AEFF_d = AMSE(\tilde{y}_{dt}^{EBLUP}) / AMSE(\hat{\hat{y}}_{dt}) \quad (13)$$

$$AEFF_s = AMSE({}_l\tilde{y}_{dt}^{EBLUP}) / AMSE({}_s\tilde{y}_{dt}^{EBLUP}) \quad (14)$$

To describe the results, we start considering the unit level models. Their performances in terms of *ARMSE*, *ARB*, *AEFF_d* and *AEFF_s* are illustrated in Table 1. At first, it can be observed that, when no covariate is present in the model, the impact of considering more waves is very important and allows for a significant gain in efficiency. In fact the EBLUPs related to models for panel data (*DM1-DM4*) are the most reliable ones, with a reduction of MSE respect to the variance of the direct estimator that ranges in 15-25%. Moreover, we note that these estimators always perform better than the correspondent sectional one, the gain in efficiency of *DM1-DM4* over *DSM* in terms of *AEFF_l* ranging from 7% to 18%.

Table 1: *Performances of the EBLUPs associated to the unit-level LMMs*

Model	$ARMSE\%$	$ARB\%$	$AEFF_d \%$	$AEFF_s \%$
<i>DSM</i>	0.414	2.203	91.4	100.0
<i>DM1</i>	0.343	2.317	75.6	82.9
<i>DM2</i>	0.339	2.186	74.7	81.9
<i>DM3</i>	0.358	2.861	79.0	86.5
<i>DM4</i>	0.387	1.798	85.3	93.5
<i>DSM*</i>	0.359	3.482	79.2	100.0
<i>DM1*</i>	0.314	2.638	69.3	87.5
<i>DM2*</i>	0.333	2.295	73.5	92.8
<i>DM3*</i>	0.317	3.460	70.0	88.3
<i>DM4*</i>	0.339	3.493	74.8	94.4

Estimators derived from *DM1* and *DM2* models show quite close performances, but when these two models are compared in term of some information criteria such as *AIC* or *BIC*, the latter exhibits a much better fit. The reason could be that the better fit is not useful for the predictive purposes of SAE. A similar consideration can be made comparing *DM1* and *DM3*. Then, as a general advice, model *DM1* seems to be the most advantageous among those without covariates, as it is much simpler than *DM2*. The estimator mentioned achieves, in fact, a reduction of MSE of about 25% with respect to the direct estimator and of 17% with respect to sectional estimator *DSM*.

As expected, the introduction of a covariate increases estimators' efficiency. In all cases the effect of considering more waves does not vanish even though it is reduced. In fact, the sectional model (*DSM**) leads to a gain in efficiency of the 20% and this gain increases up to about the 30% when time effects are introduced. However, we note that the introduction of the covariate leads to a rise of estimator' bias. This happens in particular for sectional models *DSM** and for two the sub-optimal longitudinal estimators (*DM3** and *DM4**).

We have also considered some modification of models *DM1-DM4*, *DM1*-DM4**, including time effects autocorrelated according to an AR(1) process, but the related results are not reported since they are not encouraging, probably because more time points are required to obtain an efficient estimate of the autocorrelation parameter.

In conclusion we observe from the analysis of random effects' prediction (here not reported for brevity) that area effects are generally strong and significant, while time effects are smaller and not significant in many cases. This confirms the evidence of large disparities among Italian regions and highlights the slowly changing nature of EQI, at least for the time period considered in the study. Results are not reported for brevity.

As regards the area level models the following results are obtained.

As it should be expected, in general the performances of the area-level models are very close to those of the similar unit level models in the view of the fact that area and time effects and the covariate are not unit-specific. The inclusion of a covariate makes the impact of using past waves less relevant than in the case without covariate. However for area level models this evidence is even more significant with respect to the unit level ones, as the gain in efficiency provided by the longitudinal model with covariate (*AMI**) compared to the sectional one (*ASM**) is just of the 4%. The longitudinal model

AM1 without the covariate performs very much like the sectional model *ASM** including the covariate ($AEFF_d=0,77$ in both cases).

In general, however, the comparison between estimators with and without covariate should be carried out carefully. In fact, the covariate we use is a sample estimate and the consideration of its variability can increase the overall MSE of estimators derived from these models.

Table 2: *Performances of the EBLUPs associated to the area-level LMMs*

Model	$ARMSE\%$	$ARB\%$	$AEFF_d\%$	$AEFF_s\%$
<i>ASM</i>	0.420	1.243	92.5	1.00.0
<i>AM1</i>	0.348	2.198	76.8	82.9
<i>AM4</i>	0.420	1.172	92.7	100.0
<i>ASM*</i>	0.348	2.188	76.9	100.0
<i>AM1*</i>	0.334	2.394	73.8	96.0
<i>AM4</i>	0.390	2.394	86.2	112.1

8. Concluding remarks and further developments

The obtained results show us that, in general, model specifications that “borrows strength over time”, as well as over areas, lead to a significant gain in efficiency with respect to direct estimation. However the most interesting result is that the MSE of some of the considered longitudinal estimators is considerably lower when compared to the MSE of the sectional ones, leading to a significant gain in efficiency especially for the unit models.

Moreover, the more simple model specifications appears preferable. In fact the EBLUPs related to the unit or area longitudinal model, where time and area specific random effects are considered in a simple additive structure, are always the most reliable. This result holds also when a covariate is introduced.

Some further aspects of the problem could be investigated. First of all, the usual problems related to the estimation of model from panel data, such as the treatment of the attrition and the new entries. Moreover it could be interesting to extend the model to incorporate the complexity of the sample design using pseudo-EBLUPs that consider unequal sampling weights (see Rao, 2003; section 7.2.7). The problem of estimating the MSE of the EBLUPs considered has also to be faced. In addition, the estimation of period-to-period change in average income should be investigated to give a picture of the different temporal evolution of income distribution in Italian regions.

Still, an interesting aspect in the economic studies on income is the analysis of territorial disparities. In this kind of analysis it should be considered that the collection of regional EBLUPs may show a lower variability than the underlying regional population parameters. To tackle this problem estimators can be adjusted according to some modified optimality criteria.

Finally, a more complete picture of the income distribution within the regions may be given by the estimation of some other income parameters (quantiles of income distribution, concentration indexes, etc.), which require the use of different estimation strategies.

References

- Betti G., Verma V. (2002) Non-monetary or Lifestyle Deprivation, in EUROSTAT (2002) *Income, Poverty Risk and Social Exclusion in the European Union*, Second European Social Report, 87-106.
- Cannari L., D'Alessio G. (2003) La distribuzione del reddito e della ricchezza nelle regioni italiane, *Banca d'Italia: Temi di discussione del Servizio Studi*, 482.
- Eurostat (2002) *European social statistics - Income, poverty and social exclusion*, 2nd report.
- Eurostat (2003) *Regions. Nomenclature of territorial units for statistics, NUTS – 2003*, Methods and Nomenclatures series.
- Fabrizi E. (2002) Hierarchical Bayesian Models for the estimation of Unemployment Rates in Small Domains of the Italian Labour Force Survey, *Statistica*, in press.
- Falorsi P.D., Falorsi S. and Russo A. (1994) Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey, *Survey Methodology*, 20, 171-176.
- Falorsi P.D., Falorsi S. and Russo A. (1999) Small Area Estimation at Provincial Level in the Italian Labour Force Survey, *Journal of the Italian Statistical Society*, 1, 93-109
- Fantozzi D., M.R. Prisco (2004) Indicatori regionali per la valutazione delle politiche di sviluppo, relazione presentata al workshop Istat "Informazione statistica, territorio e attività di policy", Roma, 2 Febbraio 2004, www.istat.it.
- Ghellini G., A. Schizzerotto, H.P. Blossfeld, A. D'Agostino, B. Buldo, V. Napoli (2001) *L'esperienza in tema di indagini Multiscopo e ECHP: lezioni ed indicazioni nella prospettiva di una indagine panel sulle famiglie*, Rapporti di ricerca della Commissione di Garanzia sull'Informazione Statistica, Presidenza del Consiglio.
- Ghosh M., Nangia N. and Kim D. (1996) Estimation of Median Income of Four-Person Families: a Bayesian Time Series Approach, *Journal of the American Statistical Association*, 91, 1423-1431.
- Gallo F., Siciliani I. (2002) La situazione finanziaria delle famiglie e degli individui in Italia e in Europa- Anni 1994-1997, Istat, Collana Informazioni (in corso di stampa).
- Heady P., Hennell S. (2001) Enhancing small area estimation techniques to meet European needs, *Statistics in transition*, 5, 2, 195-203.
- Jiang J. (2001) Goodness-of-Fit Tests for Mixed Model Diagnostics, *The Annals of Statistics*, 29, 1137-1164.
- Peracchi F. (2002) The European Community Household Panel: A Review, *Empirical Economics*, 27, 63-90.
- Pfeffermann D. (2002) Small Area Estimation – New Developments and Directions, *International Statistical Review*, 70, 125-143.
- Rao J.N.K. (2003) *Small Area Estimation*, Wiley, New York.
- Robinson G. K. (1991) That BLUP is a Good Thing: the Estimation of Random Effects, *Statistical Science*, 6, 15-31.
- Singh A.C., Mantel HJ., Thomas B.W. (1994) Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20, 1, 33-43.