STATISTICS IN TRANSITION, December 2005
Vol. 7, No. 3, pp. 587—608

ESTIMATION OF POVERTY INDICATORS AT SUB-NATIONAL LEVEL USING MULTIVARIATE SMALL AREA MODELS

Enrico Fabrizi[1], Maria Rosaria Ferrante[2], Silvia Pacei[2]

ABSTRACT

In recent years, the measure of regional disparities on poverty and social exclusion has received increasing attention by policy makers and this has produced a growing demand of sub-national statistical information on income parameters. Taking into account the multidimensionality of this phenomenon, in the Laeken European Council (Eurostat, 2003) a set of financial poverty indicators were suggested to monitor the progress in fighting inequality. In the European Union regional statistical information on income can be obtained from the European Community Household Panel, a survey designed to provide reliable estimates for large regions in the countries. The aim of this work is to estimate at sub-national level some of the income indicators suggested in the Laeken council. We propose Bayesian small area estimators based on multivariate area level models exploiting the correlation between different indicators. The tendency of model based estimates to over-shrink towards the synthetic component can be a draw-back for policy makers interested in capturing regional disparities in financial poverty. To preserve the relationship between different indicators and disparities over areas, we adopt a multivariate constrained Bayes estimator. The comparison between results based on different models allows us to select the estimator realizing the best compromise between gain in efficiency and reduction of the over-shrinkage.

***Key words*:** Regional disparities, financial poverty measures, European Community Household Survey, multivariate Hierarchical Bayes model, over-shrinkage.

[1] E. Fabrizi, e-mail: enrico.fabrizi@unibg.it. DMSIA, University of Bergamo, via dei Caniana 2, 24127, Bergamo, Italy.

[2] M.R. Ferrante, e-mail: ferrante@stat.unibo.it (corresponding author); S. Pacei, e-mail: pacei@stat.unibo.it. Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy.

1. Introduction

In recent years, the assessment and the reduction of the territorial disparities in the distribution of income and the promotion of an homogeneous economical development have become a priority for the European Union (EU).

Since the focus of many policies in this field is regional, the need for reliable estimates of poverty indicators at sub-national level rose (European Commission, 2004).

In EU the measurement of income and poverty is based on the information collected in the European Community Households Panel (ECHP), an annual panel survey coordinated by Eurostat (Betti and Verma, 2002; Eurostat, 2002) covering a wide range of topics on economic, social and living conditions of households. From 2003 the ECHP is being replaced by a new survey, the "European Union — Statistics on Income and Living Condition" (EU-SILC). Unfortunately, EU-SILC data has not been published yet, at least for Italy. Due to the similarity between the two surveys, the proposals defined in this work could be used in the EU-SILC context too.

The ECHP is designed to provide estimates for large areas within countries, known as NUTS1 (NUTS is European Union's "Nomenclature of Units for Territorial Statistics"; see Eurostat, 2003 or

http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html).

Since this geographical detail is too large to meet the need of the policy makers in measuring income territorial disparities, Small Area Estimation tools have to be used to derive estimates of adequate precision at a finer level. Using information provided by the last wave of the ECHP survey (2001 year) referred to Italy, we propose a model based small area estimation strategy to obtain reliable estimates of poverty indicators for the 21 NUTS2 Italian regions.

Poverty is a multi-dimensional phenomenon which is better described by a set rather than a single indicator. Even focusing on financial poverty, to consider more than one indicator is advisable. In particular we focus on the most popular financial poverty indicators endorsed by the Laeken European Council: the *Per-Capita Income*, the *Poverty Threshold*, the *At-risk-of-poverty rate* based on a regional Poverty Threshold, the *At-risk-of-poverty rate* based on a national Poverty Threshold, the *Gini coefficient.*

To take advantage of the sampling correlations between these indicators we propose to use estimators based on aggregated multivariate small area models (Fay, 1987; Datta *et al*., 1991; Datta *et al*., 1996; Ghosh *et al*., 1996; Rao, 2003), that borrow strength not only across areas but also from correlations between survey estimates of different parameters.

Considering the complex survey design and the fact that some of the indicators considered are very complicated functions of data, to estimate

variances as well as covariance of the direct estimators we adopt a solution based on a bootstrap algorithm.

Small area estimators usually take advantage from covariate information known without uncertainty from Censuses or other administrative sources. Unfortunately in Italy, Censuses are conducted only once every ten years, and fiscal and administrative data are not available timely. For these reasons in this work, we introduce as covariate the regional unemployment rate estimates, produced by the Italian Statistical Institute. As these survey estimates are characterized by their own variability, we devise a method to incorporate it into the evaluation of the variability of the estimators we propose.

As estimation method we adopt a hierarchical Bayesian approach. In this field relatively complex models, as the multivariate one, can be easily implemented (approximation of posterior distributions via MCMC algorithms is computationally feasible when area-level models are carried out). Moreover posterior variances provide a natural measure of uncertainty accounting for all sources of uncertainty or of variation in the estimation process, as the sampling variability of the covariate. The consideration of all this features in a frequentist context could make difficult the derivation of the estimators and, above all, of their MSE.

In economic analysis of poverty the usual goal is to obtain estimates of the parameters for the whole ensemble of the areas, well representing the distribution of the parameters between areas. This request is generally displayed when small area estimates are used to develop economical policies at the local level or to plan resource allocation on areas when some parameters fall below some specified poverty threshold. To this aim we need a set of estimates with good "ensemble" properties. Unfortunately, most small area estimators, such as EBLUP or posterior means in Hierarchical Bayes models improve the precision of area-specific estimates at the price of shrinking estimates, with possible effect of making disparities look less than they really are (see Louis (1984), Ghosh (1992), Heady and Ralphs (2004), Zhang (2003)). To limit this loss of between-area variation and to preserve the covariance between estimates of different poverty indicators as close as possible to the covariance distribution referred to the parameters distribution, we adopt a multivariate constrained Bayes estimator (Ghosh and Maiti, 1999).

The outline of the paper is as follows. In section 2 the strategy adopted to derive direct estimates of the income indicators and their standard error is described. Sections 3 and 4 provide a description of the adopted multivariate hierarchical Bayes model and of its estimation. In section 5 the small area estimators' performances are compared in terms of efficiency and shrinkage. Section 6 gives a description of the constrained estimators and of their performance. In Section 7 some concluding remarks and possible further development of this work are presented.

2. Direct estimates derived from European Household Community Panel data

The target population of the ECHP survey is given by all resident households of EU member countries. Although general guidelines were issued by Eurostat for drawing the first wave sample, some flexibility was also allowed, so some differences in the design across countries exist. As far as Italy is concerned, the first wave's design is a stratified two stage design, in which strata were formed grouping the PSUs (municipalities) according to geographic region (NUTS2) and demographic size (for more details see Eurostat (2002)). The ECHP deals with unit non-response, sample attrition and new entries by providing weights and using imputation. As attrition could lead to biased estimates on income if it does not appear at random, the effect of poverty on dropout propensity has been investigated (Rendtel *et al.,* 2003; Vandecasteele and Debels, 2004). Results provided by these studies show that for some countries (including Italy), this effect disappears under control of weighting variables.

2.1. The financial poverty indicators

As anticipated, the aim of this work is to estimate at sub-national level some of the financial poverty indicators suggested in the Laeken council. The set of selected indicators, each of whom evaluating a particular aspect of the income distribution, allows us to obtain a picture of the characteristics of income distribution and of poverty's diffusion and intensity in each region.

All indicators are obtained on the basis of the "personal equivalent total net income", that in the following will be called simply "income" or "EQ_INC". It is calculated dividing household total net income by equivalent household size according to the OECD scale (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 aged less then 14). Consequently, the same equivalent total net income is assigned to each person in the same household.

A brief description of the indicators selected and their meaning is reported below, with reference to the i-th region (i =1,…,m). As the ECHP is a complex survey, sampling weight has to be considered.

1. *Average of EQ_INC* (PCI_i). It represents the personal equivalent total net income owned by each member of the household under the hypothesis of uniform distribution. Hence it does not take into account of the income distribution.
2. *Regional at risk of poverty threshold* (RPT_i). It is 60% regional median of personal equivalent total net income. It is an indicator of income distribution.
3. *At risk of poverty rate, regional threshold* ($ARPR1_i$). It is the share of persons with an equivalent total net income below 60% regional median income (RPT_i):

$$\text{ARPR1}_i = \frac{\sum_{j:\text{EQ\_INC}<\text{RPT}_i} w_{ij}}{\sum_{j \in s_i} w_{ij}}$$

where "s_i" denotes the sample of region i (i=1, …, m) and w_{ij} the sampling weight of the j-th unit belonging to the i-th region. It is an indicator of the intensity of the poverty.

4. *At risk of poverty rate, national threshold* (ARPR2_i). It is the share of persons with an equivalent total net income below 60% national median income (that is the National Poverty Threshold, NPT):

$$\text{ARPR2}_i = \frac{\sum_{j:\text{EQ\_INC}<\text{NPT}} w_{ij}}{\sum_{j \in s_i} w_{ij}}$$

Looking at this indicator, the regional intensity of poverty is compared on the basis of the same threshold for each region, while ARPR1_i provides a different information, because regional intensity of poverty is measured with reference to different regional thresholds.

5. *Gini coefficient* (G_i). It is the most commonly used concentration index. It expresses the relationship of cumulative shares of the population arranged according to the level of income (1 = poorest person and N_i = richest person belonging to the i-th region), to the cumulative share of the equivalent total net income received by them.

$$G_i = \left(\frac{2 \times \sum_{j=1}^{N_i} \left(w_{ij} \times \text{EQ\_INC}_{ij} \times \sum_{l=1}^{j} w_{il} \right) - \sum_{j=1}^{N_i} \left(w_{ij}^2 \times \text{EQ\_INC}_{ij} \right)}{\left(\sum_{l=1}^{j} w_{il} \right) \times \sum_{j=1}^{N_i} \left(w_{ij} \times \text{EQ\_INC}_{ij} \right)} - 1 \right) \times 100$$

The Gini coefficient measures the income concentration and varies between 0% and 100%. It is equal to 0% in the case of perfect equality and to 100% in the case of maximum concentration (all the regional income is owned by one person).

2.2. The bootstrap estimation of the direct estimates covariance matrix

The first step in our analysis is to evaluate the variability associated to direct estimates of the five financial poverty parameters previously described. This is made difficult by the complexity of the underlying sampling design and by the fact that some of the indicators are very complex functions of data (as the two rates ARPR1 and ARPR2 which depend on the sample based threshold RPT). Moreover, to adopt a multivariate model we are interested in estimating covariances as well as variances of estimators. For these reasons we opt for a

solution based on re-sampling algorithms. In particular, consistently with some other work in this field (Betti and Verma, 2004), we propose a bootstrap estimation strategy.

Standard bootstrap theory for stratified multistage designs (see Shao and Tu, 1995 chap. 6) requires that at least two PSUs are selected from each stratum either with replacement or without replacement provided that the sampling fraction is negligible.

In this case the bootstrap sample is obtained drawing a with-replacement random sample of $u_h - 1$ PSUs out the u_h sampled from each stratum ($h = 1,...,L$ $u_h \geq 2$). The bootstrap design weights are given by:

$$w_{hzj}(b) = w_{hzj} \frac{u_h}{u_h - 1} n_{hz}(b) \tag{1}$$

where w_{hzj} is the survey weight of secondary sampling unit *j* within PSU *z* in stratum *h* and $n_{hz}(b)$ is the number of times that PSU (hz) is selected into the bootstrap sample *b*.

For the particular case of one-stage stratified designs the bootstrap sample is obtained by drawing *L* independent random samples of size $n_h - 1$ (where n_h is the sample size in stratum *h*) from each stratum. Survey weights w_{hj} $j = 1,...,n_h$, $h = 1,...,L$ are then modified similarly to (1):

$$w_{hj}(b) = w_{hj} \frac{n_h}{n_h - 1} n_j(b) \tag{2}$$

where $n_j(b)$ is the number of times in which unit *j* within stratum *h* is included into the bootstrap sample.

The discussed assumptions are not met in the ECHP case, because the 23 large municipalities are sampled with probability 1, that is they form an auto-representative or certainty stratum.

The ECHP sampling design (for the first wave) can then be represented as mixed: households residing in large cities are stratified according to municipality and a (one-stage) stratified sample is drawn. The rest of the population is sampled according to a stratified two stage design in which PSUs are sampled without replacement but with negligible sampling fraction.

We propose to treat the two parts of the sample separately, extracting two sub-samples for each bootstrap replicate. In the first, secondary sampling units are selected according to a stratified design, and in the second municipalities are selected. Survey weights are modified according to (1) and (2) respectively. The two bootstrap samples are then merged, and a unique bootstrap samples is formed. This sampling procedure is replicated $R = 500$ times.

Let's denote with $\boldsymbol{\theta}$ the vector of the population quantities of interest and $\hat{\boldsymbol{\theta}}$ its design consistent estimator (described before) The weights defined in (1) and (4) are used to obtain the bootstrap estimate $\boldsymbol{\theta}_r^*$; the variance estimator of $\hat{\boldsymbol{\theta}}$ is obtained by:

$$\Sigma^{BOOT} = R^{-1}\sum_{r=1}^{R}\left(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}^*\right)\left(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}^*\right)' \tag{3}$$

where $\boldsymbol{\theta}^* = R^{-1}\sum_{r=1}^{R}\boldsymbol{\theta}_r^*$.

As it represents ad-hoc extension of a known estimator, formal results on the statistical properties of (3) are not available. By the way, for a survey with a similar design we found the estimates obtained by (3) consistent to those obtained by the linearization method (Ferrante *et al*., 2004).

2.3. Direct estimates

Regional direct estimates, their Coefficient of Variation (CV) and the sample dimensions associated to those estimates are reported in Table 1.

Focusing on the estimates' reliability, we may observe how it varies among different indicators. The estimates of ARPR1 and ARPR2 are the least efficient (the CV are on average equal to 0.25 and 0.28 respectively). This is probably due to a further source of sampling variability of the first two indicators, based on the estimates of RPT (at regional or national level). On the other hand, the estimates of PCI and RPT are the most reliable, and may be considered sufficiently efficient for various regions.

As already noted, to estimate a multivariate small area model we need the sampling correlation matrix between the different poverty indicators. From the Table 2, where the estimated correlation coefficients averaged over the areas are reported, we note that the correlation is large in almost all cases. This features should make multivariate model based estimators more effective in reducing the variability of direct estimators than those based on simpler univariate models.

Table 1. Direct estimates and their coefficients of variation

Region	**PCI**		**RPT**		**ARPR1**		**ARPR2**		**G**		**n**
	Estimate[a]	CV	Estimate[a]	CV	Estimate[b]	CV	Estimate[b]	CV	Estimate[b]	CV	
Piemonte	25.32	0.03	14.26	0.06	12.43	0.23	7.51	0.36	23.60	0.10	591
Valle D'Aosta	25.20	0.07	13.74	0.03	13.67	0.33	13.18	0.34	28.34	0.10	239
Liguria	23.08	0.06	12.94	0.05	11.49	0.29	9.43	0.28	22.18	0.10	441
Lombardia	26.98	0.03	14.92	0.03	15.76	0.17	6.28	0.35	25.03	0.07	1,515
Bolzano	25.12	0.10	13.82	0.06	12.83	0.39	9.73	0.41	26.12	0.11	369
Trento	22.00	0.07	13.81	0.09	18.63	0.43	16.98	0.52	21.66	0.20	201
Veneto	26.47	0.05	14.01	0.02	12.88	0.14	6.39	0.17	28.52	0.09	887
Fr.-V. Giulia	26.44	0.03	14.56	0.03	5.66	0.45	2.71	0.34	17.99	0.10	228
Em. Romagna	27.95	0.04	15.51	0.03	11.14	0.17	4.47	0.24	23.70	0.07	707
Toscana	24.32	0.05	13.92	0.06	14.64	0.28	9.09	0.56	23.59	0.15	657
Umbria	22.81	0.06	13.28	0.07	8.92	0.23	6.87	0.41	20.95	0.06	482
Marche	24.87	0.04	13.44	0.09	11.37	0.49	7.10	0.21	26.45	0.08	495
Lazio	19.92	0.05	10.91	0.05	19.55	0.17	24.62	0.17	30.54	0.08	849
Abruzzo	21.58	0.03	12.42	0.02	14.55	0.38	14.55	0.40	24.62	0.13	601
Molise	17.62	0.04	9.36	0.04	12.15	0.12	26.84	0.20	27.32	0.05	412
Campania	16.16	0.02	8.79	0.03	16.70	0.15	34.18	0.10	27.42	0.04	1,837
Puglia	16.49	0.06	8.64	0.08	20.78	0.15	37.70	0.16	32.61	0.06	1,337
Basilicata	19.37	0.10	10.56	0.12	20.30	0.29	26.60	0.36	29.32	0.09	696
Calabria	15.77	0.11	8.04	0.07	18.02	0.15	43.72	0.13	31.92	0.12	561
Sicilia	15.58	0.07	7.74	0.06	18.56	0.09	41.28	0.11	32.62	0.04	1,614
Sardegna	16.20	0.09	8.26	0.11	18.77	0.13	39.46	0.13	32.26	0.07	1,224
Min	*15.58*	*0.02*	*7.74*	*0.02*	*15.76*	*0.09*	*2.71*	*0.11*	*17.99*	*0.04*	*201*
Max	*27.95*	*0.11*	*15.51*	*0.12*	*20.78*	*0.49*	*43.72*	*0.56*	*32.92*	*0.20*	*1,837*
Average		*0.06*		*0.06*		*0.25*		*0.28*		*0.09*	

[a] *in thousands of euro,* [b] *in percentage*

Table 2. Estimated correlation matrix (average over the areas)

	PCI	RPT	ARPR1	ARPR2	G
PCI	1	0.590	0.486	0.447	0.396
RPT		1	0.610	0.618	0.612
ARPR1			1	0.384	0.413
ARPR2				1	0.756
G					1

The basic data to obtain model based estimates consists in the direct estimates (D) of the five parameter above presented, $\hat{\boldsymbol{\theta}}_i = \left(\hat{PCI}_i, \hat{RPT}_i, \hat{ARPR1}_i, \hat{ARPR2}_i, \hat{G}_i \right)^T$, and the associated covariance matrix $\boldsymbol{\Psi}_i$. For computational convenience it is useful to rescale all indicators in order to express them in about the same scale. For this reason we consider PCI, RPT in thousand of euros, the ratios and G in percentage.

3. The multivariate hierarchical Bayes model

To estimate the financial poverty indicators at NUTS2 level, we propose estimators based on multivariate area level models that relate small area survey estimates to area specific covariates and borrow strength not only of areas but also of the sampling correlation between survey estimates of different parameters. Similar models are considered in Datta *et al*., 1996 and in Ghosh *et al*., 1996.

Let $\boldsymbol{\theta}_i = \left(\theta_{i1}, ..., \theta_{ik}, ..., \theta_{iK} \right)^T$ be the vector of K parameters of interest for the i-th small area (i=1,…,m) and $\hat{\boldsymbol{\theta}}_i$ the corresponding vector of survey estimates. $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$ are linked by the following sampling model:

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \mathbf{e}_i \tag{4}$$

where the sampling errors $\mathbf{e}_i = \left(e_{i1}, ..., e_{ik}, ..., e_{iK} \right)^T$ are independent K-variate normal, $N_K \left(\mathbf{0}, \boldsymbol{\Psi}_i \right)$ with mean $\mathbf{0}$, the null vector, and known covariance matrix $\boldsymbol{\Psi}_i$ conditional on $\boldsymbol{\theta}_i$. Consistently with the bootstrap estimates described in previous section we assume $E\left(\mathbf{e}_i \mathbf{e}_{i*}^T \right) = \mathbf{0} \ \forall i \neq i*$. Normality of $\mathbf{e}_i$ is justified invoking central limit effects.

The $\boldsymbol{\theta}_i$'s is related to area specific auxiliary data through the following linking linear model:

$$\boldsymbol{\theta}_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{v}_i \tag{5}$$

where the area-specific random effects $\mathbf{v}_i$ are independent $N_K \left(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{v} \right)$, $\tilde{\mathbf{X}}_i$ is an $K \times KP$ matrix of the P covariates, $\boldsymbol{\beta}$ is an KP-vector of regression coefficient and $\boldsymbol{\Sigma}_v = diag\left(\sigma_{v,k}^2 \right)$ a matrix containing the variances of the area-specific effects (k=1,…,K). Normality of $\mathbf{v}_i$ is chosen for simplicity. In fact departures from Normality of random effects are difficult to check (see Sinharay and Stern, 2003) and, moreover, models with non-normal random effects are more complicated and difficult to estimate, particularly in the multivariate case.
The formulas (4) and (5) can be combined in:

$$\hat{\boldsymbol{\theta}}_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{v}_i + \mathbf{e}_i$$

Our model can be re-written in a more compact notation:

$$\text{(a)} \quad \hat{\boldsymbol{\theta}}_i \,|\, \boldsymbol{\theta}_i \overset{ind}{\sim} N_K\left(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i\right)$$

$$\text{(b)} \quad \boldsymbol{\theta}_i \,\big|\, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{v}} \overset{ind}{\sim} N_K\left(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{v}}\right)$$

In the following the prior specification for the described model is reported:

$$f\left(\boldsymbol{\beta}, \sigma^2_{v,k}\right) = f\left(\boldsymbol{\beta}\right) f\left(\sigma^2_{v,k}\right), \qquad k = 1,\ldots,K$$

$$\boldsymbol{\beta} \sim N(0, a_1 \mathbf{I}_{KP})$$

$$f\left(\sigma_{v,k}\right) \sim Unif\,(0, a_2)$$

where a_1, a_2 are large with respect to the scale of data, thus defining "diffuse proper priors". We select conjugate priors for the regression parameters in order to simplify Markov Chain computation and we choose $\sigma^2_{v,k}$ as uniform to avoid convergence problems of the Gibbs sampler. Moreover we adopt diffuse priors to reflect the lack of prior information about the model parameters.

For comparison purposes and mainly to evaluate possible advantages associated to a lower dimension of the parameter space, we consider a special case of the outlined model for which $\sigma^2_k = \sigma^2$ (MM1). This is made possible by the reduction of all indicators on a similar scale. The unrestricted model will be referred to as MM2. We also consider separate normal univariate Fay-Herriot models for each indicator:

$$\hat{\theta}_{ik} \sim N\left(\theta_{ik}, \psi_{i(kk)}\right)$$

$$\theta_{ik} \sim N\left(x_i\beta, \sigma^2_k\right)$$

Note that these univariate models (UM) overlook the sampling correlation existing between different indicators calculated for the same area.

As covariate we use the estimates of *regional unemployment rate*, obtained from the Italian Labour Force Survey. The covariate is the same for each of the parameters of interest because of its good or acceptable predictive power in each case (R^2 ranging from 0.36 to 0.88). As the covariate is a survey estimate, it is characterized by sampling variability that could increases the variance of the final estimates. In the HB approach this further source of uncertainty can be simply taken into account modelling the $\mathbf{X}_i$ as random, being $\tilde{\mathbf{X}}_i \sim N\left(\mathbf{X}_i, \hat{V}\left(\mathbf{X}_i\right)\right)$, with $V\left(\mathbf{X}_i\right)$ calculated according to Istat (2003).

4. The small area model based estimates

The approximate posterior distributions of the parameters of the proposed models and of the vector $\boldsymbol{\theta}_i$ are obtained using Monte Carlo integration via the Gibbs sampling algorithm. In particular we used the MCMC software WINBUGS (Spiegelhalter *et al.*, 1995). We run three parallel chains with 25,000 runs each, with starting point drawn from an over-dispersed distribution. The convergence of the Gibbs sampler is monitored by visual inspection of chains plots and of autocorrelation diagrams and by means of the Gelman and Rubin statistic (Gelman and Rubin, 1992). Although all models show fast convergence according to these monitoring tools, we discard the first 5,000 iteration from each chain. In multivariate models the quite strong autocorrelation of chains is reduced by thinning the chain (1 every 3 values are considered for posterior summaries). To check the adequacy of the specified models via posterior predictive checks, we consider the following discrepancy measure (Datta *et al*., 1999):

$$d\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\Psi}\right) = \sum_{i=1}^{m} \psi_{i(kk)}^{-1} \left(\hat{\theta}_{ik} - \theta_{ik}\right)^2$$

The fit of the model result to be adequate on average for all the three models, the posterior predictive checks ranging between 0.20 and 0.68, with the most suitable values for MM2.

The posterior means $\boldsymbol{\theta}_i^{HB} = E\left(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i, \boldsymbol{\Psi}_i\right)$ are taken as estimators of area parameters and the posterior variance $V\left(\boldsymbol{\theta}_i \mid \hat{\boldsymbol{\theta}}_i, \boldsymbol{\Psi}_i\right)$ as uncertainty measures.

5. The Bayes estimators performance

5.1. The gain in efficiency

Let $\hat{\theta}_{ikt}$ and $\hat{\theta}_{iks}$ be the estimates obtained from two different estimators, *t* and *s*, referred to the *i*-th area and the *k*-th parameter (*k*=PCI, RPT, ARPR1 ARPR2, G), being compared (*t,s*=D, UM, MM1, MM2). As measure of the improvement precision we propose the percent reduction of the Coefficient of Variation realized by the $\hat{\theta}_{ikt}$ versus the $\hat{\theta}_{iks}$, evaluated on average on areas ($ACVr_{k,t/s}$):

$$ACVr_{k,t/s} = 100 - \frac{ACV_{k,t}}{ACV_{k,s}} 100$$

where $ACV_{k,t} = \frac{1}{m}\sum_{i=1}^{m} CV\left(\hat{\theta}_{ik,t}\right)$ and $CV\left(\hat{\theta}_{ik,t}\right) = \frac{\sqrt{MSE\left(\hat{\theta}_{ik,t}\right)}}{\hat{\theta}_{ik,t}}$ (similarly for *s* estimator).

Table 3 contains the results concerning $ACVr_{k,t/s}$ for the five parameters being estimated.

Table 3. Average Coefficient of Variation reduction (*ACVr*)

	PCI	*RPT*	*ARPR1*	*ARPR2*	*G*
UM/D	20.2	22.8	40.9	16.1	23.5
MM1/D	41.2	39.9	47.4	48.5	40.1
MM2/D	38.5	40.0	45.2	34.4	37.0
MM1/UM	26.3	22.3	11.0	38.7	21.7
MM2/UM	23.0	22.3	7.4	21.8	17.7

As expected, all the model based estimators lead to a substantive gain in efficiency with respect to the direct one (D), the *ACVr* ranging from a minimum of 16.1% to a maximum of 48.5%. What is more interesting, multivariate model based estimators perform always better than the univariate one leading to a relevant further reduction of the coefficient of variation (from 7.4% to 38.7%).

We note that reduction of the coefficient of variation achieved by the multivariate models is noticeable in particular for the two indicators whose direct estimates show the greatest variability (ARPR1 and ARPR2). This result is probably due to the large correlation (Table 2) that estimates of both ARPR1 and ARPR2 show with those poverty indicators having a lower level of sampling variability (a.e. RPT and G). More in detail the use of multivariate models is very advantageous for ARPR1 estimates. In this case, in fact, due to the low predictive power of the covariate, the univariate model leads to a limited coefficient of variation reduction (*ACVr*=16.1%). Finally, the two multivariate model based estimates (MM1 and MM2) show comparable performances for all indicators except for ARPR2, where MM1 is slightly better.

In summary, multivariate models seem preferable than the univariate one on the whole and it is interesting to note that the consideration of direct estimates showing a low variability (as PCI and RPT) can be very helpful in a multivariate context. In fact these estimates can "give strength" to the estimates less reliable by means of their sampling correlation.

5.2. The evaluation of the shrinkage

As anticipated in the introduction, model based estimators such EBLUPs and Bayes estimators tend to "shrink" estimates towards the synthetic component, introducing bias. The loss of between-area variation of model-based estimates with respect to the true variance of underlying population parameters is known as over-shrinkage. In particular, as Ghosh (1992) pointed out, posterior means of

Hierarchical Bayes models over-shrinks estimates toward the prior expectation. On the other hand, direct estimates tend to show a larger variance than those of population parameters. With the purpose to select those model based estimators realizing the best compromise between gain in efficiency and capability to limit the shrinkage, at first we compare the model based estimator *t*, $\hat{\theta}_{ikt}$ (*t*=UM, MM1, MM2), with the direct ones, $\hat{\theta}_{ikD}$, calculated according to the methodology described in section 2. The comparison is carried out for *k*-th parameter on average on areas using $ASHR_{k,t}$ an indicator based on the relative absolute difference between direct and model based estimate, that provides an evaluation of the shrinkage connected with the last estimator:

$$ASHR_{k,t} = \frac{1}{m}\sum_{i=1}^{m} SHR_{ik,t}$$

where $SHR_{ik,t} = \dfrac{\left|\hat{\theta}_{ikt} - \hat{\theta}_{ikD}\right|}{\hat{\theta}_{ikD}} 100$.

To evaluate the impact of the model based small area estimators on the variance between areas, we propose to use the following indicator of Variance Reduction (*VARr*) based on the Ratio of the variance referred to the small area model based estimates *h,* and the same variance calculated on direct estimates D:

$$VARr_{k,t/D} = 100 - 100\frac{Var\left(\hat{\theta}_{ikt}\right)}{Var\left(\hat{\theta}_{ikD}\right)}$$

where $Var\left(\hat{\theta}_{ikt}\right) = \dfrac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_{ikt} - \hat{\bar{\theta}}_{.kt}\right)^2$ (similarly for the *D* estimates), being $\hat{\bar{\theta}}_{.kt}$ the average over areas.

Values of the average shrinkage indicators (*ASHR*) and of the Variance Reduction (*VARr*) are reported in the Table 4.

Table 4. Average shrinkage (*ASHR*) and Variance Reduction (*VARr*) of the model based estimates respect to direct ones

		PCI	*RPT*	*ARPR1*	*ARPR2*	*G*
ASHR	UM	2.2	2.1	13.5	9.4	4.6
	MM1	6.9	6.8	19.8	28.6	8.3
	MM2	4.3	4.3	15.2	17.0	6.4

VARr	UM	6.5	8.1	58.6	23.1	32.9
	MM1	52.1	45.3	92.7	76.7	74.8
	MM2	36.1	29.8	82.6	43.4	64.8

The higher the *ASHR* are, the more relevant is the shrinkage and then the bias connected to the model based estimates. In other words, thinking to the small area estimator as a composite estimator, that is a linear combination between a direct and a synthetic estimator, higher values for *ASHR* imply an higher weight given to the synthetic component. At first we note that on average for all the poverty indicators, MM1 leads to the greatest shrinkage, followed by MM2 and UM, being the former more flexible than MM1 and the latter more close to direct estimates. The difference between direct and model based estimates results more relevant in the case of the two poverty rates, ARPR1 and ARPR2: *ASHR*'s values respectively ranges from 13.5% (UM) to 19.8% (MM1), and from 9.4% (UM) to 28.6% (MM1). The difference with the direct estimates appears much less important for the other parameters, where *ASHR* does not overcome 10%.

As far as the reduction of the variability between-areas is concerned we note that, as expected, estimates obtained from UM, MM1 and MM2 show always a reduction of the variance between areas respect to the direct one. In particular MM1 produce the greatest variance reduction for the estimates of all the considered parameters, the *VARr* ranging from 45.3% to 92.7%.

In summary, the results obtained show that small area estimator based on MM2 realizes the best compromise between the gain in efficiency and the control of shrinkage. However, even for MM2 estimates the reduction of the variance between areas remains substantive. Considering that, we decide to tempt to improve the performance of the MM2 estimator affording the partially solved problems of the over-shrinkage by means of constrained estimators.

6. The multivariate constrained Bayes estimators

In dealing with the over-shrinkage we need to consider that our aim is to obtain a picture of a multidimensional phenomenon. In other words, beside the aim to obtain reliable estimates of each of the five parameters considered, we wish that variance between areas and also covariances between parameters would close as much as possible to those referred to the distribution of parameters.

To control shrinkage, Bayes estimators can be modified by making the empirical distribution function of the Bayes estimates close to the empirical distribution function of the unknown parameters. In this context, Ghosh (1992) proposed, under general assumptions, the "constrained Bayes estimators" obtained by matching the first two moments from the histogram of estimates with the corresponding moments from the posterior histogram of m normal means. Since we are estimating a multivariate parameter we adopt the multivariate Bayes constraining procedure that Ghosh and Maiti (1999) proposed by extending the

univariate one. Let $\mathbf{e}_1^B(\mathbf{x}),....,\mathbf{e}_m^B(\mathbf{x})$ the vector of the Bayes estimates based on data **x** under any quadratic loss. Note that $\mathbf{e}_i^B(\mathbf{x})$ represents the vector of the MM2 estimates for the *i*-th NUTS2. Let us write $E(\boldsymbol{\theta}_i \mid \mathbf{x}) = \mathbf{e}_i^B(\mathbf{x})$, *i*=1,…, *m* for simplicity and let $\overline{\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}_i$. We get that:

$$E\left[\sum_{i=1}^{m}(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})^T \mid \mathbf{x}\right] = \mathbf{H}_1(\mathbf{x}) + \mathbf{H}_2(\mathbf{x})$$

where $$\mathbf{H}_2(\mathbf{x}) = \sum_{i=1}^{m}\left[\mathbf{e}_i^B(\mathbf{x}) - \overline{\mathbf{e}}^B(\mathbf{x})\right]\left[\mathbf{e}_i^B(\mathbf{x}) - \overline{\mathbf{e}}^B(\mathbf{x})\right]^T$$

$$\overline{\mathbf{e}}^B(\mathbf{x}) = m^{-1}\sum_{i=1}^{m}\mathbf{e}_i^B(\mathbf{x}).$$

The objective is to find $\mathbf{t}_1,...,\mathbf{t}_m$ which minimize $E\left[\sum_{i=1}^{m}(\boldsymbol{\theta}_i - \mathbf{t}_i)(\boldsymbol{\theta}_i - \mathbf{t}_i)^T \mid \mathbf{x}\right]$ subject to the following two constraints:

$$E(\overline{\boldsymbol{\theta}} \mid \mathbf{x}) = m^{-1}\sum_{i=1}^{m}\mathbf{t}_i(\mathbf{x}) = \overline{\mathbf{t}}(\mathbf{x})$$

$$E\left[\sum_{i=1}^{m}(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})^T \mid \mathbf{x}\right] = \sum_{i=1}^{m}\left[\mathbf{t}_i(\mathbf{x}) - \overline{\mathbf{t}}(\mathbf{x})\right]\left[\mathbf{t}_i(\mathbf{x}) - \overline{\mathbf{t}}(\mathbf{x})\right]^T$$

The multivariate constrained Bayes estimator derived from this procedure is:

$$\mathbf{t}_i^{CB} = \overline{\mathbf{e}}^B(\mathbf{x}) + (\mathbf{H}_1 + \mathbf{H}_2)^{\frac{1}{2}}\mathbf{H}_2^{-\frac{1}{2}}\left[\mathbf{e}_i^B(\mathbf{x}) - \overline{\mathbf{e}}^B(\mathbf{x})\right]$$

with the following associated measure of uncertainty matrix:

$$E\left[(\boldsymbol{\theta}_i - \mathbf{t}_i^{CB})(\boldsymbol{\theta}_i - \mathbf{t}_i^{CB})^T \mid \mathbf{x}\right] = V\left[\boldsymbol{\theta}_i \mid \mathbf{x}\right] + \left[\mathbf{e}_i^B(\mathbf{x}) - \mathbf{t}_i^{CB}\right]\left[\mathbf{e}_i^B(\mathbf{x}) - \mathbf{t}_i^{CB}\right]^T$$

To evaluate the behaviour of the constrained estimates in comparative terms respect to the unconstrained MM2 estimates, we consider ratios already used in sect. 5.1 and 5.2. In Table 5, values of *ASHR,* of *VARr* and of *ACVr* referred to the multivariate constrained model based estimates (MM2C) respect to the direct ones are reported.

Table 5. Average Shrinkage (*ASHR*), Variance reduction (*VARr)* and Average Coefficient of Variation reduction (*ACVr*) of the constraint model based estimates (MM2C) respect to direct ones (D)

	PCI	*RPT*	*ARPR1*	*ARPR2*	*G*

ASHR	4.1	4.4	11.5	16.0	5.8
VARr	32.8	27.3	63.6	39.4	53.2
ACVr	35.2	34.9	38.1	32.3	33.0

The average shrinkage indicator (*ASHR*) referred to the constrained estimates (Table 5) shows values comparable to those observed for unconstrained MM2 estimates (Table 4). Looking at each poverty indicator, we note that ARPR1 estimates are influenced by the constraints more that others indicators, the shrinkage realized by MM2C being less heavy than that one carried out by MM2 (about 11% for the constrained and 15% for the unconstrained).

As expected, looking at the variance reduction we note that, in general, it is slightly lower than that realized by the MM2 estimates. Even in this case for ARPR1 a more marked influence of the constraining can be observed, being *VARr* equal to 63.6% for MM2C and 82.5% for MM2 estimates. The results singled out for each region (Fig. 1) confirm that constrained and unconstrained estimators have very similar behaviour, except for ARPR1, which show a not negligible difference between them for most of the regions.

Regarding the gain in efficiency, from the comparison of values on average reported in Table 3 (MM2/D row) with those reported in Table 5 we can note that, in general, the constrained estimates produce a reduction of the coefficient of variation (*ACVr*) slightly lower than that connected to MM2. Again this difference is more marked for ARPR1, being the *ACVr* connected to MM2 equal to 45% and that realized by MM2C equal to 38%. Considering regional values (Fig. 2), both model based estimates realize a marked coefficient of variation reduction respect to the D estimator in almost all regions. Moreover MM2C performs worse than MM2 for some regions. For completeness also the coefficient of variation of the standard Fay-Herriot model (UM) is reported in Fig.2.

This different behaviour of ARPR1 is likely due to the weak predictive power that the covariate has on it ($R^2 = 0.36$) and to the very low variability between areas of direct estimates, giving both these features major power to the constraint. On the other hand the constraining has a limited impact on estimates where the covariate is a good predictor.

Figure 1. NUTS2 regional estimates

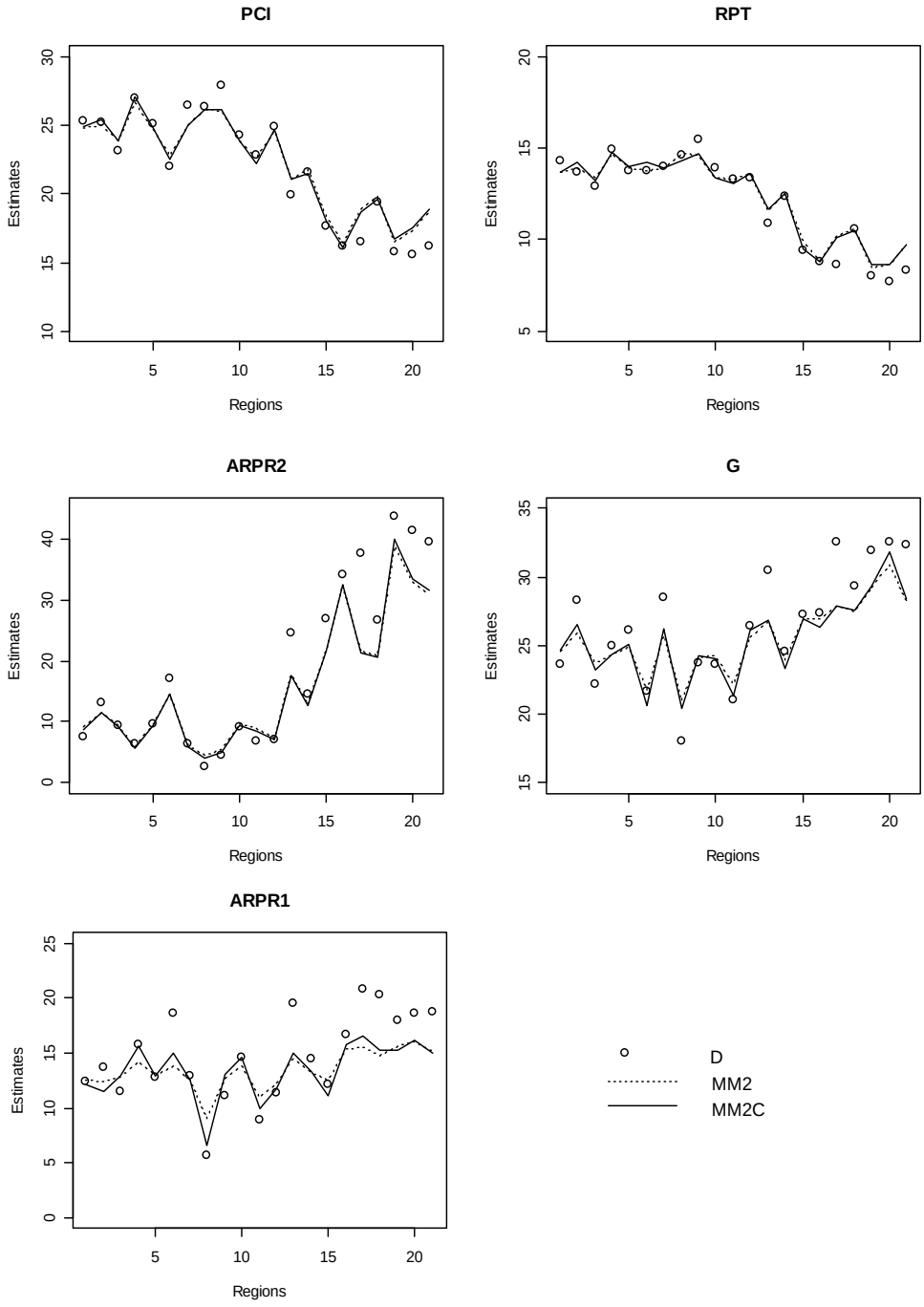
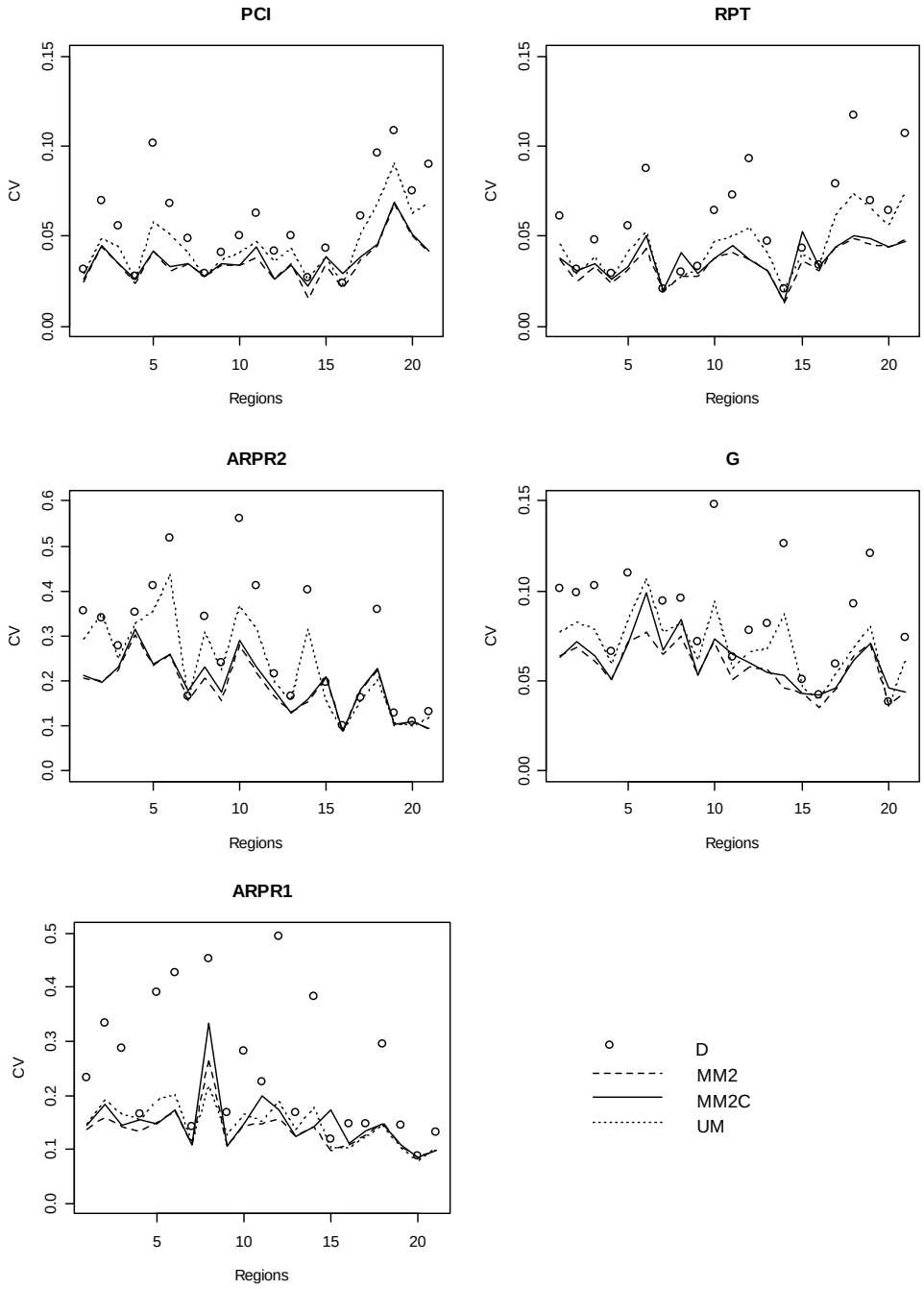


Figure 2. Coefficient of variation for NUTS2 regions



Finally, even if the MM2C estimates show a minor reduction of the coefficient of variation respect to MM2, it realizes the best compromise between gain in efficiency respect to the direct estimator and the need to limit the shrinkage respect to the standard Fay-Herriot model. Hence, we judge that it represent the best solution to obtain the NUTS2 estimates of the five poverty indicators considered. Results show that the estimation strategy adopted allows us to improve considerably the estimates reliability, assuming the coefficient of variation of the estimates acceptable values for almost all regions.

7. Concluding remark and further development

In this work we dealt with the problem of estimating a set of financial poverty indicators by means of Hierarchical Bayes estimators derived from a multivariate normal model. The results are encouraging, since the multivariate model adopted allow for a large gain in efficiency when compared to univariate Fay-Herriot estimators. Moreover, we note that the consideration of direct estimates showing a low variability (as PCI and RPT) can be very helpful in a multivariate context. In fact these estimates can give strength to the estimates less reliable by means of the sampling correlation between estimates.

As expected, estimates derived from the multivariate model adopted show more shrinkage than those obtained from the Fay-Herriot one. As we need to obtain suitable estimates for the analysis of regional disparities, we dealt also with the problem of over-shrinkage. In particular we implemented a multivariate constraint on the first two moments of the ensemble of HB estimates following to a methodology proposed by Ghosh and Maiti (1998). The impact of constraining varies across parameters and seems to be larger where the covariate used in the model has weaker predictive power.

Compared with the standard Fay-Herriot model, multivariate constrained Bayes estimators realize a very good compromise between the gain in efficiency and the need to limit the shrinkage.

By the way many problems are left to be further investigated. At first, the behaviour of the so called “triple goal estimators” (Shen and Louis, 1998) or of the “simultaneous estimators” proposed by Zhang (2003), could be evaluated as alternative solutions to deal with over-shrinkage. Moreover, as the ECHP is a panel survey, the suggested estimator could be modified to take advantage also of the repetition of the observations on the same areas for successive waves which may eventually lead to a further increment of the reliability of the small area estimates.

Regarding the estimated parameters, in this work we focused on some Laeken indicators which seemed to us able to provide a first picture of the characteristics of income distribution and of poverty’s diffusion in the regions of our country. This picture could be widened by considering other financial poverty indicators, maybe giving to the multivariate model more power in improving

reliability estimates. At the end, the strategy proposed could be extended to all the European regions.

Acknowledgements

Partially supported by Miur-PRIN "Statistical analysis on changes of the Italian productive sectors and their territorial structure — Coordinator Prof. C. Filippucci, 2003".

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

REFERENCES

DATTA, G. S., FAY, R. E. and GHOSH, M. (1991). Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, *Proceedings of Bureau of the Census 1991 Annual Research Conference, U. S. Bureau of the Census*, Washington, DC, 63—79.

DATTA, G. S., GHOSH, M., NANGIA, N., and NATARAJAN, K. (1996). Estimation of median income of four-person families: a Bayesian approach, *Bayesian Analysis in Statistics and Econometrics*, (D. A. Berry, K. M. Chaloner and J. M. Geweke Eds.), Wiley, New York, 129—140.

DATTA, G. S., LAHIRI, P., MAITI, T. and LU, K.L. (1999), Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S., *Journal of the American Statistical Association,* 94, 488, 1074—1082.

EUROPEAN COMMISSION (2004), *Third Report on Economic and Social Cohesion*, Brussels.

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 (2003), Laeken' Indicators — Detailed Calculation Methodology, *Working Paper of the Working Group "Statistics on Income. Poverty & Social Exclusion*", 28—29 April 2003.

FAY, R. E. (1987). Application of Multivariate Regression of Small Domain Estimation, in R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh (Eds.), *Small Area Statistics*, Wiley, New York, 91—102.

FERRANTE, M.R., PACEI, S. and FABRIZI E. (2004), *Estimation of household income frequency distribution in small areas,* paper presented in

IMS/ASA'S SRMS Joint Mini Meeting on Current Trends in Survey Sampling and Official Statistics, Calcutta, India, January 1—3, 2004.

GELMAN, A. and RUBIN, D. B. (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457—72.

GHOSH, M. (1992), Constrained Bayes Estimation with Application, *Journal of the American Statistical Association*, 87, 533—540.

GHOSH, M. and MAITI, T. (1999), Adjusted Bayes Estimators with Applications to small Area Estimation, *Sankhyā,* Series B, 61, 71—90.

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.

HEADY, P. and RALPHS, M. (2004), Some Findings of the Eurarea Project — and their Implications for Statistical Policy, *Statistics in Transition,* vol. 6, 5, 641—654.

ISTAT (2003), Forze di lavoro – Media 2002, *Annuari*.

JIANG, J., LAHIRI, P. and WAN, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *The Annals of Statistics*, 30, 1782—1810.

LOUIS T. A. (1984), Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods, *Journal of the American Statistical Association*, 79, 393—398.

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

RENDTEL, U., BEHR, A. and SISTO, J. (2003), Attrition effect in the European Community household panel, *CHINTEX PROJECT, European Commission*.

SHAO, J. and TU, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag New York, Inc.

SHEN, W., and LOUIS, T.A. (1998), Triple-goal Estimates in Two-stage Hierarchical Models*, Journal of the Royal Statistical Society, Series B,* 60, 455—471.

SINHARAY, S., and STERN, H. S. (2003), `Posterior Predictive Model Checking in Hierarchical Models, *Journal of Statistical Planning and Inference.*, 111, 209—221.

SPIEGELHALTER D. J., THOMAS, A., BEST, N. G., and GILKS, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50,* Medical Research Council Biostatistics Unit, Cambridge.

VANDECASTEELE, L. and DEBELS A. (2004), Modelling attrition in the European Community Household Panel: the effectiveness of weighting, *2nd International Conference of ECHP Users, EPUNet* 2004, Berlin, June 24—26.

VERMA, V. and BETTI, G. (2005), Sampling Errors for Measures of Inequality and Poverty, Invited paper in *Classification and Data Analysis 2005, Book of Short Papers*, CLADAG, 175—178.

ZHANG, L.C. (2003) Simultaneous estimation of mean of a binary variable from a large number of small areas, *Journal of Official Statistics*, 19, 253—263.