STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 423—439

COMPARING ALTERNATIVE DISTRIBUTIONAL ASSUMPTIONS IN MIXED MODELS USED FOR SMALL AREA ESTIMATION OF INCOME PARAMETERS

Enrico Fabrizi, Maria Rosaria Ferrante, Silvia Pacei

ABSTRACT

Linear Mixed Models used in small area estimation usually rely on normality for the estimation of the variance components and the Mean Square Error of predictions. Nevertheless, normality is often inadequate when the target variable is income. For this reason, in this paper we consider Linear Mixed Models for the log-transformed income (which require back-transformation for prediction of means and totals on the variable's original scale) and a Generalized Linear Mixed Model based on the Gamma distribution. Various prediction methods are compared by means of a simulation study based on the ECHP data. Standard predictors obtained from Linear Mixed Model for the untrasformed income are shown to be preferable to the considered alternatives, confirming their robustness with respect to the failure of the normality assumption.

Key words: European Community Household Panel; Average Equivalized Income; Lognormal Linear Model; Prediction; Gamma Distribution.

1. Introduction

In the European Union, the demand for estimates about the distribution of income at the sub-national level, a fundamental tool for the implementation of social cohesion policies, has grown rapidly in recent years (Stewart, 2003). For the period from 1994 to 2001, income distribution parameters and poverty indicators may be estimated consistently across most of the member states using information collected by the European Community Households Panel (ECHP), a sample survey on households' income and social conditions, coordinated by Eurostat (Betti and Verma, 2002; Eurostat, 2002). This panel survey was designed to provide reliable estimates of main parameters of interest for large areas within

countries called NUTS1 (NUTS stands for the "Nomenclature of Territorial Units for Statistics"; Eurostat, 2003).

The ECHP survey was substituted in 2004 by a new rotating panel survey called EU-SILC (European Union — Statistics on Income and Living Conditions), based on new measurement methodologies and a larger sample (Eurostat, 2005). The two surveys are very similar under many aspects and ECHP data pertaining to Italy is used for the purposes of this paper.

We are interested in estimating the mean of equivalized household income for sub-national regions defining a partition of the country, for which direct estimators, that is, those applying standard weighted estimators to the region-specific part of the sample, lead to estimates with too large a variance. The solution to this problem involves the application of a 'Small Area' estimator, that is, an estimator using relevant auxiliary information to improve the precision of direct estimates (see Rao, 2003, for a general review). The auxiliary information may be exploited by specifying a (sometimes implicit) model that relates all the areas being studied.

In particular, in Fabrizi *et al.* (2007), we discuss several models within the class of 'unit level' Linear Mixed Models, where a linear relationship is assumed between the target variable and a set of auxiliary variables whose total is accurately known from the Census or some other sources, and random effects are introduced to model the correlation of residuals. In this approach, the models are linear for the equivalized household income considered on its original scale, and normality is assumed for the random effects and the residuals. In Fabrizi *et al.* (2007), we recognize that the normality assumption may not hold exactly for the considered data, but we find it to have a moderate impact on small area point predictors of equivalized mean income; moreover, provided that a robust strategy for the estimation of MSE is followed (for instance, the jackknife estimator of Jiang *et al.*, 2002), an estimate of MSE associated to these predictors with good properties may also be obtained.

Other authors (see for instance Elbers *et al.*, 2003) prefer to apply Linear Mixed Models to the log-transformation of income. This in principle may improve the fit of the models, but it has two related drawbacks: *i*) in order to predict area means or totals on the original scale of the study variable you need to back-transform individual predicted values, but the resulting prediction values will be biased (although several methods have been proposed to keep this bias low); *ii*) the prediction of individual values requires that the values of auxiliary variables are known for each member of the population outside the sample, whereas if the model is linear on the natural scale of the study variable, only the area means/totals of auxiliary variables are needed to predict area means/totals of the study variable.

In this paper, we do not consider this latter problem, we focus instead on the prediction of the mean of the equivalized income for a subset of the population by considering several alternative options. One is to consider a linear mixed model

on the natural scale of the equivalized income. To keep things simple, we will discuss the well-known nested error regression model introduced by Battese, Harter and Fuller (1988). Another is to also consider linear mixed models on the log-transformation of the equivalized income in association with different bias correction methods: naïve, smearing (Duan, 1983) and a ratio-adjusted-for-sample-total (RAST) method discussed in Chambers and Dorfman (2003). Finally, we also consider a Generalized Linear Mixed model, in which a more suitable distribution (positive and non symmetric) for the equivalized income is assumed conditionally on the auxiliary variables, the Gamma distribution.

The comparison of these options is based on a Monte Carlo exercise. To this purpose, the last wave (2001) of the ECHP survey is treated as a pseudo-population from which we bootstrap samples using the survey weights as the size variable. This solution may not be as good as that of using data from a real Census population, but it is hopefully more realistic than generating population values of household income from a parametric model.

The paper is organized as follows. Section 2 discusses the use of Mixed Linear Models for the small area estimation and describes the nested error regression model. Section 3 briefly reviews the ECHP survey and describes how we use this survey data to conduct the Monte Carlo simulation study. Section 4 presents distributional assumptions and predictors suggested as alternatives. Performances of the estimators derived from the proposed models are compared in section 5.

2. Linear mixed models in small area estimation

When the target parameter is an average or a total, Linear Mixed Models (LMM) are largely used. A brief description of LMM and the estimators they lead to is given below. For a more complete review of the application of this class of models in the context of small area estimation, see Rao (2003, ch. 5). A general linear mixed model can be described as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1 v_1 + ... + \mathbf{Z}_s v_s + \mathbf{e}\,, \tag{1}$$

where $\mathbf{y} = \left\{ y_{dj} \right\}$ is the *n*-vector of sample observations, *j* denotes the unit and *d* the small area ($j = 1,...,n_d$; $d = 1,...,m$), β a $p \times 1$ vector of regression coefficients, v_i is a $q_i \times 1$ vector of random effects ($i = 1,...,s$), $\mathbf{e} = \left\{ e_{dj} \right\}$ a vector of errors; $\mathbf{X}$ is assumed of rank *p*, $\mathbf{Z}_i = \left\{ \mathbf{z}_{idj}^T \right\}$ is a $n \times q_i$ matrix of the incidence of the *i*-th random effects. We assume that $E\left(v_i\right) = 0$, $V\left(v_i\right) = \mathbf{G}_i$, $E\left(\mathbf{e}\right) = 0$, $V\left(\mathbf{e}\right) = \mathbf{R}$ (all expectations are wrt. model (1)) and that $v_1,...,v_s,\mathbf{e}$ are mutually independent.

As a consequence, the variance-covariance matrix of $\mathbf{y}$ is given by:

$$\mathbf{V} = V(\mathbf{y}) = \sum_{i=1}^{s} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T + \mathbf{R} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R},$$

where $\mathbf{Z} = [\mathbf{Z}_1 | ... | \mathbf{Z}_s]$. It is usually assumed that matrixes $\mathbf{G}, \mathbf{R}$ depend on a *k*-vector of variance components ψ, and so we can write $\mathbf{V}(\psi) = \mathbf{Z}\mathbf{G}(\psi)\mathbf{Z}^T + \mathbf{R}(\psi)$.

Note that at the level of individual observations, the model (1) can be rewritten as $y_{dj} = \mathbf{x}_{dj}^T \beta + \mathbf{z}_{1dj}^T v_1 + ... + \mathbf{z}_{sdj}^T v_s + e_{dj}$.

In small area estimation, the aim is to predict scalar linear combinations of fixed and random effects of the type $\eta = \mathbf{m}^T \beta + \mathbf{k}^T v$ where $\mathbf{m}$ and $\mathbf{k}$ are $p \times 1$ and $q \times 1$ vectors respectively, with $q = \sum_i q_i$. The best linear unbiased predictor (BLUP) of η can be obtained by estimating fixed effects and "realized values" of random specific area effects by GLS method:

$$\tilde{\eta}^{BLUP}(\psi) = \mathbf{m}^T \tilde{\beta}(\psi) + \mathbf{k}^T \tilde{v}(\psi). \qquad (2)$$

When the variance components in ψ are unknown, they may be estimated from the data and substituted into (2), thus obtaining "empirical BLUP" $\tilde{\eta}^{EBLUP}(\hat{\psi}) = \mathbf{m}^T \hat{\beta}(\hat{\psi}) + \mathbf{k}^T \hat{v}(\hat{\psi})$ (see Rao, 2003, ch. 6, and Jiang and Lahiri, 2006, for details). As far as the estimation of ψ is concerned, a number of methods have been proposed in the literature, such as Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) which assume the normality of random terms, and the MINQUE proposed by Rao (1971) which is non-parametric. In the present work we have opted for the REML method, thus assuming normality.

One simple example within the class of Linear Mixed Models is given by the standard one-fold nested error linear regression model of Battese, Harter and Fuller (1988), which has been widely applied in the small area literature:

$$y_{dj} = \mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj} \qquad (3)$$

where y_{dj} is the *Y* value observed on unit *j* of area *d*, $\mathbf{x}_{dj}$ is the auxiliary vector for unit *j*, β is the fixed effects vector (common to all areas), α_d is the specific area *d* effect and e_{dj} is the residual term for unit *j*.

All random terms are assumed mutually independent and normally distributed with zero mean and constant variance:

$$\alpha_d \overset{ind}{\sim} N\left(0,\sigma_\alpha^2\right), \quad e_{dj} \overset{ind}{\sim} N\left(0,\sigma_e^2\right). \tag{4}$$

Therefore this random effects structure corresponds to the assumption of a constant covariance between units that belong to the same area. Note that it is a particular case of (1) obtained when $s=1$, $q_1=m$, $\mathbf{G}_1=\sigma_\alpha^2\mathbf{I}_m$ and $\mathbf{R}=\sigma_e^2\mathbf{I}_n$.

Model (3) — (4) will be considered as the benchmark in the comparison between alternative distributional assumptions on residuals. The EBLUP estimator of the small area mean ($\eta_d=\bar{\mathbf{x}}_d^T\beta+\alpha_d$) will be given by $\tilde{\eta}_d=\bar{\mathbf{x}}_d^T\tilde{\beta}+\tilde{\alpha}_d$, where fixed effects and "realized value" of random specific area effects are estimated as described above.

3. The simulation study based on the european community household panel data

We carried out a Monte Carlo simulation study using the last wave (2001) of the ECHP data available for Italy as our 'synthetic population'. The use of the sample as pseudo-population is necessitated by the fact that the information on household income is not collected by the Census. Nevertheless, this solution is hopefully more realistic than generating population values of household income from a parametric model. Similar simulation studies based on re-sampling can be found in Falorsi *et al*. (1999); Lehtonen *et al*. (2003); Singh *et al*. (1994).

We took the household to be the reference unit in our study. The households in the data set were selected from different strata (NUTS2 regions) and were given different weights that account for unequal selection probability, adjustments for non-response in the initial recruitment and subsequent attrition.

In our Monte Carlo experiment, samples were drawn with replacement from the ECHP data set using stratified probability proportional to size sampling, the size variable being given by survey weights. Strata were given by NUTS2 regions which also correspond our domain of interest, the 21 Italian Administrative Regions, and were therefore treated as 'fixed domains'. Moreover we note that replicated samples were drawn keeping the pseudo-population fixed, so that the simulation was aimed at evaluating the design-based properties of the estimators.

The size of replicated samples was fixed to $n=1{,}000$ (roughly 15% of the size of respondent households in the 2001 wave). The region-specific sample sizes we obtained ranged from 14 to 112, being on average equal to 48. The total number of simulated samples was set to 1,000. Monte Carlo errors associated with this number of replicates were small enough to ensure significance of all comparisons we discuss in section 5.

Our target variable was given by the total net household income equivalized with respect to household size and composition. Total net household income is obtained as the sum of net incomes of all members of the household. Equivalent

net income is calculated by dividing total net household income by equivalent household size according to the OECD scale used by Eurostat (which gives a weight of 1.0 to the first adult, 0.5 to the other persons aged 14 or over who are living in the household and 0.3 to children under the age of 14).

Regarding the characteristics of the obtained pseudo-population, its overall mean is 22,547 Euros and the coefficient of variation is 0.59. The distribution is positively skewed even though skewness is not extreme (skewness coefficient $\gamma_1 \cong 2.5$). The difference between mean and median is 9% of the mean. Looking at the different administrative regions, the small area averages of the target variable show very different values, thus reflecting the well-known regional disparities which characterised the country. For example, the highest regional average is about 70% higher than the lowest one. The variance varies among regions, increasing with the regional average, so that the coefficient of variation varies among small areas from a minimum of 0.28 to a maximum of 0.84. Also the skewness (γ_1 ranging from 0.1 to 4.6) shows that the distribution of our target variable is quite a bit different in different areas.

Of the many covariates available from the ECHP questionnaire, we considered only those for which area population means were available from the 2001 Italian Census results, because those means are necessary to calculate the EBLUP estimator. Thus the chosen covariates are as follows: the percentage of employed; the percentage of unemployed; the percentage of people with a high/medium/low level of education in the household; household typology (presence of children, presence of aged people, etc.); the number of rooms per-capita and the tenure status of the accommodation (rented, owned etc.).

The adjusted R^2 of the OLS regression is close to 0.35 in almost all repeated samples. This rather low figure is the result of the nature of the phenomenon under study (household income is not easy to predict) and the constraint represented by the need to include only those covariates for which the population total can be obtained from the Census.

4. Alternative predictors

We consider two classes of alternatives to the empirical best predictor (EBLUP) associated with the nested error regression model described in section 2: the first includes predictors based on the fitting of a nested error regression model onto the logarithm of the total net equivalized household income; whereas the second assumes that, conditionally on the covariates, the total net equivalized household income is Gamma distributed.

The logarithmic transformation is often used in models for income because the logarithm of values generated from a positively asymmetric distribution are generally more “normal” than the untransformed values. Also the Gamma model has often been considered for the study of income distribution. Reasons for its use

have been both theoretical (Mukerji, 1967) and practical, due to the better fit provided with respect to empirical distribution (Eltetö e Frigyes, 1968; Van Praag *et al*., 1983).

The two strategies are described in subsections 4.1 and 4.2.

4.1. Predictors based on modeling the logarithm of income

In this model the dependent variable is given by the (natural) logarithmic transformation of the household equivalised income:

$$z_{dj} = \log y_{dj} \qquad d = 1,...,m \quad j = 1,...,n_d$$

$$z_{dj} = \mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj} \tag{5}$$

$$\alpha_d \sim N\left(0, \sigma_\alpha^2\right) \quad e_{dj} \sim N\left(0, \sigma_e^2\right) \quad \alpha_d \perp e_{dj}\,. \tag{6}$$

The usual hypotheses of independence, homoschedasticity and normality of residuals hold so variance components are estimated using the REML technique. The 'naïve predictors' discussed in subsection 4.1.1 rely on this normality assumption. However, we have proof that this assumption does not hold exactly in the case of our data. To overcome this problem, non parametric solution that do not rely on this assumption have been proposed. We discuss two different options within this class in subsections 4.1.2 and 4.1.3.

4.1.1. Naïve predictor

The quantity to be estimated is given, for area *d*, by $\overline{y}_{U,d} = N_d^{-1} \sum_{j=1}^{N_d} y_{dj} = N_d^{-1} \sum_{j=1}^{N_d} \exp\left(z_{dj}\right)$. A simple back transformation of the empirical best linear unbiased predictor (EBLUP) $\hat{\overline{z}}_{U,d} = \overline{\mathbf{x}}_d^T \hat{\beta} + \hat{\alpha}_d$, i.e. $\exp\left(\hat{\overline{z}}_{U,d}\right)$ cannot be used since it would be severely biased. A slightly better predictor may be obtained as:

$$\hat{\overline{y}}'_{U,d} = \frac{\sum_{j \in s} y_{dj} + \sum_{j \notin s} \exp\left(\hat{z}_{dj}\right)}{N_d} \tag{7}$$

with $\hat{z}_{dj} = \mathbf{x}_{dj}^T \hat{\beta} + \hat{\alpha}_d$. However, also this predictor is biased low because, in general:

$$E\left\{\exp\left(\mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj}\right)\right\} \neq E\left\{\exp\left(\mathbf{x}_{dj}^T \beta + \alpha_d\right)\right\} \tag{8}$$

even when $E\left(e_{dj}\right)=0$.

In the literature, different strategies have been suggested to overcome this problem. Some of them keep the normality distribution assumption for the transformed variable, others escape this restriction. In the first group of methods there is, for instance, the naïve lognormal predictor (Chambers and Dorfman, 2003), which uses a first order bias correction. In the case of model (5) – (6) it becomes:

$$\hat{\bar{y}}'_{U,d}=\frac{\sum_{j\in s}y_{dj}+\sum_{j\notin s}\exp\left(\hat{z}_{dj}+\frac{\hat{V}\left(z_{dj}\right)}{2}\right)}{N_d} \tag{9}$$

where the estimated variance of z_{dj} is $\hat{V}\left(z_{dj}\right)=\hat{\sigma}_e^2+\hat{\sigma}_\alpha^2$. However this predictor is still biased ($O\left(n^{-2}\right)$ bias). It is possible to demonstrate that the correction for the negative bias leads to the overestimation of *Y* values (Chambers and Dorfman; 2003).

An alternative predictor, which is strongly based on the assumption of log-normality and characterized by the same order of bias as the previous one ($O\left(n^{-2}\right)$), has been discussed by Kalberg (2000).

4.1.2. The rast predictor

The aim of the method is to yield a predictor calibrated on the sample average (or total) of *Y*, that is, such that $\sum_{j\in s}y_{dj}=\sum_{j\in s}\hat{y}_{dj}$. The lognormal predictors discussed in the previous paragraph do not possess this property. The 'naïve' predictor (7) is modified so that it will possess this property (Chambers and Dorfam, 2003). It is necessary to find new formulas for GLS estimators of β and α_d so that:

$$\sum_{j\in s}y_{dj}=\sum_{j\in s}\exp\left(z^*_{dj}\right)=\sum_{j\in s}\exp\left(\mathbf{x}_{dj}^T\beta^*+\alpha_d^*\right).$$

It is easy to show that the equality holds by simply adding to the intercept a correction given by $\gamma\left(\hat{\beta},\hat{\alpha}_d\right)=\log\sum y_{dj}-\log\sum\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)$. Therefore, assuming *K* covariates, $\beta^*=\left(\hat{\beta}_0+\gamma\left(\hat{\beta}\right),\hat{\beta}_1,...,\hat{\beta}_K\right)^T$ and $\alpha_d^*=\hat{\alpha}_d$. The resulting predictor of the population mean is:

$$\hat{\bar{y}}_{U,d}^{RAST} = \frac{\sum_{j\in s} y_{dj} + \sum_{j\notin s} \exp\left(z_{dj}^{*}\right)}{N_d} = \frac{1}{N_d}\left(\sum_{j\in s} y_{dj} + \frac{\sum_{j\notin s} \exp\left(\sum_{k=1}^{K} \hat{\beta}_k x_{kdj}\right) \sum_{j\in s} y_{dj}}{\sum_{j\in s} \exp\left(\sum_{k=1}^{K} \hat{\beta}_k x_{kdj}\right)}\right) \qquad (10)$$

It is possible to note that in (10) both intercept and specific effects disappear, but their effect is taken into consideration in the estimation of the other coefficients. In effect, the ratio between the sample sum of *Y* values and the sample sum of their predictions is used to correct individual non sampled predictions.

4.1.3. The smearing predictor

With the aim of estimating the untransformed scale expectation $E\left(Y\middle|\mathbf{X}\right) = \int \exp\left(\mathbf{X}'\beta + \varepsilon\right) dF\left(\varepsilon\right)$, without knowing the error distribution function *F* or a reliable parametric form for it, Duan (1983) suggests substituting *F* by its empirical estimate $\hat{F}_n$ so as to obtain what she calls *smearing estimate*:

$$\hat{E}\left(Y\middle|\mathbf{X}\right) = \int \exp\left(\mathbf{X}^T\beta + e\right) d\hat{F}_n\left(e\right) = \frac{1}{n}\cdot\sum_{j\in s} \exp\left(\mathbf{X}^T\hat{\beta} + \hat{e}_j\right)$$

where the $\hat{e}_j$ are the sample residuals from the ordinary least squares fit of $\log y_j$ onto $\mathbf{x}_j$. Following this idea, for an arbitrary estimator $\hat{\beta}$ of β, the smearing predictor of the population mean may be written as:

$$\hat{\bar{y}}^{SMEARING} = \frac{\sum_{j\in s} y_j + \sum_{j\notin s} y_j^{*}}{N} = \frac{\sum_{j\in s} y_j + \sum_{j\notin s} n^{-1} \sum_{h\in s} \exp\left(\mathbf{x}_j^T\hat{\beta} + \hat{e}_h\right)}{N} =$$

$$= \frac{\sum_{j\in s} y_j + \sum_{j\notin s} \exp\left(\mathbf{x}_j^T\hat{\beta}\right)\cdot n^{-1} \sum_{h\in s} \exp\left(\hat{e}_h\right)}{N}$$

where the observations for the non sampled units are predicted by correcting the "naïve" back transformation by a factor given by the average of the sample residuals $sc\left(\mathbf{e}\right) = n^{-1}\sum_{j\in s} \exp\left(\hat{e}_j\right)$.

In the case of model (5) – (6), the smearing predictor for area *d* will be given by:

$$\hat{\bar{y}}_{U,d}^{SMEARING} = \frac{\sum_{j\in s} y_{dj} + n_d^{-1}\sum_{h\in s}\exp(\hat{e}_{dh})\cdot\sum_{j\notin s}\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)}{N_d} = \frac{\sum_{j\in s} y_{dj} + sc(\mathbf{e}_d)\cdot\sum_{j\notin s}\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)}{N_d} \quad (11)$$

where fixed and random effects are estimated by GLS as usual and the correction factor $sc(\mathbf{e}_d)$ is calculated as the average of the residuals within area *d*.

It is possible to show that the smearing predictor in (11) may be also obtained as:

$$\hat{\bar{y}}_{U,d}^{SMEARING} = \frac{1}{N_d}\left\{\sum_{j\in s} y_{dj} + \sum_{j\notin s}\exp\left(\sum_{k=1}^{K}\hat{\beta}_k x_{kdj}\right)\cdot\left(n_d^{-1}\sum_{h\in s}\frac{y_{dh}}{\exp\left(\sum_{k=1}^{K}\hat{\beta}_k x_{kdh}\right)}\right)\right\}. \quad (12)$$

Also in this expression, as happened for RAST predictor in (10), the intercept and the specific effects of the model disappear, but their effect is taken into consideration in the estimation of the other coefficients.

We note that the smearing method is based on the estimation of the distribution function of the study variable separately for each of the areas whose mean is being predicted. In general, these estimators will have good asymptotic properties but they may perform rather poorly when applied to small samples.

4.2. The predictor based on the gamma linear mixed model for income

Consider the following model:

$$y_{dj} \mid \alpha_d, \beta \sim Gamma\left(\nu, \nu/\mu_{dj}\right) \quad (13)$$

with $\mu_{dj} = \mathbf{x}_{dj}^T\beta + \alpha_d$. Since $E\left(y_{dj} \mid \alpha_d, \beta\right) = \mu_{dj}$ and $V\left(y_{dj} \mid \alpha_d, \beta\right) = \mu_{dj}^2/\nu$ we then have $CV\left(y_{dj} \mid \alpha_d, \beta\right) = \nu^{-1/2}$.

As a consequence, in our Gamma model we assume a constant coefficient of variation. This assumption may be useful in situations where the variance of the observations increases with the mean (McCullagh and Nelder, 1989, p. 285). This hypothesis appears very sensible in our case. It does not allow for a direct and immediate comparison either with the benchmark model (equations 3 – 4) or with the model for the logarithm of income whose residuals are assumed homoschedastic, but the aim of this work is to compare predictors rather than their related models.

The predictor of $\overline{y}_{U,d}$ associated with this model may be easily obtained as:

$$\hat{\overline{y}}_d = \frac{\sum_{j\in s} y_j + \sum_{j\notin s}\left(\overline{\mathbf{x}}_d^T\hat{\beta} + \hat{\alpha}_d\right)}{N_d}. \tag{14}$$

The estimates $\hat{\beta}$ and $\hat{\alpha}_d$ are obtained using the Maximum Likelihood method as implemented in the GLIMMIX procedure of SAS (SAS Institute, 2006). We note that this estimator differs from the EBLUP derived under the normality distribution assumption only in the variance and covariance matrix used to estimate β and α_d.

5. Results from the simulation study

In summary, the predictors for the regional averages of the equivalized per-capita income that we are going to compare in the simulation exercise are the following: *i*) the EBLUP obtained from the normal Linear Mixed Model of (3) and (4) (LMM); *ii*) the naïve and naïve lognormal predictors (respectively NAÏVE and NALOG); *iii*) the RAST and the SMEARING predictors obtained from the normal Linear Mixed Model for the logarithm of *Y* (say RAST and SMEAR); *iv*) the predictor obtained from the Linear Mixed Model for income with Gamma distributed observations (G).

The performances of estimators will be evaluated by averaging not only over the Monte Carlo replicates but also over the small areas, following an approach common in the literature (see Rao, 2003, section 7.2.6). In particular we chose to show three measures: the average absolute relative bias (*AARB*), the average relative bias (*AARB'*) and the average relative mean squared error (*ARMSE*):

$$\begin{aligned}
AARB &= m^{-1}\sum_{d=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)\right| \\
AARB' &= m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)\right\} \\
ARMSE &= m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)^2\right\}
\end{aligned} \tag{15}$$

where $\tilde{y}_{dr}$ is the prediction of the pseudo-population small area mean $\overline{y}_{U,d}$ obtained from the r^{th} simulated sample. Both *AARB* and *AARB'* summarize the bias of the predictors: *AARB* averages the size of the relative bias over the areas, while *AARB'*, which considers the sign of this bias, is helpful to understand

whether there is a systematic tendency of the predictors to under or overestimate the actual $\bar{y}_{U,d}$. The third indicator, *ARMSE*, is a measure of accuracy of the predictors.

Results related to these indicators are reported in Table 1. Besides the indicators of (15), the average relative efficiency

$$AEFF_{dir} = \frac{ARMSE_{*}}{ARMSE_{dir}} \tag{16}$$

is also reported. Note that $ARMSE_{dir}$ pertains to the direct estimator (i.e. the Horwitz-Thompson estimator of $\bar{y}_{U,d}$ based on the inverse inclusion probabilities) in order to make the evaluation of the advantages associated with the various model-based predictors more readily comparable. In (16), * stands for LMM, NAÏVE, NALOG, RAST, SMEAR or G.

From Table 1, we note how all the considered small area strategies lead to more efficient estimates, on average, than the direct estimator, except for the NAÏVE estimator as was expected. The gain in efficiency, calculating from $AEFF_{dir}$, is relevant for all the small area estimators and varies from 35% for the RAST predictor to 53% for the LMM estimator. Therefore, the EBLUP estimator derived from a normal Linear Mixed Model shows the best performance in terms of accuracy, followed by G. On the other hand, the approximation necessary to produce the parameters' estimates in their actual scale, when the logarithmic transformation is applied, causes a greater instability in the results which makes the NALOG, RAST and SMEAR predictors less reliable than the other two. We also note that the simpler NALOG is more efficient than the two non parametric solutions, RAST and SMEARING, even though it is more biased.

Table 1. Performance measures

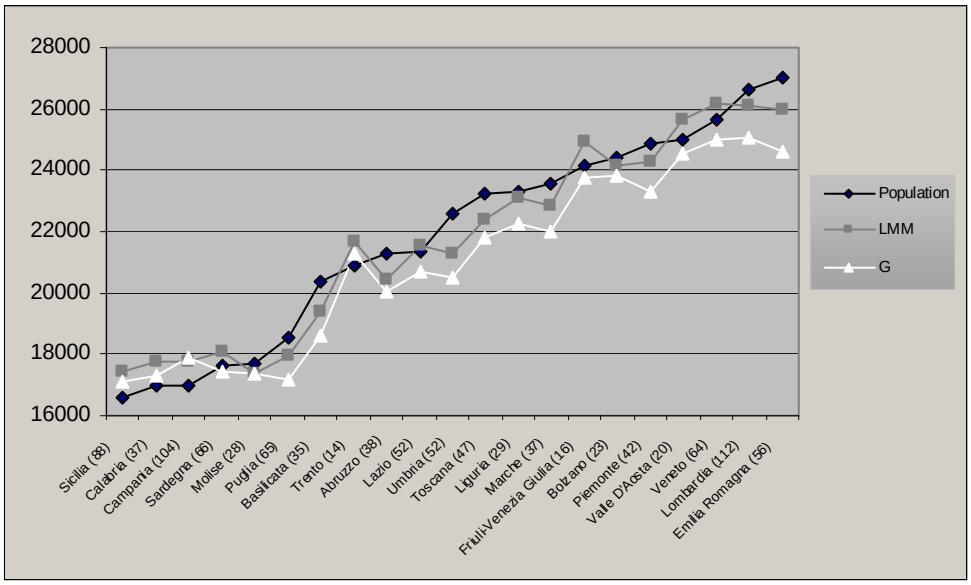
Estimator	*ARMSE%*	*AARB%*	*AARB%*	$AEFF_{Dir}$ %
Direct	0.787	0.0	0.0	100.0
NAÏVE	0.899	7.8	-7.8	114.2
NALOG	0.428	3.2	2.7	54.4
LMM	0.368	2.4	0.1	46.8
RAST	0.513	0.2	0.0	65.2
SMEAR	0.477	1.0	0.2	60.6
G	0.386	2.8	-1.2	49.0

Therefore, departure from normality seems to have a slight impact on punctual values of predictors. BLUP formulas can be derived without normality. Moreover, there are sound reasons to expect REML (and ML) estimators of variance components to perform well even if normality does not hold (see Jiang, 1996, for details).

Regarding the bias, very different results have been obtained for the different estimators. Looking at the *AARB* indicator, the least biased estimator is RAST even if, as we have already discussed, it is one of the less efficient ones. But this result is not surprising because, as explained in section 4, it is constructed as to control the transformation bias. Also the SMEAR estimator is not very biased, followed by LMM and then by G, which tends to underestimate the regional means (see *AARB´* column). Evidently the Gamma distribution is not completely suitable for the empirical income distribution that we have, even though the EBLUP derived by the Gamma model has the lowest variance. But the naïve estimators are the most biased. As expected, NAÏVE tends to seriously underestimate the parameters while NALOG corrects this bias but generates a positive bias, even though less important then the NAÏVE's one.

In order to further investigate the performance of the two preferable estimators (LMM and G), we looked at what happens in each region. To this purpose, in Figure 1, the means of the simulation replications obtained for them in the 21 regions are shown. To have a better view of the eventual effect on estimates of the value of the parameter, regions are ordered increasingly from the left to the right with respect to their population mean. This arrangement corresponds approximately to the regions arrangement form the south to the north because of the well-known greater incidence of poverty in the south of the country. We observe that, while the regional averages of the normal EBLUP are sometimes higher and sometimes lower than the population means, those related to the Gamma model are almost always lower except for the lowest levels of the population means. This means that G estimator tends to underestimate the high levels of income and to overestimate the low ones.

Figure 1. Regional means (sample size in brackets)



6. Conclusions

The main finding of the paper is that the empirical best predictor associated with the Battese Harter and Fuller model on the original scale of the study variable *Y* (given by equivalized household income) compares favourably to the back transformed predictors based on modelling the logarithmic transformation. On the one hand, this result is not completely surprising since, when obtaining the EBLUP, the normality assumption is used only in the REML estimation of the variance components, and this estimation method has been proven to be consistent even without normality (Jiang, 1996).

On the other hand, the result is relevant because the prediction of area means and totals using a linear mixed model on the original scale of *Y* requires that only the area means of the auxiliary variables are known, whereas methods considering the logarithmic transformation (as any other nonlinear transformation) need individual values for all units outside the sample. Moreover, as already noted in the introduction, in Fabrizi *et al.* (2007), we showed how the jackknife MSE estimator proposed by Jiang *et al.* (2002) is a good estimator of the design based MSE of this predictor. Besides, the extension of the work of Prasad and Rao (1990) and Datta and Lahiri (2000) to the estimation of MSE of predictors based on the non-linear transformations of the study variables have not been fully developed yet (see Slud, 2006 for more details).

The predictor based on the Gamma Generalized Linear Model, although in the case of our analysis it was outperformed by the Normal Linear Mixed Model, offers, at least in principle, an interesting alternative, since it does not require back-transformations and the MSE estimator of Jiang *et al.* (2002) may be applied.

Acknowledgements

Research was partially funded by Miur-PRIN 2003 "Statistical analysis of changes of the Italian productive sectors and their territorial structure", coordinator Prof. C. Filippucci. The work of Enrico Fabrizi was partially supported by the grants 60FABR06 and 60BIFF04, University of Bergamo.

We thank ISTAT for kindly providing the data used in this work.

REFERENCES

BATTESE G.E., HARTER R.M., FULLER W.A. (1988) An Error Component Model for Prediction of County Crop Areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28—36.

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.

CHAMBERS R.L., DORFMAN A.H. (2003) Transformed Variables in Survey Sampling, *Working paper* M03/21, Southampton Statistical Sciences Research Institute.

DATTA G.S., LAHIRI P. (2000) A Unified Measure of Uncertainty of Estimated of Best Linear Unbiased Predictors in Small Area Estimation Problems, *Statistica Sinica*, 10, 613—627.

DUAN N. (1983) Smearing Estimate: a Nonparametric Retransformation Method, *Journal of the American Statistical Association*, 78, 605—610.

ELBERS C., LANJOUW J.O., LANJOUW P. (2003) Micro-Level Estimation of Poverty and Inequality, *Econometrica*, 71, 355—364.

ÈLTETÖ O., FRYGIES E. (1968) New Income Inequality Measures as Efficient Tools for Causal Analysis and Planning, *Econometrica*, 36, 383—396.

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.

Eurostat (2005) *The continuity of indicators during the transition between ECHP and EU-SILC*, Working papers and studies, 2005 Edition.

FABRIZI E., FERRANTE M.R., PACEI S. (2007) Small Area Estimation of Average Household Income based on Unit Level Models for Panel Data, *Survey Methodology*, forthcoming.

FALORSI P.D., FALORSI S., RUSSO A. (1999) Small Area Estimation at Provincial Level in the Italian Labour Force Survey, *Journal of the Italian Statistical Society*, 1, 93—109.

JIANG J. (1996) REML Estimation: Asymptotic Behavior and Related Topics, *The Annals of Statistics*, 24, 255—286.

JIANG J., LAHIRI P., WAN S.M. (2002) A Unified Jackknife Theory for Empirical Best Predictor with M-estimation, *The Annals of Statistics*, 30, 1782—1810.

JIANG, J., LAHIRI, P. (2006) Mixed model prediction and small area estimation, Editor's invited discussion paper, *Test*, Vol. 15, 1, 1—96.

LEHTONEN R., SÄRNDAL C.-E., VEIJANEN A. (2003) The Effect of Model Choice in Estimation for Domains, Including Small Domains, *Survey Methodology*, 29, 1, 33—44.

KARLBERG F. (2000) Population Total Prediction Under a Lognormal Superpopulation model, *Metron*, 58, 53—80.

MCCULLAGH P., NELDER J.A. (1989) *Generalized Linear Models*, Chapman and Hall, London, England.

MUKERJI V. (1967) Type III Distribution and its Stochastic Evolution in the Context of Distributions of Income, Landholdings and Other Economic Variables, Sankhia, 29, A, 405—416.

PRASAD N., RAO J.N.K. (1990) Estimation of Mean-Squared Errors in Small Area Estimation, *Journal of the American Statistical Association*, 85, 163—171.

RAO C.R. (1971) Estimation of Variance Components – MINQUE Theory, Journal of Multivariate Analysis, 1, 257—275.

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

SAS Institute inc. (2006) The GLIMMIX Procedure. User Manual. June 2006. Downloadable at the address http://support.sas.com/rnd/app/papers/glimmix.pdf

SINGH A.C., MANTEL H.J., THOMAS B.W. (1994) Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20, 1, 33—43.

SLUD E.V., MAITI T. (2006) Mean-squared Error Estimation in Transformed Fay-Herriot Models, *Journal of the Royal Statistical Society, ser.* B, 68, 239—257.

STEWART K. (2003) Monitoring social exclusion in Europe's regions, *Journal of European Social Policy*, 13, 4, 335—356.

VAN PRAAG B., HAGENAARS A., VAN ECK W. (1983) The influence of classification and observation errors on the measurement of income inequality, *Econometrica*, 51, 1093—1108.