Multivariate Small Area Estimation of Multidimensional Latent Economic Wellbeing Indicators

Angelo Moretti[1], Natalie Shlomo[2] and Joseph W. Sakshaug[3]

1 (Corresponding author). University of Manchester, Social Statistics Department. Email: angelo.moretti@manchester.ac.uk or a.moretti2@outlook.com

2 University of Manchester, Social Statistics Department. Email: natalie.shlomo@manchester.ac.uk

3 German Institute for Employment Research, Nuremberg. Email: joe.sakshaug@iab.de

Abstract: Factor analysis (FA) models are used in data dimensionality reduction problems where the variability among observed variables can be described through a smaller number of unobserved latent variables. This approach is often used to estimate the multidimensionality of wellbeing. We employ FA models and use multivariate EBLUP (MEBLUP) to predict a vector of means of factor scores representing wellbeing for small areas. We compare this approach to the standard approach whereby we use SAE (univariate and multivariate) to estimate a dashboard of EBLUPs on original variables and then averaged. Our simulation study shows that the use of factor scores provides estimates with lower variability than weighted and simple averages of standardised MEBLUPs and univariate EBLUPs. Moreover, we find that when the correlation in the observed data is taken into account before small area estimates are computed multivariate modelling does not provide large improvements in the precision of the estimates over the univariate modelling. We close with an application using the EU Survey on Income and Living Conditions data.

Keywords: Factor analysis models; Latent variables, Model-based inference; Multivariate EBLUP; Multivariate multilevel models.

1. Introduction

The international scientific community, national statistical agencies, and international organisations have pointed out the multidimensional nature of wellbeing as developed under the UN initiative of the Sustainable Development Goals (United Nations, 2017). In particular, government agencies in European Union (EU) countries have been developing wellbeing measurement frameworks. One example is the Italian Statistical Institute (ISTAT) and National Council for Economics and Labour (CNEL) "Equitable and Sustainable Wellbeing (BES)" project. These frameworks generally consist of many dimensions (also called domains), each with many single indicators associated to them. To reduce data dimensionality, summary statistics in the form of a composite estimator may be helpful for policy makers to inform policies targeted to improving wellbeing. There is an ongoing debate about the appropriateness of using composite indicators versus a dashboard of single indicators: Ravallion (2011) points out that single multidimensional indicators lead to a loss of information, while Yalonetzky (2012), on the other hand, stresses that composite estimates are necessary when the goal is measuring multiple deprivations (or wellbeing) within the same unit (individual or household). In order to measure multidimensional wellbeing, analysing a dashboard of single indicators (means, totals, ratios, etc.) from the initial set of variables is a standard approach. However, if many indicators need to be analysed, the result may be difficult to interpret. Factor analysis (FA) models can be used to reduce data dimensionality and produce composite estimates. In these models, the variability among observed correlated variables is described through a smaller number of unobserved latent variables (factors).

In order to inform policies based on wellbeing measurement, there is a need to obtain reliable and accurate indicators at a local area level since wellbeing phenomena are *heterogeneous* and have different and varying features in territorial areas. This leads to the need for advanced statistical tools to provide reliable estimates of wellbeing (Lemmi et al., 2016) at the small area level. However, we face the issue of data reliability at local area levels since data on income, poverty, and quality of life

typically obtained from national surveys are usually not available or not reliable at a small area level. One way to overcome this problem is through model-based inference such as small area estimation (SAE) (Rao and Molina, 2015). Small area estimates 'borrow strength' from related small areas through the use of auxiliary variables available at the population level and other related (correlated) dependant variables. As an example, one of the most important social surveys available in EU countries for investigating social phenomena is the Statistics for Income and Living Conditions (EU-SILC). This data can be used to produce accurate direct estimates only at the Nomenclature of Territorial Units for Statistics (NUTS) 2 level (Giusti et al., 2012a) while any areas below this level are unplanned domains with small or even zero sample sizes.

Multivariate SAE is a research field still under investigation and there is an important gap about social exclusion and wellbeing measurement in a multivariate SAE setting. In the unit-level SAE approach, Fuller and Harter (1987) propose the use of multivariate mixed effects modelling in order to predict a vector of means of multiple characteristics of a finite population. Datta et al. (1999) develop a multivariate empirical best linear unbiased predictor (MEBLUP) and empirical bayes (EB) approach for small areas mean vectors. They also propose an approximation for the mean squared error and show a gain in efficiency obtainable by using multivariate mixed effects models compared to univariate models since the correlations between the vector components is taken into account. More recently, Molina (2009) deals with the multivariate mixed effects model under a logarithmic transformation, and Baillo and Molina (2009) studied a particular case of the multivariate nested error regression model for uncorrelated random effects.

In the classical univariate unit-level SAE approach, the use of the Battese, Harter, and Fuller (BHF) model (Battese et al., 1988) is widely used. The model is a mixed effects model and allows taking into account between-area variability in the prediction stage based on auxiliary information available for the population, such as a register or census. The BHF model can be naturally extended to the multivariate case, where a vector of K means becomes the new object of statistical inference.

Moretti et al. (2018b) evaluate the use of Factor Analysis (FA) models in SAE in order to reduce data dimensionality for economic wellbeing indicators and show that they can provide good estimates of multidimensional wellbeing phenomena at small areas. A dashboard of single indicators estimated at the small area level using a univariate SAE approach was compared to small area estimates of a single composite indicator arising from the FA model. They showed a gain in terms of the reduction in mean squared error when comparing the estimated mean factor scores with the use of an averaged dashboard of single indicators. According to the FA assumptions, the composite indicators derived from the latent factors are linearly related to the observed variables, and hence have the same economic interpretation (Moretti et al., 2018b). As mentioned, Moretti et al. (2018b) only consider a single latent variable. In this paper, we extend this work by studying the case of more than one latent factor using a multivariate empirical best linear unbiased predictor (MEBLUP) for factor score mean predictions. This new approach is compared to the averaging of dashboard small area estimates from the original variables using both a univariate and multivariate SAE approach.

In summary, this paper will investigate the following comparisons:

a) Comparison of EBLUP and MEBLUP of single observed response variables;
b) Comparison of EBLUP and MEBLUP of multidimensional latent factors as measured by factor scores;
c) Comparison of the use of latent factors in (b) to a dashboard of single observed response variables expressed as a simple or weighted average of standardised EBLUP and MEBLUP from (a).

This paper is organised as follows: in section 2 we introduce the multivariate SAE approach for a mean vector and review the multivariate EBLUP (MEBLUP) under the mixed effects model. In

section 3 we discuss the data dimensionality reduction problem via a Factor Analysis (FA) model. In section 4 we present a simulation study to evaluate our approach and address the comparisons (a) to (c) above. In section 5 we consider the multidimensionality issue of housing deprivation in Italy through an application using Italian EU-SILC data. We conclude our work in section 6 with a final discussion on the main findings and future work.

2. Multivariate Empirical Best Linear Unbiased Predictor (MEBLUP)

Let $d = 1, \dots, D$ denote the small areas for which we want to compute estimates, and let us consider a sample $s \subset \Omega$ of size n drawn from a target finite population Ω of size N. The non-sampled units, $N - n$, are denoted by r, hence, $s_d = s \cap \Omega_d$ is the sub-sample from the small area d of size n_d, $n = \sum_{d=1}^{D} n_d$, and $s = \cup_d s_d$. r_d denotes the non-sampled units for small area d of size $N_d - n_d$.

Considering $\boldsymbol{y}_{di} = (\boldsymbol{y}_{di1}, \dots, \boldsymbol{y}_{diK})'$, which denotes the $K \times 1$ vector of interest for $i = 1, \dots, N_d, d = 1, \dots, D$, we can write the target means vector as follows:

$$\overline{\boldsymbol{Y}}_d = N_d^{-1} \sum_{i=1}^{N_d} \boldsymbol{y}_{di}. \tag{1}$$

Hence, because of linearity of this quantity, each area means vector can be split into sampled and non-sampled (out-of-sample) elements as follows:

$$\overline{\boldsymbol{Y}}_d = N_d^{-1} \left(\sum_{i \in s_d} \boldsymbol{y}_{di} + \sum_{i \in r_d} \boldsymbol{y}_{di} \right). \tag{2}$$

The quantity $\sum_{i \in r_d} \boldsymbol{y}_{di}$ is not observed, so it needs to be predicted. In this work we propose the use of the multivariate mixed effects model, suggested in SAE by Fuller and Harter (1987) in order to predict the out-of-sample observations.

2.1 Multivariate nested-error linear regression model

We assume that unit-specific auxiliary variables $\boldsymbol{x}_{id}$ are available for all the population elements in

each small area *d* coming from a census or register. We also assume that the following linear model relates the response variables to the auxiliary variables as follows:

$$\boldsymbol{y}_{di} = \boldsymbol{x}_{di}\boldsymbol{\beta} + \boldsymbol{u}_d + \boldsymbol{e}_{di}, d = 1, \dots, D, i = 1, \dots, N_d,$$

$$\boldsymbol{u}_d \sim N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_u), \qquad \boldsymbol{e}_{di} \sim N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_e)\ \boldsymbol{u}_d \text{ and } \boldsymbol{e}_{di} \text{ independent} \tag{3}$$

where $\boldsymbol{x}_{di}$ is a *p*-dimensional row vector of auxiliary variables, $\boldsymbol{\beta}$ is a $p \times K$ matrix of unknown regression coefficients, $\boldsymbol{u}_d$ is a *K*-dimensional row vector of area effects, and $\boldsymbol{e}_{di}$ is *K*-dimensional row vector of the individual effects; $\boldsymbol{u}_d$ and $\boldsymbol{e}_{di}$ are assumed to be independent and normally distributed, N_K denotes a *K*-variate Normal distribution. Here, the $K \times K$ positive-definite matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are the variance-covariance matrices of the area effects and individual effects, respectively.

Under model (3) we can write the realised mean of area *d* as:

$$\bar{\boldsymbol{Y}}_d = \bar{\boldsymbol{X}}_d\boldsymbol{\beta} + \boldsymbol{u}_d \tag{4}$$

$\bar{\boldsymbol{X}}_d$ denotes the known population covariates means for area *d*.

2.2 Estimation and prediction of unknown parameters

For simplicity we now make use of the following notation (Fuller and Harter, 1987):

$$\boldsymbol{Y}' = (\boldsymbol{y}_{11}, \boldsymbol{y}_{12}, \dots, \boldsymbol{y}_{1,n_1}, \dots, \boldsymbol{y}_{D1}, \dots, \boldsymbol{y}_{D,n_D}),$$

$$\boldsymbol{X}' = [(\boldsymbol{I}_K \otimes \boldsymbol{x}_{11})', (\boldsymbol{I}_K \otimes \boldsymbol{x}_{12})', \dots, (\boldsymbol{I}_K \otimes \boldsymbol{x}_{1,n_1})', \dots, (\boldsymbol{I}_K \otimes \boldsymbol{x}_{D,n_D})'],$$

where $\boldsymbol{Y}$ denotes the vector of *NK* observations on $\boldsymbol{y}_{di}$ where $\boldsymbol{y}_{di}$ is defined above, and $\boldsymbol{X}$ denotes the $NK \times pK$ matrix of covariates. The operator $\otimes$ denotes the Kronecker product, and $\boldsymbol{I}$ denotes the identity matrix.

Let us now denote the covariance matrix of $\boldsymbol{Y}$ by

$$\boldsymbol{V}(\boldsymbol{Y}) = block\ diag(\boldsymbol{V}_{11}, \dots, \boldsymbol{V}_{DD}) \tag{5}$$

where $\boldsymbol{V}_{dd} = (\boldsymbol{J}_{dd} \otimes \boldsymbol{\Sigma}_u) + \left(\boldsymbol{I}_{n_d} \otimes \boldsymbol{\Sigma}_e\right)$. $\boldsymbol{J}_{dd}$ is the $n_d \times n_d$ matrix with every element equal to one and $\boldsymbol{I}_{n_d}$ is an identity matrix. The operator $\otimes$ denotes the Kronecker product. The empirical best linear unbiased estimator of the regression coefficients is given by:

$$\text{vec}\ \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{6}$$

The empirical best linear unbiased predictors of the random effects are given by the following expression

$$\widehat{\boldsymbol{u}}_d = (\overline{\boldsymbol{Y}}_d - \overline{\boldsymbol{x}}_d\widehat{\boldsymbol{\beta}})[\left(\widehat{\boldsymbol{\Sigma}}_u + n_d^{-1}\widehat{\boldsymbol{\Sigma}}_e\right)^{-1}\widehat{\boldsymbol{\Sigma}}_u, d = 1, \dots, D \tag{7}$$

$\widehat{\boldsymbol{\Sigma}}_u$ and $\widehat{\boldsymbol{\Sigma}}_e$ are estimators of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ (we refer to Schafer and Yucel (2002) for the algorithm and its implementation).

The Multivariate Empirical Best Linear Unbiased Predictor (MEBLUP) of $\overline{\boldsymbol{Y}}_d$ is given by:

$$\widehat{\overline{\boldsymbol{Y}}}_d^{\ MEBLUP} = \overline{\boldsymbol{X}}_d\,\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_d = \overline{\boldsymbol{X}}_d\,\widehat{\boldsymbol{\beta}} + (\overline{\boldsymbol{Y}}_d - \overline{\boldsymbol{x}}_d\widehat{\boldsymbol{\beta}})[\left(\widehat{\boldsymbol{\Sigma}}_u + n_d^{-1}\widehat{\boldsymbol{\Sigma}}_e\right)^{-1}\widehat{\boldsymbol{\Sigma}}_u],\ d = 1, \dots, D. \tag{8}$$

where $\overline{\boldsymbol{x}}_d$ denotes the sample auxiliary means for area d. In case of areas with $n_d = 0$ it holds that $\widehat{\overline{\boldsymbol{Y}}}_d^{MEBLUP} = \widehat{\overline{\boldsymbol{Y}}}_d^{Synthetic} = \overline{\boldsymbol{X}}_d\,\widehat{\boldsymbol{\beta}}$.

The mean squared error of (8) can be estimated via parametric bootstrap proposed by Moretti et al. (2018a). The mean squared error of $\widehat{\overline{\boldsymbol{Y}}}_d^{Synthetic}$ is given by the error due to prediction as in usual regression models, and it can be approximated via bootstrap using only the synthetic part of the model. For a complete discussion on techniques to estimate the MSE we refer to Rao and Molina (2015).

3. Data dimensionality reduction and the use of factor scores

Composite indicators are measures for multidimensional phenomena that cannot be studied by the use of single indicators. Due to their complexity, composite indicators should be based on theoretical frameworks and/or definitions to combine single indicators in a way which reflects the phenomena structure (OECD, 2004). A vast literature on multivariate statistical analysis techniques is available; for a formal review on the main methods we refer to Hardle and Simar (2012). In this paper we assume that latent constructs exist for a wellbeing domain and use FA methods to reduce the data dimensionality from the original variables.

3.1. The factor analysis model

Let us consider a $K \times 1$ vector of observed variables $\boldsymbol{Y}$ and we assume that they are linearly dependent on a vector of factors $\boldsymbol{f}$, with dimension $M \times 1$ ($M<K$). Thus, we can write the following linking model (Kaplan, 2009):

$$\boldsymbol{Y} = \boldsymbol{\Lambda f} + \boldsymbol{\epsilon} \tag{10}$$

where $\boldsymbol{\epsilon}$ denotes a vector $K \times 1$ containing both measurement and specific errors, and $\boldsymbol{\Lambda}$ is a $K \times M$ matrix of factor loadings.

It is assumed that:

i) $E(\boldsymbol{\epsilon}) = \mathbf{0}$,

ii) $E(\boldsymbol{f}) = \mathbf{0}$,

iii) $Cov(\boldsymbol{\epsilon}, \boldsymbol{f}) = \mathbf{0}$.

Therefore, the covariance matrix of the observed data is given by:

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{YY}') = \boldsymbol{\Lambda} E(\boldsymbol{ff}')\boldsymbol{\Lambda}' + E(\boldsymbol{\epsilon\epsilon}') = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}, \tag{11}$$

where $\boldsymbol{\Phi}$ is a $M \times M$ matrix of factor variances and covariances, and $\boldsymbol{\Theta}$ is a $K \times K$ diagonal matrix of specific variances.

The maximum-likelihood (ML) approach is used to estimate the model parameters. ML equations under FA models are complicated to solve, so iterative numerical algorithms are proposed in the literature (see e.g. Mardia et al., 1979). The log-likelihood function ℓ of the data $\boldsymbol{Y}$ can be written as follows (Hardle and Simar, 2012):

$$\ell(\boldsymbol{Y}; \boldsymbol{\Lambda}, \boldsymbol{\Theta}) = \frac{n}{2}\left[\log\{|2\pi(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta})|\} + \mathrm{tr}\left\{(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta})^{-1}\widehat{\boldsymbol{\Sigma}}\right\}\right], \tag{12}$$

where $\widehat{\boldsymbol{\Sigma}}$ denotes the empirical covariance of $\boldsymbol{Y}$ (estimator of $\boldsymbol{\Sigma}$).

After the model parameters are estimated, the factor scores are also estimated. Factor scores are defined as estimates of the unobserved latent variables for each unit *i.* For a review of estimated factor scores we refer to Johnson and Wichern (1998).

Using the regression method, the individual factor scores estimate for $i = 1, \dots, n$ are given by (Hardle and Simar, 2012) where $\widehat{\boldsymbol{\Lambda}}$ denotes the estimator of $\boldsymbol{\Lambda}$:

$$\hat{\boldsymbol{f}}_i = \widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}_i. \tag{13}$$

In the presence of both binary and continuous observed variables under a maximum likelihood estimation approach, the factor scores may be estimated via the expected posterior method described in Muthén (2004) and applied in Mplus7.4 (Muthén and Muthén, 2012).

4. Simulation study

This simulation study is designed to assess the feasibility of the multivariate MEBLUP compared to the univariate EBLUP when considering the problem of data dimensionality reduction and the comparisons (a) to (c) mentioned in the introduction.

The overall results of the simulation study are evaluated via the empirical root mean squared error (RMSE) described in Section 4.2.

4.1. Generating the population

We generate a single population with $N = 20{,}000$, $D = 80$, and $130 \leq N_d \leq 420$. N_d are generated from the discrete uniform distribution, $N_d \sim \mathcal{U}(a = 130, b = 420)$ with $\sum_{d=1}^{D} N_d = 20{,}000$. $\boldsymbol{y}_{di}$ observations are generated according to the multivariate mixed effects model shown in formula (3). The simulation parameters $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\beta}$ are estimated from real Australian Agricultural and Grazing Industries Survey data (Australia, Bureau of Agricultural Economics, 1978; Molina, 2009). We define the following covariance matrix $\boldsymbol{\Sigma}_e$:

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} 0.386 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & 0.414 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & 0.213 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & 0.301 \end{bmatrix}.$$

Let r_u and r_e denote the correlation coefficients associated with the covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ respectively. Hence, σ_{lj} with $l \neq j$ in $\boldsymbol{\Sigma}_e$ varies according to r_e. For example, $\sigma_{12} = r_e\sqrt{0.386 \cdot 0.414}$ in the above matrix $\boldsymbol{\Sigma}_e$. The intra-class correlation coefficients are fixed as follows: $ICC_k = \{0.05, 0.1, 0.3\}$. Therefore the variances of $\boldsymbol{\Sigma}_u$ are generated as functions of the variances of $\boldsymbol{\Sigma}_e$ as follows: $ICC_k = \sigma^2_{uy_k}/(\sigma^2_{uy_k} + \sigma^2_{ey_k})$, where *k=1,…,4* denote the k^{th} component of $\boldsymbol{y}_{di}$.The covariances for $\boldsymbol{\Sigma}_u$ are then calculated as described above for $\boldsymbol{\Sigma}_e$.

In this simulation we study the following combinations of r_u and r_e: $r_u = r_e = 0.2$, $r_u = 0.2$ and $r_e = 0.7$, $r_u = -0.2$ and $r_e = 0.7$.

The $\boldsymbol{\beta}$ regression coefficients matrix is given by the following:

$$\boldsymbol{\beta} = \begin{bmatrix} 1.001 & 0.386 & 0.141 \\ 1.187 & 0.377 & 0.133 \\ 1.086 & 0.035 & 0.024 \\ 0.114 & 0.009 & 0.002 \end{bmatrix}.$$

The uncorrelated covariates are generated from discrete Uniform distributions, $X_1 \sim dUnif(20000, 145{,}459)$, $X_2 \sim dUnif(20000{,}55{,}345)$.

On the generated population, we run two Confirmatory Factor Analysis (CFA) models described in

(6): the first model for one latent factor and the second model for two latent factors. This is based on an initial explanatory analysis where we identified that both CFA models provide a good fit to the generated population. We show in Appendix A the goodness of fit statistics of the two CFA models on the generated population for the simulation study.

Figure 1 shows how the factors relate to the observed variables for the case of two latent factors in the CFA model. For each latent factor in both CFA models, we estimate the population factor scores $\boldsymbol{f}_i, i = 1, \dots, N$ from (13).

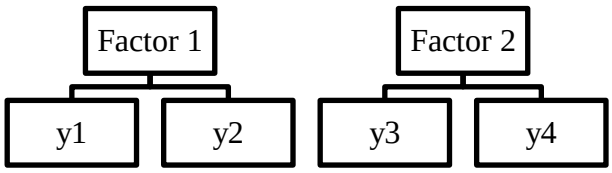

Figure 1 Relationship between the factors and observed variables two-factor CFA.

As mentioned in Moretti et al. (2018b), although FA models have been developed to account for multilevel structures, it is not possible to fit these models for *unplanned domains* given small and zero sample size domains. We leave this for future work.

We also calculate the following true values based on the generated population for each of the small areas *d:* the factor score means, simple averages of the standardised observed variable means, and weighted averages using the CFA loadings denoted by $\bar{Y}_{dm}^{S\_Averages}$ and $\bar{Y}_{dm}^{W\_Averages}$, respectively, where m denotes the m^{th} factor and the averages are taken over those variables associated to the m^{th} factor. The true means are calculated from the generated population to be used in evaluations of the RMSE and BIAS.

For example, the weighted average (based on the factor loadings) of standardised EBLUPs (which have been transformed with zero mean and unit variance) for area d for the variables $k = 1, \dots, K$ that contribute to the m^{th} factor is given by:

$$\hat{\bar{Y}}_{dm}^{EBLUP\_W\_Averages} = \frac{\sum_{k=1}^{K}\left(\hat{\bar{Y}}_{dk}^{standard\_EBLUP}\lambda_{km}\right)}{\sum_{k=1}^{K}\hat{\lambda}_{km}}, d = 1, \ldots, D, m = 1, \ldots, M \qquad (14)$$

where $\hat{\lambda}_k$ is the estimated factor loading for variable *k* related to factor *m*.

4.2. Simulation steps

1. Draw $S = 500$ samples using simple random sampling without replacement from the generated population;
2. Fit the one-factor and two-factor CFA model on each sample and estimate the EBLUP factor score means from each model for each area *d* in each sample. In addition to the separate EBLUP factor score means for each of the factors under the two-factor CFA model, estimate the MEBLUP factor score means;
3. The EBLUP and MEBLUP for each of the observed variables and vectors ***Y*** are also estimated in order to construct simple averages of the standardised small area EBLUPs and MEBLUPs, and a weighted average using the factor loadings estimated in 2;
4. As the true values are known from the generated population, we can calculate the root mean squared error and the bias for each area *d* for the different types of estimates: EBLUPs and MEBLUPs of factor score means; and simple and weighted averages of EBLUPs and MEBLUPs. For example, for the univariate EBLUPs of the observed variable mean *k,* the root mean squared error is:

$$RMSE\left(\hat{\bar{Y}}_{dk}^{EBLUP}\right) = \sqrt{S^{-1}\sum_{s=1}^{S}\left(\hat{\bar{Y}}_{dks}^{EBLUP} - \bar{Y}_{dk}^{TRUE}\right)^2} \qquad (15)$$

where $\bar{Y}_{dk}^{TRUE}$ denotes the true mean of the Y_k variable for the d^{th} area.

4.3. Results of the simulation study

4.3.1. Comparison (a) of EBLUP and MEBLUP of single observed response variables

Table 1 shows the percentage relative reduction (in terms of RMSE) of the multivariate MEBLUP over the univariate EBLUP under comparison (a) for single observed response variables. The percentage relative reduction for each area is calculated as follows $\Delta_{dk} = \frac{RMSE(\hat{Y}_{dk}^{MEBLUP}) - RMSE(\hat{Y}_{dk}^{EBLUP})}{RMSE(\hat{Y}_{dk}^{EBLUP})} \cdot 100$, $k = 1, \dots, K$, *d=1,...,D.* Δ_{dk} estimates are then averaged across the areas to provide summary statistics for each variable *k*: $\Delta_k = \frac{1}{D}\sum_d \Delta_{dk}$.

		Scenario		
$\boldsymbol{ICC_k}$		$r_e = 0.7, r_u = 0.2$	$r_e = 0.7, r_u = -0.2$	$r_e = 0.2, r_u = 0.2$
0.05	***y1***	-3.50	-9.21	-1.04
	y2	-3.00	-10.81	-1.02
	y3	-3.00	-12.22	-0.30
	y4	-2.00	-12.01	0.00
0.1	***y1***	-6.00	-18.42	-0.31
	y2	-3.41	-18.33	-0.20
	y3	-6.00	-19.20	-0.03
	y4	-6.02	-16.90	-0.09
0.3	***y1***	-8.00	-20.00	0.00
	y2	-7.51	-19.20	0.00
	y3	-7.03	-21.11	0.00
	y4	-6.52	-18.90	0.00

Table 1 Percentage relative reduction (%) in RMSE of MEBLUP over EBLUP (Δ_k) for single observed response variables averaged over all areas

When the correlations r_e and r_u are equal to 0.2, we see that the MEBLUP does not provide much improvement over the univariate EBLUP. Indeed, when r_e and r_u tend to 0 we are close to the independence case, whereby univariate analysis provide the same results as the multivariate one (Datta et al., 1999). When correlation coefficients associated to $\boldsymbol{\Sigma}_e$ are large, MEBLUP provides more efficient predictions than EBLUP. As it has already been noted by Datta et al. (1999), these gains tend to become large when the signs of r_e and r_u are opposite. The gains in efficiency are good even when the intra-class correlation is low, although we have bigger improvements with

respect to the RMSE when the intra-class correlation increases.

4.3.2. Comparison (b) of EBLUP and MEBLUP of multidimensional latent factors (two-factor CFA model) as measured by factor scores

Table 2 shows the estimates of the correlation terms and the intra-class correlations resulting from the multivariate modelling of latent factors that were estimated by the two-factor CFA model. It can be seen that the estimated correlation terms and *ICC* of the two latent factors increase compared to the correlation structure of the original variables when $r_e = 0.7, r_u = 0.2$ and $r_e = 0.7, r_u = -0.2$. Under the case $r_e = 0.2,\ r_u = 0.2$ there are mixed results for the correlation term of $r_u$between the two factors and we see a decrease in the estimated *ICC*.

			Scenario								
			$\boldsymbol{r_e = 0.7, r_u = 0.2}$			$\boldsymbol{r_e = 0.7, r_u = -0.2}$			$\boldsymbol{r_e = 0.2, r_u = 0.2}$		
		$\boldsymbol{ICC_k}$	**0.05**	**0.1**	**0.3**	**0.05**	**0.1**	**0.3**	**0.05**	**0.1**	**0.3**
Factor scores		$\hat{r}_e$	0.85	0.70	0.60	0.75	0.70	0.62	0.53	0.63	0.75
		$\hat{r}_u$	0.95	0.90	0.95	0.95	0.88	0.90	0.00	0.20	0.59
		$\widehat{ICC}_{f_1}$	0.16	0.24	0.51	0.20	0.20	0.53	0.04	0.06	0.09
		$\widehat{ICC}_{f_2}$	0.15	0.19	0.50	0.20	0.18	0.48	0.06	0.04	0.09

Table 2 $\hat{r}_e$, $\hat{r}_u$, and $\widehat{ICC}$ of factor scores under *multivariate MEBLUP averaged across samples.*

Table 3 shows the percentage relative reduction (in terms of RMSE) of the multivariate MEBLUP over the univariate EBLUP of the factor scores. The case $r_e = 0.2, r_u = 0.2$ produces smaller *ICCs.* This means that the MEBLUP has little gain over the univariate EBLUP. The case of $r_e = 0.7, r_u = -0.2$ and $r_e = 0.7, r_u = 0.2$ produce high factor correlations and higher ICCs; thus, increased efficiency of MEBLUP over the EBLUP. Note that the values of the RMSE of the factor scores means SAE predictions are shown in Table 6.

		Scenario		
$\boldsymbol{ICC_k}$		$\boldsymbol{r_e = 0.7, r_u = 0.2}$	$\boldsymbol{r_e = 0.7, r_u = -0.2}$	$\boldsymbol{r_e = 0.2, r_u = 0.2}$
0.05	*Factor scores 1*	-2.44	-2.50	0.00
	Factor scores 2	-2.50	-2.56	0.00
0.1	*Factor scores 1*	-2.56	-3.13	0.00
	Factor scores 2	-3.33	-2.86	0.00

0.3	*Factor scores 1*	-4.48	-5.56	0.00
	Factor scores 2	-5.56	-6.67	0.00

Table 3 Percentage relative reduction (%) in terms of RMSE of MEBLUP over EBLUP (Δ_k), two-factor CFA model

4.3.3. Comparison (c) of the use of latent factors (b) to simple and weighted averages of standardised EBLUP and MEBLUP estimates

One-Factor CFA Model

Table 5 provides the values of the RMSE of the estimates under consideration in comparison (c): simple and weighted averages of standardised original variables for EBLUPs and MEBLUPs and the one-factor CFA factor score means from the univariate SAE EBLUP. *Table 4* shows the percentage relative reduction in RMSE for the simple and weighted averages of standardised MEBLUPs over EBLUPs shown in *Table 4*.

		Scenario					
$\boldsymbol{ICC_k}$		$r_e = 0.7, r_u = 0.2$		$r_e = 0.7, r_u = -0.2$		$r_e = 0.2, r_u = 0.2$	
		EBLUP	MEBLUP	EBLUP	MEBLUP	EBLUP	MEBLUP
0.05	*Factor scores*	**0.081**	-	**0.080**	-	**0.079**	-
	simple	0.267	0.244	0.231	0.181	0.230	0.228
	weighted	0.230	0.220	0.207	0.164	0.185	0.184
0.1	*Factor scores*	**0.070**	-	**0.061**	-	**0.063**	-
	simple	0.246	0.225	0.250	0.180	0.207	0.205
	weighted	0.180	0.190	0.224	0.162	0.190	0.189
0.3	*Factor scores*	**0.065**	-	**0.039**	-	**0.078**	-
	simple	0.200	0.177	0.181	0.160	0.198	0.197
	weighted	0.175	0.157	0.163	0.144	0.185	0.185

Table 4 RMSE of factor scores means from one-factor CFA model, and simple and weighted averages of standardised original variables EBLUP/MEBLUP(Bold values highlight smaller RMSE for factor score means under EBLUP).

		Scenario		
$\boldsymbol{ICC_k}$		$r_e = 0.7, r_u = 0.2$	$r_e = 0.7, r_u = -0.2$	$r_e = 0.2, r_u = 0.2$
0.05	*simple*	-8.61	-21.65	-0.87
	weighted	-4.35	-20.77	-0.54
0.1	*simple*	-8.54	-28.00	0.00
	weighted	-5.56	-27.68	0.00
0.3	*simple*	-11.50	-11.60	-0.51
	weighted	-10.29	-11.66	0.00

Table 5 Percentage relative reduction (%) in terms of RMSE of simple and weighted averages of standardised MEBLUP over EBLUP (Δ_k).

Looking at *Table 5*, we can see that the EBLUP of the factor scores under the one-factor CFA model are all smaller than the simple and weighted averages of single variables under both the EBLUP and MEBLUP approaches. This confirms findings in Moretti et al. (2018b), which showed that factor score means estimated through EBLUP are more efficient compared to the dashboard approach of taking averages of indicators while both approaches have the same economic interpretation. In addition, the MEBLUP approach for the single variables provides estimates of simple and weighted averages with lower variability than the case where the single variables are estimated under the univariate EBLUP from *Table 5*. We do not see MSE reductions when the correlations in the variance-covariance matrices are small, which is the case when $r_e = 0.2, r_u = 0.2$.

Two-Factor CFA Model

Table 7 provides the values of the RMSE of each of the estimates under consideration in comparison (c): simple and weighted averages of standardised original variables for EBLUPs and MEBLUPs associated to each of the factors, and the two-factor CFA factor score means from the univariate and multivariate SAE. *Table 6* shows the percentage relative reduction in RMSE for simple and weighted averages of standardised MEBLUPs over EBLUPs for those variables associated to each of the factors in the two-factors CFA model as shown in *Table 6*. Note that the results of the percentage relative reduction in RMSE for the factor score means estimated by EBLUP and MEBLUP are shown in *Table 3* and discussed in Section 4.3.2.

			Scenario					
	$\boldsymbol{ICC_k}$		$r_e = 0.7, r_u = 0.2$		$r_e = 0.7, r_u = -0.2$		$r_e = 0.2, r_u = 0.2$	
			EBLUP	MEBLUP	EBLUP	MEBLUP	EBLUP	MEBLUP
Factor 1	0.05	*Factor scores*	**0.082**	**0.080**	**0.080**	**0.078**	**0.032**	**0.032**
		simple	0.380	0.360	0.360	0.340	0.350	0.340
		weighted	0.378	0.358	0.353	0.330	0.340	0.330
	0.1	*Factor scores*	**0.078**	**0.076**	**0.064**	**0.062**	**0.034**	**0.034**
		simple	0.450	0.410	0.450	0.330	0.400	0.402
		weighted	0.430	0.390	0.440	0.340	0.395	0.394
	0.3	*Factor scores*	**0.067**	**0.064**	**0.036**	**0.034**	**0.048**	**0.048**
		simple	0.600	0.530	0.600	0.500	0.356	0.355

		weighted	0.589	0.519	0.530	0.435	0.346	0.345
Factor 2	0.05	*Factor scores*	**0.040**	**0.039**	**0.039**	**0.038**	**0.012**	**0.012**
		simple	0.487	0.468	0.443	0.350	0.462	0.460
		weighted	0.485	0.462	0.440	0.344	0.450	0.449
	0.1	*Factor scores*	**0.030**	**0.029**	**0.035**	**0.034**	**0.022**	**0.022**
		simple	0.400	0.364	0.470	0.350	0.400	0.400
		weighted	0.388	0.345	0.465	0.341	0.375	0.375
	0.3	*Factor scores*	**0.036**	**0.034**	**0.030**	**0.028**	**0.028**	**0.028**
		simple	0.360	0.310	0.312	0.250	0.258	0.258
		weighted	0.350	0.305	0.305	0.253	0.245	0.245

Table 6 RMSE of factor score means from two factor CFA model and simple and weighted averages of standardized original variables EBLUP/ MEBLUP (Bold values highlight smaller RMSE for factor score means under EBLUP/MEBLUP).

		Scenario					
ICC_k		$r_e = 0.7, r_u = 0.2$		$r_e = 0.7, r_u = -0.2$		$r_e = 0.2, r_u = 0.2$	
		Simple	*Weighted*	*Simple*	*Weighted*	*Simple*	*Weighted*
0.05	*Factor 1*	-5.26	-5.29	-5.56	-6.52	-2.86	-2.94
	Factor 2	-3.90	-4.74	-20.99	-21.82	-0.43	-0.22
0.1	*Factor 1*	-8.89	-9.30	-26.67	-22.73	-0.50	-0.25
	Factor 2	-9.00	-11.08	-25.53	-26.67	0.00	0.00
0.3	*Factor 1*	-11.67	-11.88	-16.67	-17.92	-0.28	-0.29
	Factor 2	-13.89	-12.86	-19.87	-17.05	0.00	0.00

Table 7 Percentage relative reduction (%) in terms of RMSE for simple and weighted averages of variables associated to each of the factors of MEBLUP over EBLUP, (Δ_k) two-factors CFA model.

Table 6 shows that factor scores produce composite estimates with lower variability than simple and weighted averages for the two-factors case similar to the findings for the one-factor case. In *Table 7*, the MEBLUP provides estimates with lower variability than EBLUP for simple and weighted averages of those variables associated to each of the two factors in the two-factor CFA model. The percentage relative reduction is larger in the case of opposite signs in r_e and r_u. We also see no gains in efficiency when correlations are small.

4.4. Final remarks on simulation study

In this simulation study we investigated the use of CFA models in data dimensionality reduction and the application of multivariate SAE for small area indicators. It can be seen that, in line with the

general multivariate SAE literature, the use of multivariate mixed effects models provides estimates with lower variability than the univariate BHF model when variables are highly correlated with high intra-cluster correlations. In particular, the percentage of MSE reduction becomes larger when r_e and r_u have opposite signs. The use of factor score means provide more efficient estimates than the use of the simple and weighted averages of standardised EBLUPs and MEBLUPs of original variables for multidimensional phenomena. Interestingly, we can see that if the correlations in the original data are low, we see little or no gain in using an MEBLUP approach compared to the univariate EBLUP. The CFA model produces factor scores to represent latent variables which changes the correlation structures compared to the original variables. In particular, if the intra-cluster correlation reduces as a result of the CFA model, we see little gain in using the MEBLUP compared to the EBLUP. On the other hand, when correlations in the original data are high, and the correlation structure between factor scores remains high with an increased intra-cluster correlation, this leads to larger gains in the MEBLUP approach. However, in both cases we see that the MEBLUP approach has less reduction of RSMEs over the univariate EBLUP on factor score means estimation compared to a much larger reduction of RSMEs when comparing simple and weighted averages of small area estimates on the original variables. Thus it appears that when accounting for the correlation structure in the original data a priori through the use of CFA models, we can use a simpler univariate EBLUP approach on each of the factor scores means since there are little gains in using the MEBLUP approach.

5. Application

In this section we present an application using real data on housing quality in Italy, focusing on one of the key dimensions in the multidimensional Italian "Economic Wellbeing" of the BES framework. Housing quality is also an important determinant of wellbeing in other Organisation for Economic Co-operation and Development (OECD) countries (Andrews et al. 2011). Data from EU-SILC 2009 and the Italian Census 2001 (for the auxiliary variables) are used. Although the 2009 EU-SILC data were collected in 2008 (seven years after the census), the years 2001–2007 were a

period of relatively slow growth and low inflation in Italy (Giusti et al., 2012b). Future work will take into account more recent data for comparisons.

5.1. Data and variables

We focus on the following sub-dimensions of housing quality (Eurostat, 2016): housing deprivation and problems related to the residential area. Due to data availability, a limited number of variables are selected: severe material deprivation, smog, noise, crime, housing ownership, presence of humidity, darkness inside the house, absence of rubbish in the street, and absence of damages in public buildings. Income is another factor related to wellbeing, although monetary measurement is not always exhaustive for measuring poverty and wellbeing phenomena (Stiglitz et al., 2008). However, income has an interesting effect on housing quality. As Fusco (2015) notes, income and housing deprivation are negatively associated and, in the long run, this relationship becomes stronger. Therefore, it is reasonable to consider income in the analysis of multidimensional housing quality. In our work we use equivalised disposable income denoted by I^{DE}, which is calculated as follows (Atkinson et al., 2002):

$$I_i^{DE} = \frac{I_i^D}{n_i^E}, i = 1, \dots, N, \tag{16}$$

where $i = 1, \dots, N$ denotes households, I_i^D is the disposable household income, and n_i^E is the equivalised household size calculated in the following way:

$$n_i^E = 1 + 0.5 \cdot (HM_{14+} - 1) + 0.3 \cdot HM_{13-}, \tag{17}$$

where HM_{14+} is the number of household members aged 14 and over at the end of the income reference period, and HM_{13-} is the number of household members aged 13 or younger at the end of the income reference period.

The explanatory variables used in the model (following model-fit diagnostics not shown here) relate to the head of the household and are common to both EU-SILC and Census data. They are gender, age, year of education, household size, size of the flat (in squared metres), and status of employment. Appendix B shows descriptive statistics of the observed variables and auxiliary

variables used in the application.

The EU-SILC is conducted yearly by ISTAT for Italy, and coordinated by EUROSTAT at the EU level. For the Italian geography, the survey is designed to produce accurate estimates only at the national and regional levels (NUTS-2) and provinces, whereas municipalities (NUTS-3 and LAU-2 levels), and lower geographical levels are unplanned domains (Giusti et al., 2012a). The regional samples are based on a stratified two-stage sample design as follows: the Primary Sampling Units (PSUs) are the municipalities within the provinces and households are the Secondary Sampling Units (SSUs). The PSUs are stratified according to their population size. The SSUs are then selected by systematic sampling in each PSU. We use the EU-SILC 2009 dataset for Tuscany. The 14th Population and Housing Census 2001 surveyed 1,388,252 households of persons living in Tuscany permanently or temporarily, including the homeless population and persons without a dwelling.

5.2. Factor analysis and composite estimates

First, we show results of the unrestricted factor analysis model, also known as Explanatory Factor Analysis (EFA), on the observed variables to investigate their contribution to the total variability (Kaplan, 2009). *Table 8* shows the factor structure of the first two factors and how the variables relate to the factors via the factor loadings. According to the factors' structure , the following two latent variables can be defined: residential area deprivation (factor 1) and housing material deprivation (factor 2) as shown in *Figure 2*. *Figure* shows the scree plot of the EFA eigenvalues where it can be seen that indeed the first two factors explain a good amount of the total variability. Therefore, we keep two factors and carry out the Confirmatory Factor Analysis (CFA) model estimation stage. The factor scores are estimated from the CFA model using Mplus 7.4. For technical issues on the estimators we refer to Muthén (2004).

Variable	Factor 1	Factor 2
Severe material deprivation	0.010	**0.733**
Smog	**0.757**	0.025
Noise	**0.617**	0.154
Crime	**0.659**	0.130

Housing ownership	0.096	**-0.589**
Presence of humidity	0.010	**0.596**
Darkness inside the house	-0.002	**0.551**
Absence of rubbish in the street	**-0.843**	0.084
Absence of damages in public buildings	**-0.810**	0.012
Log equivalised disposable income	0.139	**-0.398**

Table 8 Factor structure for two latent factors using EFA.

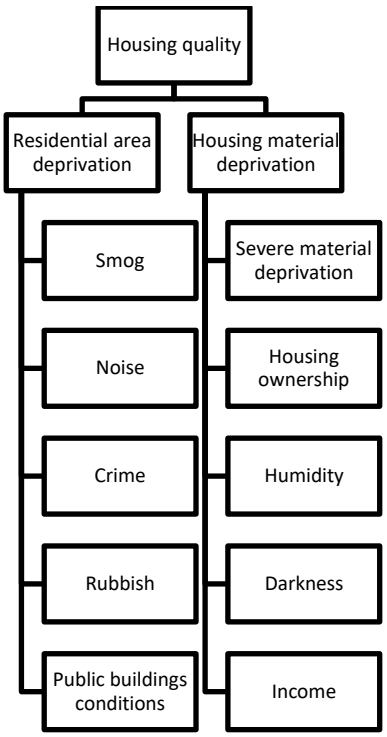

Figure 2 Housing quality sub-dimensions

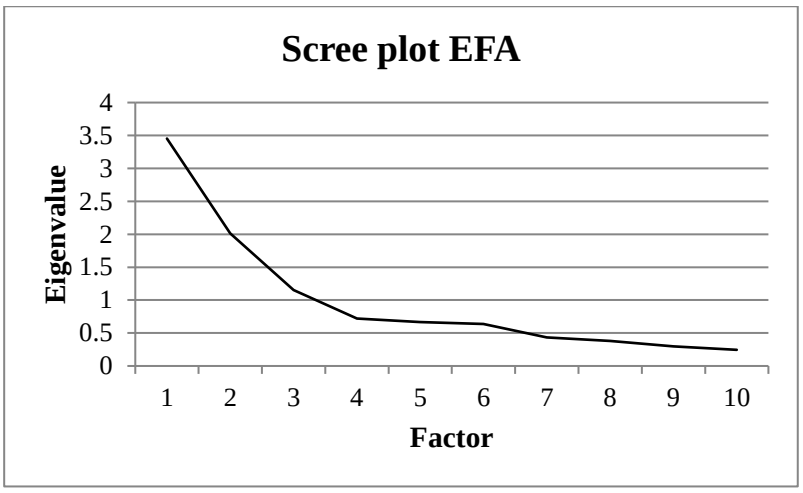

Figure 3 Scree plot EFA.

The goodness of fit statistics, root mean square error of approximation (RMSEA), the comparative fit index (CFI), and Tucker-Lewis index (TLI) show good results according to Hu and Bentler (1999): *RMSEA*=0.040*, CFI*=0.925*,* and *TLI*=0.901. The estimated correlation coefficient between factor 1 and factor 2 is 0.4. *Figure 2* shows the distributions of the factor scores for each of the latent variables arising from the CFA model following the use of the Box-Cox transformation with a parameter δ (Box and Cox, 1964) in order to approximate the normal distribution assumption needed for the SAE models. For Factor 1 we used δ=3.2 and for Factor 2 we used $\delta = 3.0$.

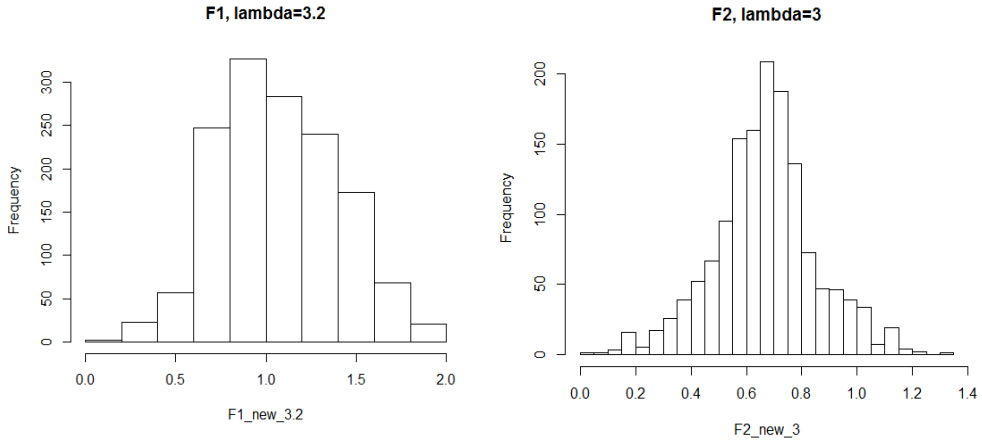

Figure 2 Factor scores histograms from CFA two-factor model after transformations.

5.3. Small area estimates and model diagnostics

Tuscany municipalities are defined as the EU-SILC small areas, with sample sizes ranging from 0 to 135 households. We assume a hierarchical structure in the data with households (level 1) nested within municipalities (level 2). The total number of households in the sample is 1,448 and 59 out of 287 municipalities were sampled. We build two different types of SAE models: first, we apply the univariate BHF approach and consider the factor scores as two separate dependent variables to obtain estimates of the univariate EBLUPs of the single factor means. Also, the multivariate approach is applied and the vector of the factor score means is predicted by MEBLUP. The MSEs of the EBLUPs of factor score means are estimated as in Moretti et al. (2018b). The MSEs of the MEBLUPs are estimated as in Moretti et al. (2018a), taking into account the variability arising from the CFA model as proposed in Moretti et al. (2018b).

In the case of areas where $n_d = 0$ it holds that:

$$\begin{matrix} \hat{\bar{f}}_{dm}^{EBLUP} = \hat{\bar{f}}_{dm}^{Synthetic} = \bar{\boldsymbol{X}}_d'\hat{\boldsymbol{\beta}}, m = 1,2 \\ \hat{\bar{\boldsymbol{f}}}_d^{MEBLUP} = \hat{\bar{\boldsymbol{f}}}_d^{Synthetic} = \bar{\boldsymbol{X}}_d'\hat{\boldsymbol{\beta}} \end{matrix} . \quad (18)$$

where $\hat{\bar{f}}_{dm}^{EBLUP}$ and $\hat{\bar{\boldsymbol{f}}}_{dm}^{MEBLUP}$ denote the EBLUP of the mean of the factor scores for the m^{th} factor and the MEBLUP of the mean vector of factor scores, respectively.

The final EBLUP and MEBLUP factor score means are then transformed for enabling interpretation and mapping using the 'Min-Max' criterion (OECD, 2008), which transforms the estimates to the interval [0,1]. For example, for the EBLUP of the *m=1,2* factors, the factor scores mean is transformed to a value given by:

$$\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP*} = \frac{\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP} - \min(\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP})}{\max\left(\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP}\right) - \min(\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP})}, \hat{\bar{f}}_{dm}^{EBLUP*} \in [0,1]. \quad (19)$$

where $\hat{\bar{\boldsymbol{f}}}_{dm}^{EBLUP}$ denotes the EBLUP of factor score means for the m^{th} factor for small area *d*, the minimum and maximum are across all EBLUPs in areas *d=1,…,D.*

We proceed with the MEBLUP of factor score means and interpret our findings. *Table 9* shows the percentiles for the transformed latent housing quality indicators based on MEBLUP of factor score means. *Figure 3* shows the maps of residential area deprivation and housing material deprivation, respectively.

	MEBLUP Percentile				
	0%	*25%*	*50%*	*75%*	*100%*
Residential area deprivation	0.000	0.261	0.266	0.270	1.000
Housing material deprivation	0.000	0.418	0.457	0.502	1.000

Table 9 Percentiles for transformed latent housing quality indicators based on MEBLUP of factor score means.

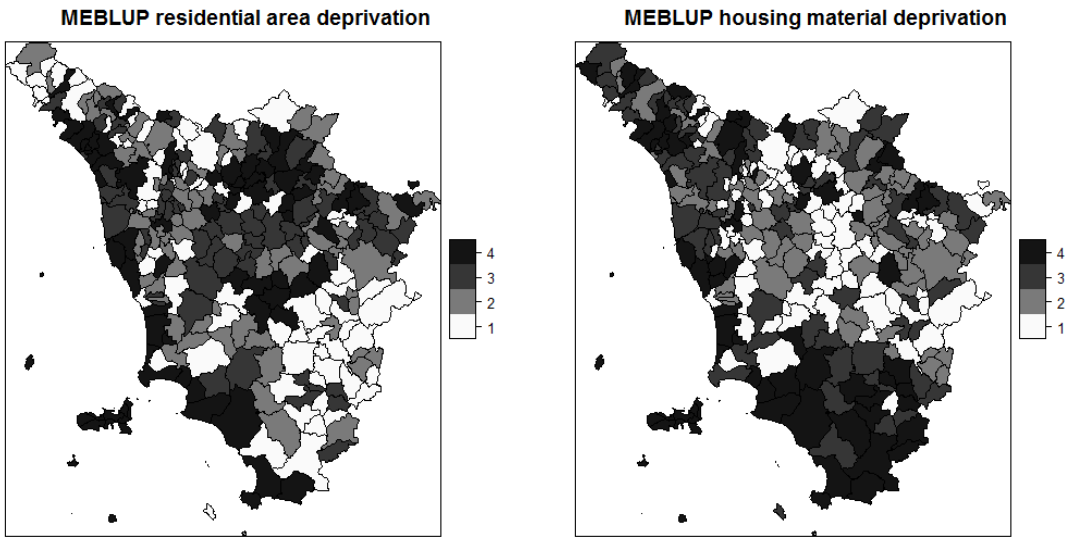

Figure 3 Housing quality indicators based on transformed MEBLUP factor score means {1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile}.

Although the residential area deprivation dimension is positively correlated with the housing material deprivation dimension, there are important differences at the area level between the two sub-dimensions. These differences can be seen in the maps. Looking at residential area deprivation estimates (*Figure 3*; left panel) it can be seen that the municipalities located in Massa e Carrara and Siena provinces have the lowest values of the residential area deprivation indicators. Low levels of

residential area deprivation are estimated for some municipalities of the south Grosseto province (Manciano and Magliano in Toscana). The highest values in residential area deprivation areas are estimated for municipalities located in the north of the Florence province and north Livorno province. The second map in *Figure 3* (right panel) depicts the housing material deprivation indicator. Interestingly, although the correlation between the two indicators is 0.4, there are noteworthy differences in some areas: Massa e Carrara, north Siena, Florence, Grosseto and south Siena provinces. For the municipalities located in these provinces the estimates of the housing material deprivation indicator belong to the 4th quantile, denoting high levels of housing material deprivation and belong to the 1st and 2nd quantiles denoting low levels of residential area deprivation.

A multilevel analysis (using SAE in this case) shows that one housing dimension can be low and the other housing dimension can be high for some areas. This gives important guidelines for informing policies.

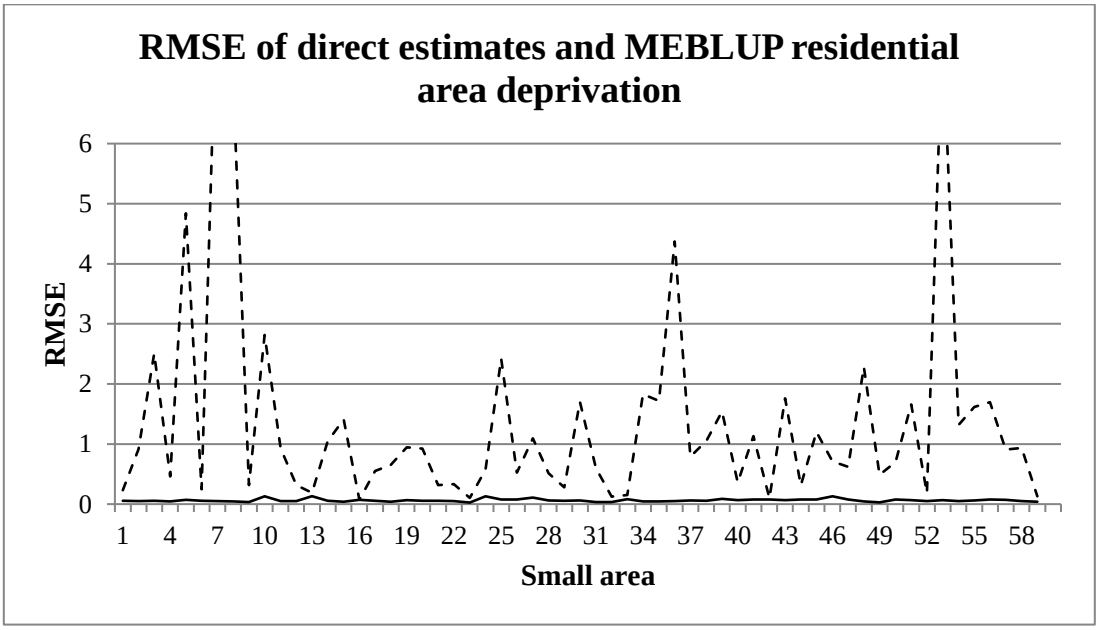

Figure 4 Root Mean Squared Error (RMSE) of MEBLUP (__) and direct estimates (---) of residential area deprivation small areas with $n_d > 0$.

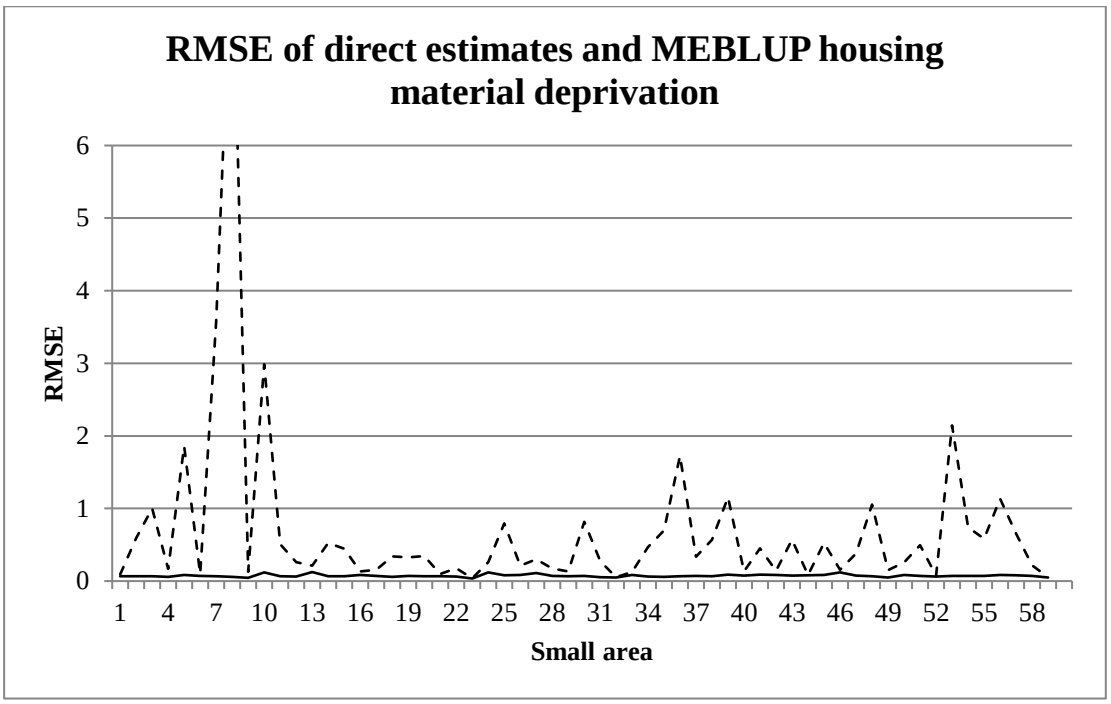

Figure 5 Root Mean Squared Error (RMSE) of MEBLUP (__) and direct estimates (---) of housing material deprivation small areas with $n_d > 0$.

Figure 4 and *Figure 5* show the Root Mean Squared Error (RMSE) of MEBLUP and direct estimates calculated via the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) for those small areas with $n_d > 0$ for residential area deprivation and housing material deprivation, respectively. *Figure 6* and *Figure 7* show the RMSEs of residential area deprivation and housing material deprivation comparing the EBLUP and MEBLUP estimates for those small areas with $n_d > 0$, respectively.

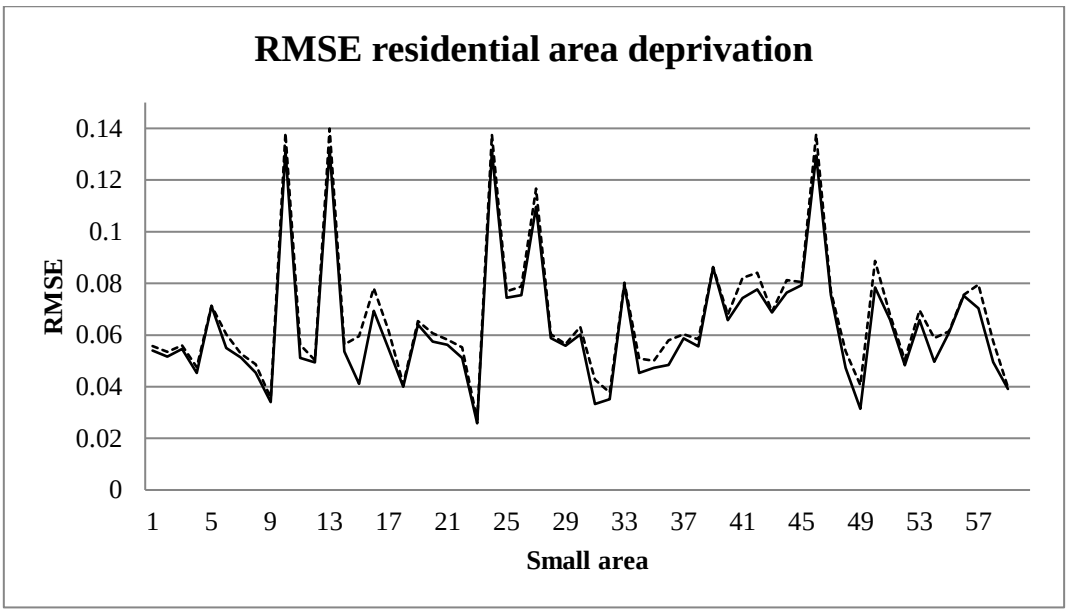

Figure 6 Root Mean Squared Error (RMSE) of MEBLUP (__) and EBLUP (---) of residential area deprivation small areas with $n_d > 0$.

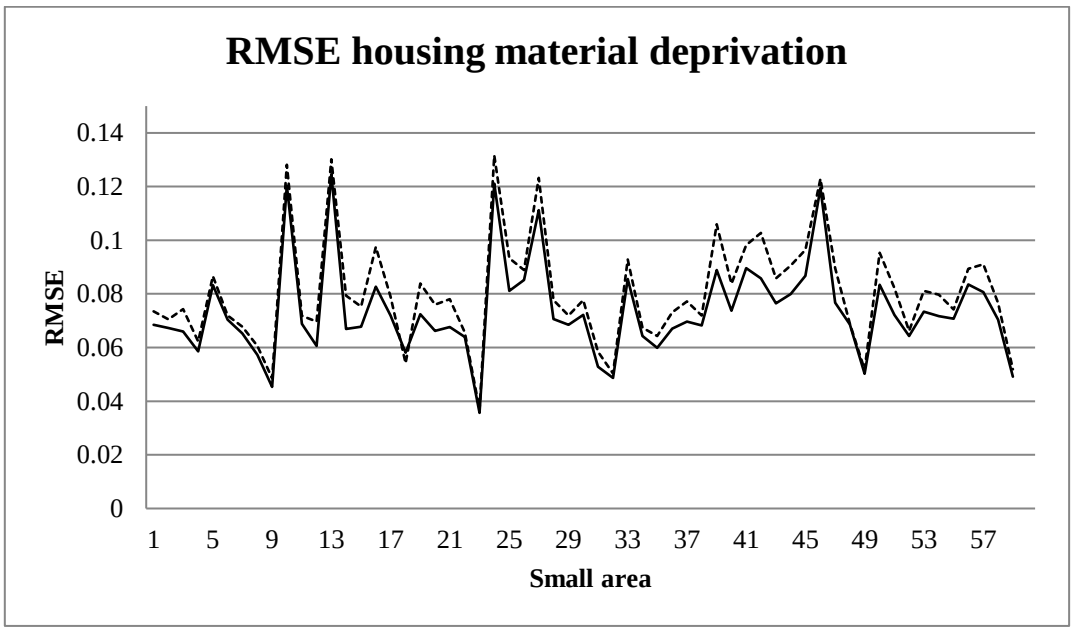

Figure 7 Root Mean Squared Error (RMSE) of MEBLUP (__) and EBLUP (---) of housing material deprivation small areas with $n_d > 0$.

It can be seen from the figures that the MEBLUP approach provides smaller RMSE over the univariate EBLUP approach. The percentage reduction in terms of *RMSE* across all areas is 6.41% and 7.90% for residential area deprivation and housing material deprivation, respectively.

The model estimates of the variance components and correlations of the latent factors are:

$$\hat{\sigma}^2_{e,f_1} = 0.086, \sigma^2_{u,f_1} = 0.023,$$

$$\hat{\sigma}^2_{e,f_2} = 0.170, \sigma^2_{u,f_2} = 0.017,$$

$$\hat{\boldsymbol{\Sigma}}_e \begin{bmatrix} 0.086 & 0.012 \\ 0.012 & 0.169 \end{bmatrix} with\ \hat{r}_e = 0.10,$$

$$\hat{\boldsymbol{\Sigma}}_u \begin{bmatrix} 0.023 & 0.015 \\ 0.015 & 0.016 \end{bmatrix} with\ \hat{r}_u = 0.78.$$

The estimated *ICCs* are 0.21 and 0.09 for factor 1 and factor 2, respectively.

Figure 8 and *Figure 9* show the Q-Q plots of the residuals (level-1 and level-2) from the BHF and multivariate models, respectively, for both of the factors. It can be seen that the residuals are approximately normally distributed and, in the case of the multivariate mixed effects model, they behave slightly better.

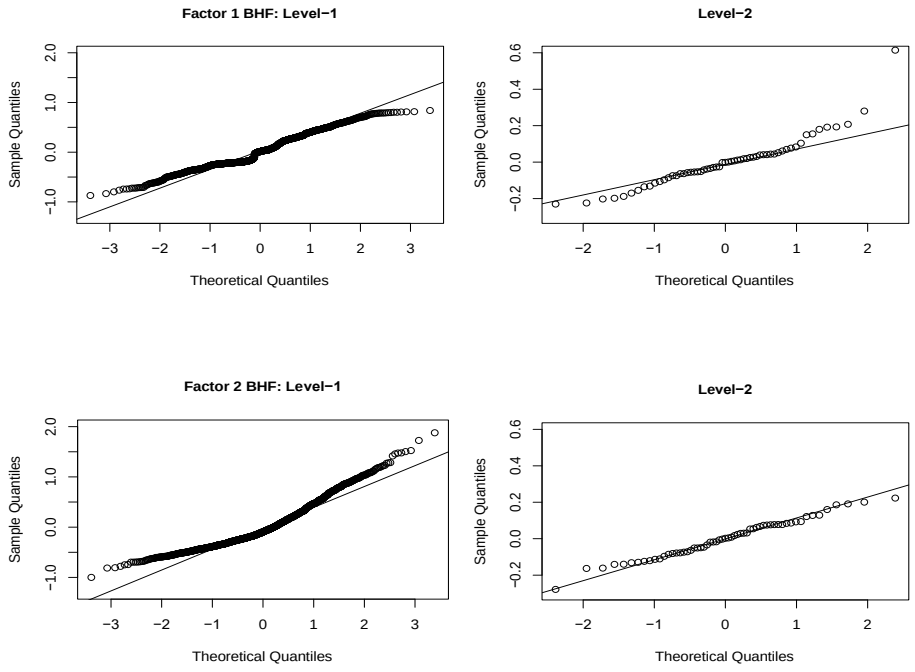

Figure 8 Q-Q plots of the residuals estimated from the univariate BHF model.

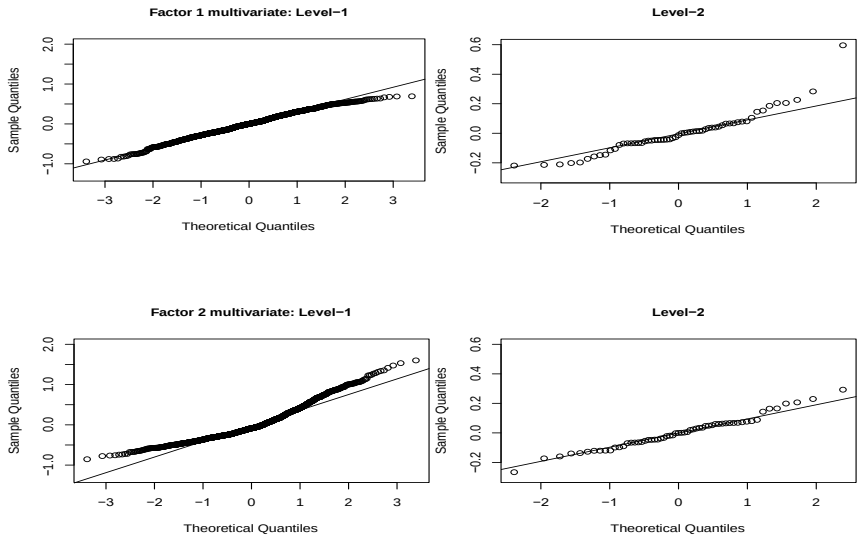

Figure 9 Q-Q plots of the residuals estimated from the multivariate model.

6. Discussion

In this paper we evaluated the use of a multivariate empirical best linear unbiased predictor (MEBLUP) for data dimensionality reduction. In particular, we compared the use of factor score means with the use of simple and weighted averages of standardised EBLUPs and MEBLUPs of original variables in a large-scale simulation study.

The reduction in terms of MSE of the multivariate analysis over the univariate analysis depends on the correlation coefficients (r_e and r_u) associated to the variance-covariance matrices and intra-class correlation of the original variables and in particular how these change when accounting for the correlations a priori through Factor Analysis models.

Our work contributes to the data dimensionality issue in small area estimation. To summarise, we can state that when factor score means on several latent variables are used in data dimensionality reduction, these may be calculated using univariate EBLUPs, since the correlation structure is accounted for a priori via the factor analysis model. This is shown in the simulation study under comparison (c), where percentages of reduction in terms of RMSE for the factor scores case are small compared to the weighted and simple averages of the original variables.

We note that factor scores are still crucial in data dimensionality reduction where different types of

variables may arise (binary, continuous, categorical etc.). In fact, in the real data application, we have variables measured on different scales, hence, multivariate EBLUP would require joint multivariate mixed effects models, which have not been studied in SAE so far and is a topic for future work. Factor scores estimated by a FA analysis model overcome this issue and allow the study of multidimensional well-being phenomena. Another area of future work is the study of MSE estimation in multivariate SAE models.

Acknowledgements

This analysis was carried out on confidential data released by ISTAT. Data were analysed by respecting all of the Italian confidential restriction regulations (D.Lgs. 196/03 – Codice Privacy). Therefore, we are not able to release the data. The authors thank Dr. Luca Faustini and Dr. Linda Porciani from the ISTAT regional office of Florence for their kind help and suggestions during the data request process. This work was financially supported by the following grant: ESRC DTC Award and Advanced Quantitative Methods (also known as AQM).

Appendix A: Goodness of Fit for CFA Models on Generated Population for Simulation Study

		One-factor model			Two-factor model		
Correlation structure	ICC_k	**SRMR**	**CFI**	**TLI**	**SRMR**	**CFI**	**TLI**
$\boldsymbol{r_e = 0.2, r_u = 0.2}$	0.05	0.016	0.985	0.956	0.026	0.985	0.956
	0.1	0.016	0.986	0.957	0.016	0.986	0.957
	0.3	0.016	0.991	0.972	0.016	0.991	0.972
$\boldsymbol{r_e = 0.7, r_u = 0.2}$	0.05	0.040	0.969	0.908	0.035	0.989	0.978
	0.1	0.038	0.971	0.912	0.032	0.975	0.925
	0.3	0.028	0.985	0.955	0.020	0.985	0.955
$\boldsymbol{r_e = 0.7, r_u = -0.2}$	0.05	0.040	0.970	0.909	0.038	0.978	0.978
	0.1	0.032	0.975	0.924	0.029	0.968	0.927
	0.3	0.020	0.985	0.955	0.024	0.987	0.978

Table A1 Confirmatory factor analysis goodness of fit statistics, one-factor and two-factor model, on the generated population

Appendix B: Description of variables on EU-SILC 2009 Tuscany dataset for Application in Section 5

Variable	Mean	S.D.
Severe material deprivation	4%	0.0384
Smog	17%	0.373
Noise	23%	0.424
Crime	13%	0.341
Housing ownership	74%	0.439
Presence of humidity	15%	0.358
Darkness inside the house	8%	0.277
Equivalised disposable income	20,090	13,990.88
Rooms per household component	1.989	1.239

Table B1 Descriptive statistics of the observed variables (EU-SILC, Tuscany 2009).

Access to public services		
	Absolute frequency	*Relative frequency %*
Very difficult	133	9.19
Some difficulties	249	17.20
Easy	631	**43.58**
Very easy	290	20.03
Not needed	145	10.01
Total	1448	100.00

Table B2 Frequency distribution of access to public services (EU-SILC, Tuscany 2009)

Perception of damages to public buildings		
	Absolute frequency	*Relative frequency %*
Always	65	4.49
Often	83	5.73
Sometime	294	20.30
Never	1006	**69.48**
Total	1448	100.00

Table B3 Frequency distribution of damages to public buildings (EU-SILC, Tuscany 2009)

Perception of rubbish in the street		
	Absolute frequency	*Relative frequency %*
Always	75	5.18
Often	82	5.66
Sometime	308	21.27
Never	983	**67.89**
Total	1448	100.00

Table B4 Frequency distribution of perception of rubbish in the street (EU-SILC, Tuscany 2009)

Variable	Mean	S.D.
Household size	2.43	1.18
Gender (female)	70%	0.46
Status of employment (employed)	50%	0.50
Age	57.39	16.86
Years of education	9.76	4.56
Flat (or house) size in squared metres	97.54	38.43

Table B5 Descriptive statistics of the auxiliary variables (EU-SILC, Tuscany 2009).

Appendix C: Specification of the R functions used

Here we describe the main R packages that can be to replicate the analysis.

C.1 Estimation of small area means and MSE under univariate EBLUP approach. Although we programmed our functions manually, the *sae* package (Molina and Marhuenda, 2015) may be used:

- Required packages: nlme, MASS
- Functions: eblupBHF( ) and pbmseBHF( ),

nlme and MASS are still required.

C.2 Running Mplus models in the R environment via *MplusAutomation* (Muthén and Muthén, 2012; Hallquist and Wiley, 2014)

- Functions: mplusObject( ), mplusModeler( ).

Mplus is required.

C.3 Mapping using *spdep, maptools, sp, Hmisc*

- Functions: readShapePoly( ), spplot( )

C.4 Multivariate mixed effect model ML fitting via *mlmmm* (Yucel, 2010)

- Function: mlmmm.em()

All the other analysis can be programmed easily.

References

Andrews, D., Caldera Sánchez, A., and Johansson, A. (2011) Housing markets and structural policies in OECD countries. OECD Economics Department Working Papers, no 836. OECD Publishing.

Atkinson, T., Cantillon, B, Marlier, E., Nolan, B. (2002) *Social Indicators: The EU and Social Inclusion*. Oxford University Press, Oxford.

Australia. Bureau of Agricultural Economics (1978), Australian agricultural and grazing industries survey. Grazing industry - sheep and beef cattle industries, *Australian Government Publishing Service*, Canberra.

Baillo, A. and Molina, I. (2009) Mean Squared Errors of Small-Area Estimators Under a Unit-Level Multivariate Model, Statistics, **43**, 553-569.

Battese, G. E., , R., Harter, R. M. and Fuller, W. A. (1988) An Error-Components model for Prediction of County crop areas using Survey and Satellite data. *Journal of the American Statistical Association*, **83** (401) ,28-36.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *Journal of Royal Statistical Society Series B* **26**, 211-246.

Datta, G. S., Day, B. and Basawa, I. (1999) Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference,* **75**, 269-279.

Eurostat (2016) Housing statistics. Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php/Housing_statistics.

Fuller, W. A. and Harter, R. M. (1987) The Multivariate Components of Variance Model for Small Area Estimation, in R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh (Eds.), *Small Area Statistics*, New York: Wiley, 103-123.

Fusco, A. (2015) The relationship between income and housing deprivation: A longitudinal analysis.

Economic Modelling, **49**, 137-143.

Giusti, C., Marchetti, S. and Pratesi, M. (2012a) Estimation of Income Quantiles at the Small Area Level in Tuscany in A. Di Ciaccio et al. (eds.), Advanced Statistical Methods for the Analysis of Large Data-Sets, Studies in Theoretical and Applied Statistics, Springer-Verlag Berlin Heidelberg 2012.

Giusti, C., Marchetti, S., Pratesi, M. and Salvati, N. (2012b) Robust Small Area Estimation and Oversampling in the Estimation of Poverty Indicators. *Survey Research Methods*, **6**(3), 155-163.

Hallquist, M and Wiley, J. (2014) MplusAutomation: Automating Mplus Model Estimation and Interpretation. R package version 0.6-3. Retrieved from http://CRAN.R-project.org/package=MplusAutomation.

Hardle, W. K. and Simar, L. (2012) *Applied Multivariate Statistical Analysis*. Springer - Verlag Berlin Heidelberg.

Horvitz D.G. and Thompson, D. J. (1952) A Generalization of Sampling without Replacement from Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

Hu, L., and Bentler, P. M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6** (1), 1-55.

ISTAT and CNEL (2015) BES 2015. Il benessere equo e sostenibile in Italia. *Report* published by ISTAT and CNEL in 2015. Retrieved from http://www.istat.it/it/files/2015/12/Rapporto_BES_2015.pdf. (In Italian).

Johnson, R., Wichern, D. W. (1998) Applied Multivariate Analysis, 4th edition, Prentice.

Kaplan, D. (2009) Structural Equation Modeling. Foundations and Extensions. Second Edition. Madison, USA: Sage.

Lemmi, A., and Panek, T. (2016) Regional and Local Poverty Measures. In: Pratesi (Editor), *Analysis of Poverty Data by Small Area estimation*. Wiley.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) Multivariate Analysis, Academic Press, Duluth,

London.

Molina, I. and Marhuenda, Y. (2015) sae: An R Package for Small Area Estimation. *The R Journal*, **7**(1), pp. 81-98. Retrieved from http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf>.

Molina, I. (2009) Uncertainty under a multivariate nested-error regression model with logarithmic transformation. *Journal of Multivariate Analysis*, **100**, 963-980.

Moretti, A., Shlomo, N., and Sakshaug, J. (2018a) Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP. *Communications in Statistics – Simulation and Computation.* In press.

Moretti, A., Shlomo, N., and Sakshaug, J. (2018b) Small Area Estimation of Latent Economic Wellbeing. *Sociological Methods & Research*. In press.

Muthén, B.O. (2004) Mplus Technical Appendices. Los Angeles, CA: Muthén and Muthén.

Muthén, L.K. and Muthén, B.O. (2012) *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén and Muthén.

OECD, (2004) The OECD-JRC Handbook on Practices for Developing Composite Indicators, paper presented at the OECD Committee on Statistics, 7-8 June 2004, OECD, Paris.

OECD (2008) Handbook on constructing composite indicators: methodology and user guide. *OECD Statistics report*. Retrieved from http://www.oecd.org/std/42495745.pdf.

Rao, J. N. K. and Molina, I. (2015) *Small area estimation*. New York: Wiley.

Ravallion, M. (2011) On multidimensional indices of poverty. *Journal of Economic Inequality* **9**(2), 235-248.

Schafer, J. L., and Yucel, R. M. (2002) Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics.* **11**, 437-457.

Stiglitz, J.E., Sen, A., and Fitoussi J.P. (2008) Report by the Commission on the Measurement of Economic Performance and Social Progress, available online at: www.stiglitz-sen-

fitoussi.fr/documents/rapport_anglais.pdf.

United Nations (2017) United Nations sustainable development agenda. Available at http://www.un.org/sustainabledevelopment/development-agenda/.

Yalonetzky, G. (2012) Conditions for the most robust multidimensional poverty comparisons using counting measures and ordinal variables. ECINEQ working paper, 2012-2257.

Yucel, R. (2010) mlmmm: ML estimation under multivariate linear mixed models with missing values. R package version 0.3-1.2. Retrieved from http://CRAN.R-project.org/package=mlmmm.