MULTIVARIATE SMALL AREA ESTIMATION FOR MULTIDIMENSIONAL WELL-BEING INDICATORS

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2018

ANGELO MORETTI

School of Social Sciences, Department of Social Statistics
# Table of Contents

Table of Contents.................................................................................................................................2

List of Tables .............................................................................................................................................6

List of Figures ...........................................................................................................................................8

Abstract ....................................................................................................................................................11

Acknowledgments....................................................................................................................................14

1 Introduction ..........................................................................................................................................15

  1.1 Multidimensional Well-being .........................................................................................................15

  1.2 The problem of composite indicators .............................................................................................17

  1.3 The need for local estimates: small area estimation methods...........................................................20

  1.4 Multivariate Small Area Estimation and Data Dimensionality Reduction ........................................23

  1.5 Research Questions ..........................................................................................................................24

  1.6 Rationale for Alternative Format .....................................................................................................25

  1.7 The Aims and Structure of this Thesis..............................................................................................26

2 Literature Review and Methods Overview ............................................................................................31

  2.1 Composite Indicators .......................................................................................................................31

  2.2 Factor analysis model and the use of factor scores ..........................................................................36

    2.2.1 The linear factor analysis model ...............................................................................................36

    2.2.2 The choice of the number of factors .........................................................................................39

  2.3 Two-level factor analysis models .......................................................................................................40

  2.4 Model- versus design-based inference ...............................................................................................42

  2.5 The use of the multivariate nested error regression model .................................................................43

    2.5.1 Notation .....................................................................................................................................43

    2.5.2 The multivariate SAE problem for a vector of means ...............................................................44
2.6 Mean Squared Error Estimation........................................................................................................... 47
  2.6.1 Resampling techniques to approximate the MSE ................................................................. 49
2.7 The data used in this thesis .................................................................................................................. 50
2.8 Gaps in the literature: multivariate SAE, composite indicators and multidimensional social indicators .............................................................................................................. 51

3 Small Area Estimation of Latent Economic Well-being ........................................................................ 53
  3.1 Introduction ........................................................................................................................................ 55
  3.2 Using Factor Scores for Data Dimensionality Reduction ............................................................... 59
    3.2.1 Issues in Composite Indicators ............................................................................................ 60
    3.2.2 The Linear Single-factor Analysis Model .............................................................................. 61
  3.3 Small Area Estimation using Empirical Best Linear Unbiased Prediction (EBLUP) ....................... 64
    3.3.1 Notation .................................................................................................................................. 65
    3.3.2 Model based prediction using EBLUP .................................................................................... 65
    3.3.3 Mean Squared Error Estimation ............................................................................................ 67
  3.4 Simulation Study ............................................................................................................................... 68
    3.4.1 Generating the population ....................................................................................................... 69
    3.4.2 Simulation steps ....................................................................................................................... 74
    3.4.3 Results: factor scores versus weighted and simple averages of standardized EBLUPs ............ 76
    3.4.4 Bootstrap MSE Estimation ....................................................................................................... 78
    3.4.5 Final remarks on the simulation study .................................................................................... 80
  3.5 Economic Well-being in Tuscany: a Multidimensional Approach .................................................. 81
    3.5.1 Data and variables .................................................................................................................... 81
    3.5.2 The construction of the factor scores ....................................................................................... 85
    3.5.3 Small Area Estimates .............................................................................................................. 87
Appendix A: Goodness of Fit for CFA Models on Generated Population for Simulation Study
Appendix B: Description of variables on EU-SILC 2009 Tuscany dataset for Application in Section 4.5
Appendix C: Tuscany region map
Appendix D: Specification of the R functions used and issues in computation

5 Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP
5.1 Introduction
5.2 Multivariate Small Area Estimation of a Means Vector
  5.2.1 Multivariate nested-error linear regression model
  5.2.2 Estimation and prediction of unknown parameters
5.3 Parametric bootstrap
5.4 Simulation study
  5.4.1 Generating the population
  5.4.2 Results
  5.4.3 Final remarks on the simulation study
5.5 Application to Corn and Soy Bean Data
5.6 Conclusion

Appendix A Issues in computations

6 Discussion
  6.1 Summary of this work and relationship to the literature
  6.2 Future work and limitations
  6.3 Final remarks

References

WORD COUNT: 34287
List of Tables

Table 1  Eigenvalues from the EFA of the simulation population. .................................72
Table 2  Average eigenvalues across 500 samples from EFA model.................................76
Table 3  Average intra-class correlation $\hat{\rho} = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_e}$ estimates across 500 samples ......76
Table 4  Spearman's correlation estimates for the three approaches. .................................77
Table 5  RMSE estimates: comparison across 500 samples for the three approaches. ....78
Table 6  Descriptive statistics of factor scores......................................................................86
Table 7  Percentiles for the transformed latent economic well-being indicator based on the EBLUP of factor score means and simple and weighted averages.........................90
Table 8  Percentage relative reduction (%) in RMSE of MEBLUP over EBLUP ($\Delta_k$) for single observed response variables averaged over all areas. .................................121
Table 9  $\hat{r}_e$, $\hat{r}_u$, and $\hat{I}_{CC}$ of factor scores under multivariate MEBLUP averaged across samples...........................................................................................................123
Table 10  Percentage relative reduction (%) in terms of RMSE of MEBLUP over EBLUP ($\Delta_k$), two-factor CFA model....................................................................................124
Table 11  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)..................................125
Table 12  Percentage relative reduction (%) in terms of RMSE of simple and weighted averages of standardised MEBLUP over EBLUP ($\Delta_k$). .................................................................126
Table 13 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)........................................ 128

Table 14 Percentage relative reduction (%) in terms of RMSE for simple and weighted averages of variables associated to each of the factors of MEBLUP over EBLUP, \( \Delta_k \) two-factors CFA model. .................................................................................................................. 129

Table 15 Factor structure for two latent factors using EFA. ................................................................. 134

Table 16 Percentiles for transformed latent housing quality indicators based on MEBLUP of factor score means. ................................................................................................................ 139

Table 17 Descriptive statistics and relative bias of the Prasad-Rao and bootstrap estimators for univariate EBLUP MSE across small areas, \( EMSE(\hat{Y}_1^{EBLU}) = 18.35 \), \( EMSE(\hat{Y}_2^{EBLU}) = 13.42 \), for \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \)................................................................. 170

Table 18 Empirical mean squared error, bootstrap MSE, relative bias across the small areas: EBLUP and MEBLUP estimates – parametric bootstrap. \( \Delta_k \) shown in parenthesis). .................................................................................................................. 173
List of Figures

Figure 1 Scree plots from the EFA of the simulation population. ........................................73

Figure 2 RMSE for Direct estimates and EBLUP of factor score means for small areas with \( n_d > 0 \)........................................................................................................................................77

Figure 3 Ratios between bootstrap RMSE and empirical RMSE estimated via bootstrap taking into account the FA model variability (---) and bootstrap ignoring the FA model variability (__). ..................................................................................................................................80

Figure 4 Factor scores distribution graphs..................................................................................86

Figure 5 RRMSE direct estimates (__) and EBLUPs (---) for small areas with \( n_d > 0 \) ordered by growing sample size. .................................................................................................................88

Figure 6 Latent economic well-being indicator based on transformed EBLUP of factor scores means \{1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile\}. ..............90

Figure 7 Latent economic well-being indicator based on simple and weighted averages of single EBLUPs \{1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile\}. .....91

Figure 8 Q-Q plots for the level-1 and level-2 residuals of the BHF model fitting...........93

Figure 9 Standardized residuals versus leverage measure...........................................................93

Figure 10 Relationship between the factors and observed variables two-factor CFA...118

Figure 11 Housing quality sub-dimensions. ..................................................................................135

Figure 12 Scree plot EFA. ..............................................................................................................135

Figure 13 Factor scores histograms from CFA two-factor model after transformations. ................................................................................................................................136
Figure 14 Housing quality indicators based on transformed MEBLUP factor score means \{1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile\}. ...................... 139

Figure 15 Root Mean Squared Error (RMSE) of MEBLUP (___) and direct estimates (---) of residential area deprivation small areas with \( n_d > 0 \). ............................................. 141

Figure 16 Root Mean Squared Error (RMSE) of MEBLUP (___) and direct estimates (---) of housing material deprivation small areas with \( n_d > 0 \). ............................................. 141

Figure 17 Root Mean Squared Error (RMSE) of MEBLUP (___) and EBLUP (---) of residential area deprivation small areas with \( n_d > 0 \). ............................................. 142

Figure 18 Root Mean Squared Error (RMSE) of MEBLUP (___) and EBLUP (---) of housing material deprivation small areas with \( n_d > 0 \). ............................................. 143

Figure 19 Q-Q plots of the residuals estimated from the univariate BHF model. ....... 144

Figure 20 Q-Q plots of the residuals estimated from the multivariate model.......... 145

Figure 21 Relative bias of univariate EBLUPs’ MSEs of \( y_1 \) estimated via Prasad-Rao approximation and parametric bootstrap for \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) .................... 171

Figure 22 Relative bias of univariate EBLUPs’ MSEs of \( y_2 \) estimated via the Prasad-Rao approximation and parametric bootstrap for the \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) .......... 171

Figure 23 Bootstrap MSEs \( y_1 \): comparison between EBLUP and MEBLUP \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) ................................................................................................. 174

Figure 24 Bootstrap MSEs \( y_2 \): comparison between EBLUP and MEBLUP \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) ................................................................................................. 174

Figure 25 Coverage rates (%) comparisons: MSEs of MEBLUP and EBLUP estimated via bootstrap and via Prasad-Rao approximation (PR) for EBLUP case: \( y_1 \) for \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) .......................................................... 175
Figure 26 Coverage rates (%) comparisons: MSEs of MEBLUP and EBLUP estimated via bootstrap and via Prasad-Rao approximation (PR) for EBLUP case: $y_2$ for $\rho_e = -0.7$ and $\rho_u = 0.3$. ..........................................................176

Figure 27 Bootstrap RMSEs $y_1$ - corns: comparison between EBLUP (---) and MEBLUP (___).................................................................180

Figure 28 Bootstrap RMSEs $y_2$ – soy beans: comparison between EBLUP (---) and MEBLUP (___)........................................................................................................180
Abstract

Using multivariate statistical models in small area estimation (SAE) may improve the efficiency of the small area estimates over the univariate SAE. In this thesis, we study the multivariate SAE problem of multidimensional well-being indicators. We first investigate the univariate EBLUP for a single latent variable estimated through confirmatory factor analysis. We use factor scores as composite estimates and calculate the EBLUP of factor score means and compare the use of these with the traditional approach of weighted and simple averages of standardized univariate EBLUPs of a dashboard of single observed indicators. Our simulation studies show that the use of factor scores provides more accurate and efficient estimates than weighted and simple averages in SAE. We also propose a bootstrap algorithm that accounts for the factor analysis model variability in the mean squared error (MSE) estimation of an EBLUP of factor score means. Next, we examine the use of multivariate EBLUP to estimate factor score means (for two latent factors) and compare to the use of weighted and simple averages of standardized EBLUPs of a dashboard of single observed indicators that are estimated in a univariate approach and in a multivariate SAE. We show that in general the multivariate EBLUP is more efficient than the univariate EBLUP, however, when the data correlation is taken into account before SAE estimates are computed (the case of factor scores) multivariate EBLUP does not provide large improvements in efficiency over the univariate case. Finally, we propose an MSE bootstrap estimator of a multivariate EBLUP. The results are in line with the SAE literature in terms of MSE comparisons of the multivariate EBLUP over the univariate EBLUP.
Declaration

I, Angelo Moretti, declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

1. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

2. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
4. Further information on the conditions under which disclosure, publication and
commercialisation of this thesis, the Copyright and any Intellectual Property
and/or Reproductions described in it may take place is available in the University
relevant Thesis restriction declarations deposited in the University Library, The
University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s
policy on Presentation of Theses.
Acknowledgments

Firstly, I would like to express my sincere gratitude to my Ph.D. supervisors, Prof. Natalie Shlomo and Dr. Joseph Sakshaug for their continuous support during my Ph.D., for their patience, motivation, and knowledge.

Besides my supervisors, I would like to thank my Master of Science supervisor, Prof. Monica Pratesi who introduced me to the fascinating ‘Statistics world’ and to Small Area Estimation methods.

I would like to thank my internal reviewer, Dr. Johan Koskinen, and my first year internal reviewer Prof. Mark Tranmer, for their precious comments and suggestions during my Ph.D.

My most heartfelt thanks go to my boyfriend Robin for his immense support and encouragement.

I am also thankful to my best friends Annarita, Caterina and Chiara. Our friendship has been extremely important in my life and especially during the last three years. Caterina provided great help in the software developed and used in this thesis.

I would like to thank my parents, for giving me the opportunity to study, and my grandmother Clema who had so much influence in my life.

Thanks to my friends and fellow Ph.D. students, they have been a source of support and friendship: Adrian, Bo, Dan, David, Georgia, Jen, José, Rihab, Sixten and Yizhang.

I gratefully acknowledge the Economic and Social Research Council for providing funding for this research (grant number ES/J500094/1).
1 Introduction

In this chapter we discuss the topic of multidimensional well-being indicators and the need of small area estimation methods in this field. We also outline the research questions and thesis structure.

1.1 Multidimensional Well-being

In recent years measuring well-being has become a key issue for European Union (EU) governments. Some important initiatives in this context were due to the Organisation for Economic Co-operation and Development (OECD) since 2004. In 2008, the French Government launched the well-known Stiglitz-Sen-Fitoussi Commission with the aim of studying the measurement of economic performance and social progress. Since the 1980s, the research community has been involved in defining new measures of well-being. In particular, social scientists became aware that the monetary dimension (measured by monetary variables only) is not the only dimension that needs to be studied (Betti and Lemmi, 2013).

One of the initial important steps was the so-called Index of Sustainable Economic Welfare (ISEW) developed by Herman Daly and John B. Cobb in 1989. In 1990, the economists Mahbub ul Haq and Amartya Sen developed the composite indicator Human
Development Index (HDI). The index detects the average results achieved by a country on three basic dimensions of human development: life expectancy, education, and income (Anand, and Sen, 1994).

A good illustration of a well-being framework can be seen in Italy. ISTAT (the Italian National Institute of Statistics) and CNEL (National Council for Economics and Labour) developed a well-being framework for Italy, the BES (Equitable and Sustainable Well-being). This framework has been heralded as the “result of an inter-institutional initiative which places Italy in the forefront of the international panorama for the development of well-being indicators going beyond GDP” (ISTAT March 2013). It is a very complex framework made up of 12 dimensions and 129 indicators. The following well-being dimensions are considered in the framework:

1. Health;

2. Education and training;

3. Work and life balance;

4. Economic well-being;

5. Social relationships;

6. Politics and Institutions;

7. Safety;

8. Subjective well-being;
9. Landscape and cultural heritage;

10. Environment;

11. Innovation, research and creativity;

12. Quality of services.

Each of these dimensions has a variety of single indicators. This is the well-being framework we are following in this thesis and in particular we will consider the “Economic well-being” dimension (number 4).

1.2 The problem of composite indicators

Although it is important to provide estimates (or predictions) of single indicators within framework dimensions (dashboard approach), e.g. single indicators for the economic well-being dimension, it is also crucial to provide composite estimates of dimensions or sub-dimensions. There is an ongoing debate about the suitability of bringing together multiple indicators of well-being into a composite one. Those against this argument stress the potential loss of social information, therefore they prefer a dashboard approach (see Ravaillon, 2011). Those in favor emphasize the fact that when we want to consider deprivations or well-being within the same individuals it is unavoidable to consider a composite index (Yalonetzky, 2012). Composite estimates of variables are needed when we have complex phenomena. For instance, if we have many well-being dimensions with ten single indicators in each dimension, composite statistics for each
dimension can be helpful to identify critical areas, and then to implement appropriate interventions. Furthermore, as Noll (2009) points out composite estimates are crucial in order to address the needs of policy makers asking for aggregate measures. Composite indicators are also important in order to present results in the media, where complex indicator systems may be difficult to be used. Aggregate measures are also helpful in order to give simple answers to the following general questions on living conditions:

- Are living conditions improving over time?
- Do citizens living in a particular area enjoy a better quality of life than those living in another area?

In order to build a composite indicator many steps need to be followed (OECD-JRC, 2008):

1. **Define theoretical framework**: this step provides the theoretical theory where elementary indicators are grouped into dimension to get an understanding of the multidimensional phenomenon. Here, experts and stakeholders need to be involved in order to avoid building a useless framework;

2. **Data selection**: data sources with variables need to be selected here for each indicator and dimension. Proxy variables are considered when data are scarce;

3. **Imputation of missing data**: this step is important to provide a complete dataset and to assess the impact of the imputation on the composite indicator results;
4. **Applied multivariate statistical methods**: this step is necessary to study the dataset structure and identify correlation relationships between elementary indicators. This step is crucial before weighing and aggregation apply. For instance, factor analysis techniques may be applied here;

5. **Normalisation**: here we make indicators comparable, for example, by scale adjustments. This step can also be used at the end, in order to rescale the composite indicators in some intervals (e.g. 0,1). This makes values comparable and interpretable between geographical areas;

6. **Weighting and aggregation**: this step should follow the theoretical framework developed in 1. Weights or scores estimated from multivariate statistical methods used in 4 can be used;

7. **Uncertainty and sensitivity analysis**: here we address the robustness and uncertainty of the composite indicators. In particular, the problem of variance estimation (uncertainty) of the composite indicators needs to be addressed here. Comparisons with other indicators developed in the literature may help in the sensitivity analysis;

8. **Back to data**: here we find the drivers of the composite indicator and causality analysis may be conducted. Moreover, it has to be checked whether some elementary indicators dominate the composite indicator;
9. **Links to other indicators**: composite indicators should be correlated with other indicators to find out relationship and support the sensitivity analysis;

10. **Visualisation of the results**: this step is crucial since can influence the interpretation of the composite indicators. The visualization technique needs to be chosen carefully.

Since the issue of building a composite indicator is a difficult one, involving a large variety of expertise and stakeholders, in this thesis we will focus on step 4, 6 and 7 only. Small area estimation methods will be integrated with these steps in order to provide small area estimates of multidimensional well-being indicators. Step 5 will be used to make the indicators comparable and we will visualise the final estimates in a way that is appealing to a larger audience, the map (step 10). Issues related to the other steps will be studied elsewhere.

### 1.3 The need for local estimates: small area estimation methods

In this thesis, the word “local” refers to any sub-regional geographical level identified as unplanned domain for which direct estimation may return large variance. The unplanned domains may have small or even zero sample sizes.

The European Commission has paid particular attention on measurement methods of well-being phenomena. In the last decade, several projects have been funded by the EU. For instance, two of the most important initiatives were SAMPLE (Small Area Methods
for Poverty and Living Condition Estimates) and AMELI (Advanced Methodology for European Laeken Indicators). Related to the SAMPLE project, an important point which is crucial to the construct of stress is the measurement of poverty and well-being indicators at the local level. Policy makers should use detailed information on the geographical distributions of social indicators in order to set policies against poverty and hunger. Hence, they should establish their decisions based on detailed information referring to appropriate geographical domains. As a consequence, poverty mapping has become a tool in order to design ad-hoc policies and interventions, such as providing social services and transfers in kind (Pratesi, 2016).

When we aim to provide estimates of target parameters for specific domains (such as means, totals and proportions), the problem of unplanned domains arises. This is because large-scale sample surveys are not designed to be representative for small domains. In order to manage this problem, we can follow two strategies (Chambers and Pratesi, 2013):

1. Increase the sample size of the domain;
2. Consider the use of small area estimation methods.

In this work we follow the second strategy. In the presence of small areas, direct estimation methods (e.g. the well-known Horvitz and Thompson estimator) of population parameters might not provide accurate and reliable estimates at a local level. In this case indirect small area estimation (SAE) methods provide more accurate estimates than direct design-based estimators (Pratesi, 2016). According to Rao and
Molina (2015: p. 1) a small area is any domain of a target population for which precise direct estimates cannot be produced. The idea of small areas is very broad and relative. In fact, these can be defined by cross-classification of geographical districts by social, economic, demographic characteristics or other subpopulations. Thus, in this context SAE techniques might be relevant for a state, province, municipality or school district for example. For instance, we can also identify a small area considering a particular ethnic group in a specific neighborhood within a municipality. SAE methods try to fill the gap between the released official statistics and the local governments’ demand for detailed local data.

A key resource for measuring well-being is the European Union Statistics on Income and Living Conditions (EU-SILC). It is well known that this data source has not been designed to be representative at sub-regional levels. In particular, the survey design for the EU-SILC data in Italy is such that it can be used to produce accurate direct estimates only at the NUTS (Nomenclature of Territorial Units for Statistics) 2-level, the regional level in Italy (Pratesi, 2016). In practical terms, this leads to serious consequences. Considering the case of Italy, this means that there might be some municipalities, the ones with very small population size, which have not been sampled, or with very few sampled units (the problem of unplanned domains). Traditional inferential methods do not allow calculating accurate estimates of small area parameters. Since EU-SILC is one of the most important EU surveys for investigating social exclusion and well-being phenomena in many EU countries, new inferential methods are strongly needed and necessary to overcome the technical estimation problems.
1.4 Multivariate Small Area Estimation and Data Dimensionality Reduction

Multivariate SAE is a research field still under investigation. Fuller and Harter (1987) propose the use of the multivariate mixed effects model in SAE. Datta et al. (1999) propose an empirical best linear unbiased predictor (EBLUP) and its mean squared error estimator for a mean vector of multiple characteristics. An interesting work on the integration of SAE with multivariate statistical methods is Fabrizi et al. (2016). Here, the problem of latent class models is studied. Although there are some works on the multivariate SAE problem, there is still an important gap in the literature on SAE and multivariate statistical methodologies (e.g. factor models) in the study of well-being. Namely, there is a gap about SAE methods for data dimensionality reduction. Satorra and Bentler (2014) deal with longitudinal factor models and produced small area estimates using an area-level approach. However, the issue of multivariate empirical best linear unbiased prediction (MEBLUP) for factor score means compared to the use of a dashboard of single indicators under the unit-level SAE approach has not been investigated so far.

In this thesis we deal with the problem of composite indicators when multidimensional conceptual frameworks are already available, and a set of variables (and indicators) for each well-being dimension has already been defined and approved by governments and Official Statistics (e.g. the case of the Italian BES). Therefore, in order to provide composite estimates, which show an economic interpretation, we consider well-being as
a multidimensional phenomenon, and use factor analysis models in a confirmatory approach only. In this view, the factor analysis models are based on substantive theory and/or previous exploratory research. From a confirmatory factor analysis model we estimate the latent factor scores and we treat these as composite estimates. At this stage, EBLUP approach is used to estimate the mean of the factor scores at small area level. Here, univariate and multivariate EBLUPs are investigated in the context of data dimensionality reductions using factor scores and compared to the case of weighted and simple averages of single small area estimates. Our procedure is a two-step procedure, where factor scores can be seen as plug-ins in a mixed-effect model for small area estimation. However, the mean squared error estimation needs to take this into account. In order to evaluate our method, in the simulation studies in chapters 3 and 4 the observed variables are first generated from the multivariate mixed effect model by Fuller and Harter (1987), this is the true model; the latent factor scores are estimated and considered as “plug-ins”. This choice has been done in order to take control for the intra-class correlation and correlations in the covariance-variance matrices of the observed variables.

1.5 Research Questions

In this thesis we address the following research questions:

1. Does the use of factor scores in data dimensionality reduction for multidimensional latent well-being measurement provide better estimates (in
terms of bias and variability) compared to the traditional use of a dashboard of single indicators?

2. How can we estimate the mean squared error (MSE) of EBLUP of factor score means taking into account the factor analysis model variability? If this variability is ignored, is the MSE biased?

3. Similar to the first research question, we consider the issue of multiple latent factors and their comparison to a dashboard of single indicators estimated via a multivariate SAE. Does the multivariate SAE improve on the univariate SAE? If correlations are accounted for a priori through a factor analysis model, how will this impact on the multivariate SAE approach compared to the univariate SAE for factor scores means?

4. The final research question is related to the problem of mean squared error estimation in multivariate SAE under a parametric bootstrap method. How does our proposed parametric bootstrap algorithm perform for approximating the MSE of a multivariate EBLUP?

1.6 Rationale for Alternative Format

Each of the main chapters, Chapters 3, 4, and 5 are stand-alone research papers submitted (and under review) to international journals. The papers focus on the same research area: multivariate SAE in multidimensional well-being measurement, and they
are strictly connected, as is described in the following section 1.6. The papers are co-authored with my supervisors, Natalie Shlomo and Joseph Sakshaug. The main author, Angelo Moretti, designed the research questions at the time of the presentation of the Ph.D. proposal. He conducted the analysis and prepared the paper drafts to submit to the supervisors prior to the supervision meetings. Natalie Shlomo and Joseph Sakshaug helped him to shape the research questions and writing the final draft to be submitted to international journals.

In order to be able to have a list of figures and tables, we did not interrupt the numbering of these throughout the thesis. Equations’ numbers and appendices refer to each chapter.

1.7 The Aims and Structure of this Thesis

In this thesis we address the problem of providing reliable small area estimation of multidimensional economic well-being phenomena starting from an established well-being measurement framework (e.g. the Italian BES) already developed and used by specific countries. Therefore, the issue of providing composite estimates is treated in a confirmatory way: latent dimensions are already identified in advance. We are following a two-step procedure: first the identified latent variables are estimated and second the small area estimates along with measures of uncertainty are computed. Particularly, in the simulation, the observed variables were generated from a super population model (true model). This has been done to take in control for the intra-class correlation coefficient (ICC) which changes after a confirmatory factor analysis model is employed.
due to changes in the correlation structure. The role of the ICC is important to check the performance of the multivariate EBLUP compared to the univariate EBLUP.

Chapter 1 is the introduction and the background of the research problem.

Chapter 2 is a literature review and an overview of SAE methods. First, we discuss the problem of the dimensionality reduction of a dataset using factor models. We also describe the multivariate small area estimation problem for a vector of means as a target parameter. Only the most relevant research papers available in the current literature are reviewed here as each stand-alone paper in the subsequent chapters also contains relevant literature review.

Chapter 3 is about the use of factor scores for data dimensionality reduction. Particularly, we address the issue of using a factor analysis model in the context of small area estimation. We consider a single latent variable as a proxy of multidimensional economic well-being indicators. The use of factor scores is compared to weighted and simple averages of standardized univariate EBLUPs. In this chapter, we also propose a bootstrap procedure for mean squared error estimation of the EBLUP of factor score means. We show that factor scores provide more accurate estimates of well-being phenomena than weighted and simple averages. Since we use a two-step approach to provide small area estimates of multidimensional well-being, this needs to be included in the mean squared error estimator. Indeed, we first estimate the factor scores from a factor analysis model and this variability needs to be considered.
Chapter 4 is about the use of multivariate SAE under the EBLUP approach in data dimensionality reduction. The use of a dashboard of univariate and multivariate EBLUPs (MEBLUP) is compared with the use of factor score means (estimated via EBLUPs and MEBLUPs). Here, we focus on one and two well-being dimensions (latent variables). Therefore, the use of multivariate SAE modeling can be compared to the univariate in the form of factor scores and a dashboard of observed variables (or vectors). We investigate the performance of the EBLUP before and after data dimensionality reduction.

Chapters 3 and 4 present simulation studies as well as applications using real EU-SILC data for the Italian Tuscany region. In Chapter 3 we propose an application of economic well-being indicators in the form of a single latent factor; and in Chapter 4, we study a sub-dimension of economic well-being: housing quality with two latent factors.

Chapter 5 is a methodological chapter on a proposed approach for a parametric bootstrap to estimate the mean squared error of a multivariate EBLUP. We evaluate the algorithm via a simulation study and a short example based on the well-known LANDSAT data (Batteese, et al. 1988). This bootstrap procedure is applied in Chapter 4 where it is particularized considering the variability arising from the factor analysis model with two latent factors via the main results obtained in Chapter 3 for the univariate case.

The simulation studies conducted in this thesis belong to the group of “design-based under model data” simulations (Münnich, 2014). One outcome of the multivariate
mixed-effect model (Fuller and Harter, 1987) is used as fixed universe and the rest of the study is a design-based simulation. This outcome is the vector of observed responses. The factor scores are then estimated in each sample. It was not possible to generate many universes since our computations are very intensive and we did not have available powerful computers.

Simulation studies are important when analytical results about properties of distributions (e.g. sampling distributions) cannot be provided otherwise. Also, as Münnich (2014) points out, we can learn from simulation studies about the applicability of different types of estimators in various scenarios, since in practice we have only one sample available. Moreover, simulations are helpful in order to discover peculiarities of estimators which can be hardly found via mathematical proofs (Münnich, 2014). In order to make use of simulations effectively, these have been informed by empirical data, i.e. the simulation parameters are estimated from real EU-SILC data and unplanned domains are generated, since this is typical in the Italian EU-SILC geography. Our simulations do not take into account the case of model assumptions violations (e.g. outliers, heteroscedasticity and non-normality). As stated in the conclusion of this thesis, these topics will be subject of future research since we did not encounter such issues in the data used in this thesis.

Chapters 3 and 4 give a contribution to the use of EBLUP approach (univariate and multivariate) in data dimensionality reduction literature. Chapter 3 gives a contribution to survey statistics literature in terms of mean squared error of an EBLUP of factor scores mean. In both chapters we propose the use of factor scores as composite estimates in order to provide small area estimates of multidimensional well-being
indicators where dimensions and single indicators are a priori developed through government frameworks. Chapter 5 contributes to multivariate SAE literature. The mean squared error estimation problem of a multivariate EBLUP in case of continuous variables is addressed.
2 Literature Review and Methods Overview

In this chapter we provide a literature review and discuss the issue of composite indicators estimated from factor analysis models and small area estimation methods used throughout the thesis.

2.1 Composite Indicators

In the last two decades, the literature has been focused on the measurement of multidimensional poverty and well-being indicators in a traditional dashboard approach. In this view, single well-being indicators are presented separately with no attempt to determine their relative importance (Drabsch, 2012). As it is claimed in Drabsch (2012), this approach avoids subjective decisions about the relative importance of different single indicators. Consequently, the focus has been mainly on univariate small area estimation (SAE) methodologies to provide separate indicators estimated for local areas in a model-based approach.

In the last ten years, complex well-being measurement frameworks have been developed. ISTAT (the Italian National Institute of Statistics) and CNEL (National Council for Economics and Labour) developed a well-being framework for Italy, the BES (Equitable and Sustainable Well-being). According to BES, well-being is measured
on the basis of 12 social and environmental dimensions: health, education and training, work and life balance, economic well-being, social relationships, politics and institutions, security, subjective well-being, landscape and cultural heritage, environment, research and innovation, and quality of services. This is the framework we are referring to in this thesis.

Many social concepts such as social exclusion and well-being phenomena are multidimensional. Therefore, in the measurement stage, we should be considering multiple dimensions. Amartya Sen has explored these problems in a large number of studies (e.g. Sen, 1985; Sen, 1999). In the social sciences there have been a large number of efforts to deal with the problem of the multidimensionality of well-being and poverty indicators. It is generally agreed that the monetary dimension alone is not sufficient to be analysed (Lemmi and Panek, 2016). Townsend was one of the first scholars to point out that the monetary dimension is not enough in poverty measurement. He proposed to consider also other dimensions i.e. dwelling conditions, affluence, education, professional and financial resources (Abel-Smith and Townsend, 1965; Townsend, 1979). Due to the developments in the social indicators literature there is a need for new statistical methodologies to construct multidimensional measures.

Traditional multivariate statistical techniques, such as factor and principal component analysis, as well as structural equation models can be used to reduce the data dimensionality of well-being indicators. Some examples on the use of factor analysis in latent well-being in order to reduce data dimensionality can be seen in Ferro Luzzi, Fluckiger, and Weber (2008) and Gasparini et al. (2011). In particular, Ferro Luzzi,
Flückiger, and Weber (2008) apply factor analysis techniques on a set of 32 observed variables from the Swiss Household Panel. In this work, they reduce the data dimensionality from 32 variables to four latent factors related to financial, health, neighborhood and social exclusion dimensions. After this step, cluster analysis is used in order to find the ‘poor’ group on the basis of the latent factors. Gasparini et al. (2011) investigate the Gallup World Pool starting from 15 observed variables. Three latent factors are selected: income, subjective perceptions of well-being, and basic needs. They found that income has a large impact on multidimensional well-being, although is not the only factor involved.

As mentioned in Chapter 1, there is an ongoing debate on composite indicators in well-being and poverty measurement. For example, Ravallion (2011) claims that a single multidimensional composite indicator leads to a loss of information, and on the other hand, Yalonetsky (2012) points out that composite indicators are necessary when the aim is to measure multiple deprivations within the same unit (individual or household). According to OECD-JRC (2004), when single indicators are compiled into a single indicator on the basis of an underlying model related to measured multidimensional concepts, we obtain a composite indicator. In the composite indicators literature there is an ongoing debate on the formulation of composite estimates. On the one hand, composite indicators are easier to interpret than a dashboard of single indicators; on the other hand, if they are poorly constructed or misinterpreted, they can produce serious misleading policy messages (OECD-JRC, 2008).
The process of constructing composite indicators requires several steps. As reported by OECD-JRC (2008), we need to consider ten steps. First of all, we need to understand how to select and combine variables related to well-being. In this step, many experts and stakeholders should be involved. This step is particularly important in order to get a clear comprehension and definition of the multidimensional phenomenon. Then, relevant datasets should be selected, and consequently we may need to impute missing data. An important step is the normalization due to the different scales of the observed variables. A crucial step is the weighting and aggregation process. This is the step considered in this thesis. Here, the suitable weighting and aggregation procedures are considered and formulated. In addition, we need to investigate whether the correlation among indicators should be accounted for. The other steps are related to the assessment of the robustness of the indicators and linking the indicators to each other. The last step is the visualisation of the results.

In order to monitor these phenomena Laeken indicators, which consist of 18 single indicators on social exclusion were proposed in 2001 (Lemmi and Panek, 2016). Also, in each country of the European Union, governments developed complex multidimensional well-being frameworks with many single indicators combined into dimensions such as the Italian framework BES (Equitable and Sustainable Well-being) 2015 (ISTAT 2015).

When well-established well-being frameworks are available the confirmatory factor analysis approach can be used, and this is the research that we are focusing on in this thesis. In fact, as pointed out in Sosa-Escudero, Caruso, and Svarc (2013), the vector of
unobserved (latent) variables represents a set of variables that jointly describe the underlying multivariate phenomenon. Factor analysis is commonly used in this field for data dimensionality reduction issues; see for example Ferro Luzzi, Fluckiger, and Weber (2008) and Gasparini et al. (2011).

Once calculated from factor analysis (FA) models, factor scores are used in the literature as composite measures as part of regression modelling or predictive analysis. This is a two-step procedure, where factor scores are first estimated and then used in regression modeling as we proposed in this thesis in small area models. Kawashima and Shiomi (2007) propose the use of factor scores in order to conduct an ANOVA analysis on high school students’ attitudes towards critical thinking and tested differences by grade level and gender. Bell, McCallum, and Cox (2003) study the reading and writing skills where they extracted the factors and estimated factor scores before using them in a multiple regression analysis model. Factor scores estimated from factor analysis models can be used to reduce data dimensionality and are used in small area models. As it is noted in Skrondal and Laake (2001), using factor scores as dependent variables in regression modelling produces consistent estimates of model parameters since any measurement error from the factor analysis model is absorbed into the prediction error and coefficients are not attenuated. Also, as discussed in Kaplan (2009), we can assume that the specific variances from the factor analysis model are very small compared to the prediction error. First attempts in small area estimation (SAE) and data dimensionality reduction using FA models can be seen in Smith et al. (2015). Here, the construction of
the composite indicators was on the small area EBLUPs of the single indicators, thus a dashboard approach was used.

2.2 Factor analysis model and the use of factor scores

Factor analysis (FA) is a basic multivariate statistical method which can be used to reduce the dimensionality of a multivariate random variable $Y$ (Härdle and Simar, 2012: p. 307). Each factor is interpreted as a latent characteristic of the individuals revealed by the original variables. In factor analysis models we identify a limited number of latent variables, expressed by factors. From a geometric point of view, we move from a larger space to a smaller one.

2.2.1 The linear factor analysis model

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

$$Y = \Lambda f + \epsilon,$$  \hspace{1cm} (1)

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

It is assumed that:

i) $E(\epsilon) = 0$,

ii) $\text{Var}(\epsilon) = \Theta$,

iii) $\epsilon \sim N(0, \Theta)$,

iv) $\epsilon$’s components are uncorrelated,

v) $E(f) = 0$,

vi) $\text{Cov}(\epsilon, f) = 0$,

vii) $\epsilon \sim N(0, \Theta)$.

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

$$\Sigma = \text{Cov}(YY') = \Lambda E(ff')\Lambda' + E(\epsilon\epsilon') = \Lambda \Phi \Lambda' + \Theta,$$  \hspace{1cm} (2)

where $\Phi$ is a $M \times M$ matrix of factor variances and covariances, and $\Theta$ 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 do not have closed solution, so iterative numerical algorithms are proposed in the literature (see e.g. Mardia et al., 1979). The log-likelihood function $\ell$ of the data $Y$ can be written as follows (Härdle and Simar, 2012: p. 316):
\[ \ell(Y; \Lambda, \Theta) = \frac{n}{2} \left[ \log(2\pi(\Lambda \Lambda' + \Theta)) \right] + \text{tr}\left( (\Lambda \Lambda' + \Theta)^{-1} \Sigma \right). \] 

(3)

where \( \Sigma \) denotes the empirical covariance of \( Y \) (estimator of \( \Sigma \)).

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 estimates for \( i = 1, \ldots, n \) are given by (Härdle and Simar, 2012: p. 323 and Thurstone (1935: pp. 226-231)):

\[ \hat{f}_i = \hat{\Lambda}' \hat{\Sigma}^{-1} y_i. \] 

(4)

where \( \hat{\Lambda} \) denotes the estimator of \( \Lambda \). It can be seen that (4) is a plug-in.

“Regression method” comes from the regression terminology. Here, the independent variables are the standardized observed variables. These are weighted by regression coefficients obtained by multiplying the inverse of the observed variable correlation matrix by the matrix containing the factor loadings and in case of oblique transformations by the factor correlation matrix.
Bartlett’s method can be also used to estimate the individual factor scores for $i = 1, ..., n$. The estimator is given by (Bartlett, 1937):

$$\hat{f}_i = \hat{\Phi} \hat{\Theta}^{-1} y_i. \quad (5)$$

Where $\hat{\Phi} = \hat{\Lambda} \hat{\Theta}^{-1} \hat{\Lambda}$ and $y_i$ denotes a $K$-dimensional vector of observations of $K$ components of $Y$ for $i = 1, ..., n$.

Bartlett’s method produces unbiased estimates of the true factor scores (Hershberger, 2005).

In the case of both binary and continuous observed variables under a Maximum Likelihood Estimation approach, the factor scores can be estimated via the Expected Posterior Method described in Muthén (2012).

Nicoletti et al. (2000) propose a statistical approach based on factor analysis to formulate composite indicators. FA shows, within each dimension of single indicators, which are more associated with the underlying factors.

### 2.2.2 The choice of the number of factors

According to OECD-JRC (2008) and classical multivariate statistical literature (Bartholomew et al., 2008), factors should be chosen according to the following criteria:
1. They have associated eigenvalues larger than one;

2. They contribute individually to the explanation of overall variance by more than 10%;

3. They contribute cumulatively to the explanation of the overall variance by more than 60%.

Among the goodness of fit indices for confirmatory factor analysis models, we have the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). The RMSEA ranges from 0 to 1, with smaller values indicating better model fit and a value of 0.06 or less is indicative of acceptable model fit. CFI values range from 0 to 1, with larger values indicating better fit. CFI value of .95 or higher is accepted as an indicator of good fit (Hu & Bentler, 1999).

### 2.3 Two-level factor analysis models

In the latent variables modeling literature there are several studies on multilevel factor analysis models. For a good review about these we refer to Longford and Muthén (1990), and Kaplan (2009). In this section, for the economy of space, we introduce the general idea only.

In multilevel models (two-levels) the unit $i$ is nested in the area $d = 1, \ldots, D$. This means that we have a hierarchical structure in the population. Thus, the vector of observed variables $Y_{di}$ now refers to the unit $i$ in the area $d$. The total sample covariance matrix for $Y$ can be written as $V(Y) = \Sigma = \Sigma_B + \Sigma_W$, which means that the variance can be
decomposed into a between and within part (Kaplan, 2009). Therefore, under model assumption specified in 2.2.1, model (1) can be written in a multilevel setting as

\[ Y = \Lambda_B f_B + \Lambda_W f_W + \epsilon_B + \epsilon_W. \]  

(6)

where \( \Lambda_B \) and \( f_B \) denote the loadings and factor for the between-model part, respectively, and \( \Lambda_W \) and \( f_W \) denote the same quantities for the within-model part. \( \epsilon_B \) and \( \epsilon_W \) denote the residual for the area-level and unit-level variation, respectively.

Although in SAE we need to consider the multilevel structure in the population of interest, it is not always possible to use these types of FA models described in (6) due to the issue of unplanned domains. This means that the sample sizes in the unplanned domains can be small and even zero which leads us to use a model-based approach for estimation. In case of small or zero \( n_d \) the two-level factor analysis model will encounter problems in the parameters estimation algorithms, and as a result convergence issues arise. Also, when \( n_d = 0 \) it is not possible to estimate the within area factor. Unplanned domains are typical in the EU-SILC survey at levels under NUTS-2. Earlier simulation studies (not shown here) show that the two level factor analysis model is not applicable for unplanned domains. Therefore, this requires attention in future research.
2.4 Model- versus design-based inference

Modern statistical inference faces an important debated: the use of model-based versus design-based approach. Model-based approach gives important advantages in small area estimation (Rao and Molina, 2015: 5):

- Optimal estimators can be built under an assumed model;
- Models can be validated (via model diagnostic) from the sample;
- Models are very versatile and can be formulated considering the complexity arising from the sample.

However, failures of model assumptions and misspecification need to be carefully evaluated in this approach otherwise we may introduce a bias in the estimates (Rao and Molina, 2015: 5).

In the design-based approaches, statistical models are often used in order to construct the estimators (model-assisted approach). However, bias and variance of these estimators are evaluated under the sample distribution accounting for inclusion probabilities and design features. On the other hand, model-based techniques’ inference is based on the underlying model (Pfeffermann, 2013).

A distinctive attribute common to both design- and model-based SAE approaches is the use of auxiliary variables, obtained from administrative data or the Census. The use of the auxiliary variables in SAE depends on specific approaches (Pfefferman, 2013).
2.5 The use of the multivariate nested error regression model

In the study of multidimensional well-being, we wish to estimate means vector of multivariate characteristics for a small area $d$. A further crucial work is Fuller and Harter (1987). Here, the multivariate regression model with variance components structure is considered. The problem of the prediction of mean vectors has been studied by Datta et al. (1999) as well. In this work, the multivariate variance components model by Fuller and Harter (1987) was used to derive EBLUP (Empirical Best Linear Unbiased Predictor) and EB (Empirical Best) of small area mean vectors. Furthermore, they derive the first two components of the MSE of the multivariate EBLUP. Molina (2009) studied a multivariate mixed effects model considering exponentials of mixed effects, and these are predicted using an empirical bias-corrected predictor. The methods which have been mentioned follow a unit-level approach, and unfortunately, micro data for auxiliary variables might not be available for the units of the population. Thus, an area-level approach can be used in those situations. Due to the aim of this work, we will not discuss the area-level SAE approach.

2.5.1 Notation

Let $d = 1, ..., D$ denote the small areas for which we want to compute the target estimates, and for a sample $s \subset \Omega$ of size $n$ drawn from the target finite population 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 $N_d - n_d$ dimension.

$y_{di} = (y_{d1i}, ..., y_{dki})'$, which denotes the $K \times 1$ vector of interest for $i = 1, ..., N_d, d = 1, ..., D$, we can write the target mean vector as follows:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}.$$  \hfill (7)

2.5.2 The multivariate SAE problem for a vector of means

Because of the linearity of (7), each area means vector can be split into sampled and non-sampled elements as follows:

$$\bar{Y}_d = N_d^{-1} \left( \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} y_{di} \right).$$  \hfill (8)

The quantity $\sum_{i \in r_d} y_{di}$ is not observed, so it needs to be predicted. Unit-specific auxiliary variables $x_{di}$ are available by assumption, for all the population elements in each small area $d$. Also, we assume that the following linear model relates the response variables to the covariates as follows:

$$y_{di} = x_{di} \beta + u_d + e_{di}, \quad d = 1, ..., D, i = 1, ..., N_d,$$

$$u_d \sim N_K(0, \Sigma_u), \quad e_{di} \sim N_K(0, \Sigma_e) \quad u_d \text{ and } e_{di} \text{ independent}.$$
Where \( y_{di} \) is the response vector for the \( i \)th unit from the \( d \)th small area, \( x_{di} \) is the \( K \times p \) matrix of the auxiliary variables, \( \beta \) is a \( K \times p \) matrix of unknown regression coefficients, \( u_d \) is the \( k \times 1 \) vector of the area effects, and \( e_{di} \) is the \( K \times 1 \) vector of the individual effects. Here, the \( K \times K \) positive-definite matrices \( \Sigma_u \) \( \Sigma_e \) are the covariance matrices of the area effects and individual effects, respectively.

Hence, the Best Linear Unbiased Predictor (BLUP) of (8) under model (9) is given by (e.g. Royall (1970) for univariate case, this can be extended to the multivariate case since the linearity of the vector of means):

\[
\hat{Y}_{BLUP}^d = N_d^{-1} \left( \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} \tilde{y}_{di} \right),
\]

(10)

Where \( \tilde{y}_{di} = x_{di}^T \tilde{\beta} + \tilde{u}_d \), \( \tilde{u}_d \) denotes the BLUP of \( u_d \). In practice, we calculate the empirical BLUP of \( \tilde{Y}_d \) due to unknown variance components, \( \Sigma_u \) and \( \Sigma_e \). Therefore, these quantities need to be replaced by consistent estimators \( \hat{\Sigma}_u \) and \( \hat{\Sigma}_e \). According to Schafer et al. (2002) in order to get the maximum-likelihood (ML) estimates we apply the Fisher-scoring algorithm that maximises the ML function. For more details on the algorithm we also refer to Cressie (1992). The random effects are estimated via the Empirical Bayes method (Schafer et al., 2002).

Let \( \tilde{\beta}_{GLS} \) and \( \tilde{u}_d \) be the results of replacing the estimates of the variance components, the multivariate EBLUP (MEBLUP) of \( \tilde{Y}_d \) can be written as (Harter and Fuller, 1987):
$$\hat{\mathbf{Y}}_d^{MEBLUP} = \tilde{\mathbf{X}}'_d \hat{\mathbf{B}} + \tilde{\mathbf{u}}_d = \bar{\mathbf{X}}'_d \hat{\mathbf{B}} + (\bar{\mathbf{Y}}_d - \bar{\mathbf{X}}'_d \hat{\mathbf{B}})[(\Sigma_u + n_d^{-1}\Sigma_e)^{-1}\Sigma_u],$$

where $\bar{\mathbf{X}}_d$ are the known population covariates means.

Datta et al. (1999) use the EM algorithm to find ML and REML model estimates, and proposed a simulation experiment to study the multivariate SAE problem. They point out that the percentages of reductions in terms of mean squared error (MSE) of the multivariate analysis over the univariate analysis depend on $\rho_e$, $\rho_u$ and $r_k = \sigma_{ujj}/\sigma_{ejj}$, where $\rho_e$ and $\rho_u$ are the correlation coefficients associated with $\Sigma_e$ and $\Sigma_u$ (correlations of the error terms and random effects), respectively. $k = 1,...,K$ denotes the $k^{th}$ component of the vector $\mathbf{y}_{di}$, $\sigma_{ujj}$ and $\sigma_{ejj}$ are the diagonal elements of $\Sigma_u$ and $\Sigma_e$, respectively, and $r_k$ is a function of the intra-class correlation defined as $\rho_k = \sigma_{ujj}^2/(\sigma_{ujj}^2 + \sigma_{ejj}^2)$. When the signs of $\rho_e$ and $\rho_u$ are opposite (the correlations of the random errors and area effects are different) the percentages of reductions in terms of MSE are higher than in the other cases. These refer to the correlations of the random errors and random effects; in multivariate mixed-effect models we have multiple random errors and random effects. Also, these depend on the magnitude of the correlations. These results are also showed theoretically in Datta et al. (1999).
2.6 Mean Squared Error Estimation

The MSE of the MEBLUP $\hat{Y}_{d}^{MEBLUP}$ denoted by $mse(\hat{Y}_{d}^{MEBLUP})$ is given by the following:

$$mse(\hat{Y}_{d}^{MEBLUP}) = E\left[\left((\hat{Y}_{d}^{MEBLUP} - \bar{Y}_{d})(\hat{Y}_{d}^{MEBLUP} - \bar{Y}_{d})\right)^{'}\right] =$$

$$= mse(\hat{Y}_{d}^{MBLUP}) + E\left[\left(\hat{Y}_{d}^{MEBLUP} - \hat{Y}_{d}^{MBLUP}\right)(\hat{Y}_{d}^{MEBLUP} - \hat{Y}_{d}^{MBLUP})\right)^{'} +$$

$$+ E\left[\left(\hat{Y}_{d}^{MBLUP} - \bar{Y}_{d}\right)(\hat{Y}_{d}^{MEBLUP} - \hat{Y}_{d}^{MBLUP})\right]^{'}. $$

Using the normality assumption and the main results obtained in Kackar and Harville (1984) it can be showed that the last two terms of equation (12) are equal to zero for any unbiased and translation invariant estimator of the variance components.

For $mse(\hat{Y}_{d}^{MBLUP})$, which is the MSE of the multivariate BLUP of $\hat{Y}_{d}$, the analytical estimators can be calculated, and it is given by the following (Datta et al. 1999):

$$mse(\hat{Y}_{d}^{MBLUP}) = G_{1d}(\psi) + G_{2d}(\psi).$$

(13)
where $\boldsymbol{\psi}$ denotes a $K(K+1)$ component vector of variance parameters of which $\text{mse}(\hat{\mathbf{y}}_{d}^{MBLUP})$ is a function. For details on the expressions $G_{1d}(\boldsymbol{\psi})$ and $G_{2d}(\boldsymbol{\psi})$ we refer to Datta et al. (1999).

Given these assumptions:

i) $\hat{\boldsymbol{\psi}}(-\mathbf{Y}) = \hat{\boldsymbol{\psi}}(\mathbf{Y})$

ii) $\hat{\boldsymbol{\psi}}(\mathbf{Y} + Xg) = \hat{\boldsymbol{\psi}}(\mathbf{Y}), \ \forall g, \mathbf{Y}$

The MSE of $\hat{\mathbf{y}}_{d}^{MEBLUP}$ is given by

$$\text{mse}(\hat{\mathbf{y}}_{d}^{MEBLUP}) =$$

$$= G_{1d}(\hat{\boldsymbol{\psi}}) + G_{2d}(\hat{\boldsymbol{\psi}})$$

$$+ E\left[ (\hat{\mathbf{y}}_{d}^{MEBLUP} - \hat{\mathbf{y}}_{d}^{MBLUP}) (\hat{\mathbf{y}}_{d}^{MEBLUP} - \hat{\mathbf{y}}_{d}^{MBLUP})' \right]$$

(Approximations of the last term are proposed in the literature for large $D$ (see, for example, Kackar and Harville, 1984; Prasad and Rao, 1990).

A second-order accurate estimate of $\text{mse}(\hat{\mathbf{y}}_{d}^{MEBLUP})$ is in Datta et al. (1999) and it is given by:

$$\text{mse}(\hat{\mathbf{y}}_{d}^{MEBLUP}) = G_{1}(\hat{\boldsymbol{\psi}}) + G_{2}(\hat{\boldsymbol{\psi}}) + 2G_{3}(\hat{\boldsymbol{\psi}}) + Q(\hat{\boldsymbol{\psi}}).$$

The reader may refer to Datta et al. (1999) for more details on $Q(\hat{\boldsymbol{\psi}})$ and assumptions.
In the univariate SAE literature the use of the Prasad-Rao MSE analytical approximation (Prasad and Rao, 1990) is diffused in order to approximate the MSE of the univariate EBLUP. González-Manteiga et al. (2008a) evaluates the case of estimation of the MSE for the univariate EBLUP compared to bootstrap estimators. As Datta and Lahiri (2000) highlight, the Prasad-Rao approximation considers only ANOVA estimates of variance components; however, ML and the REML methods are also used to produce EBLUP in small area estimation and in these cases, the Prasad-Rao theory is not useful (Datta and Lahiri, 2000). The Prasad-Rao approximation can be used to provide MSE analytical approximations under multivariate Fay-Herriot models (e.g. Benavent and Morales, 2016).

2.6.1 Resampling techniques to approximate the MSE

Resampling techniques, such as jackknife and bootstrap, are widely used in SAE and more generally in survey sampling to approximate the MSE of a predictor when these are not available in exact form (González-Manteiga et al., 2008a). They are also useful when analytical approximations in SAE are a challenging task. Among resampling techniques, as Lahiri (2003) points out, the bootstrap is the most flexible and efficient method in survey sampling. This is due to the fact that can be used to solve many issues in complex surveys, such as small area estimation problems and both for smooth and unsmooth target parameters. However, the procedure can be computer intensive compared to linearization methods.
González-Manteiga et al. (2008a) highlight another important advantage of the bootstrap method: the bootstrap may provide more accurate MSE estimates than the analytical approximation because of the second-order accuracy property, which is not usually satisfied by asymptotic (or linear) approximations.

2.7 The data used in this thesis

In chapters 3 and 4 data from the Italian EU-SILC 2009 and Census 2001 were used, while for the application of chapter 5 we used data from LANDSAT (Battese et al., 1988).

EU-SILC is a sample survey coordinated by EUROSTAT at the EU level and conducted yearly by ISTAT for Italy. The survey is designed to produce accurate estimates at the national and regional levels (NUTS-2). Therefore, for the Italian geography the survey is not representative of provinces, municipalities (NUTS-3 and LAU-2 levels, respectively), and lower geographical levels. The regional samples are based on a stratified two-stage sampling design and in particular, 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 and SSUs are selected by systematic sampling in each selected PSU.
The EU-SILC is an important sample survey used to provide data on the structural indicators of social cohesion and for monitoring poverty and social inclusion indicators in the EU.

Due to the methodological nature of chapter 5, data from LANDSAT only were used. The data comprises survey and satellite data for corn and soy beans for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and from land observatory satellites during the 1978 growing season.

2.8 Gaps in the literature: multivariate SAE, composite indicators and multidimensional social indicators

There are still many open questions in multivariate SAE in the measurement of multidimensional well-being indicators. First, the use of multivariate EBLUP in data dimensionality reduction has not been addressed yet. In particular the evaluation of multivariate statistical analysis techniques, such as factor analysis models, needs to be considered and compared with the traditional use of a dashboard of EBLUPs of single indicators in the multivariate SAE approach. These issues are addressed in Chapters 3 and 4. The problem of MSE estimation taking into account the variability arising from factor analysis models in the EBLUP estimation of factor score means needs attention and is addressed in Chapter 3. In Chapter 4, we investigate the case of multiple latent factors and we deal with multivariate SAE under a multivariate mixed effects model. Bootstrap estimators have been proposed in order to approximate the MSE of univariate
EBLUPs under many different types of mixed effects models; we refer to Rao and Molina (2015) for a literature review on this. The estimation of the MSE of a multivariate EBLUP under multivariate mixed effects models needs to be investigated using the bootstrap setting and is addressed in Chapter 5. Since this research field is still ongoing, it is important to evaluate the related issues under real data scenarios in order to provide guidelines to policy makers and official statistics on the potential and problems of these methodologies.
3 Small Area Estimation of Latent Economic Well-being

Introduction to the paper

This chapter is a paper submitted to Sociological Methods and Research and the reviewers recommended publication as outcome of the revision process. I am the lead author of the paper and responsible for the writing of the article and carrying out all the analysis and simulation studies. All ideas and approaches are discussed through the normal supervision process.

This paper was also submitted and accepted for oral presentation at the international conference in Small Area Estimation (Maastricht, 2016).

In this paper we introduce the problem of univariate empirical best linear unbiased prediction in data dimensionality reduction. In particular, we address the use of factor scores as composite estimates of multidimensional well-being indicators and compare these to a dashboard of single indicators. Our method is evaluated via simulation and enriched with a real data application using Italian EU-SILC 2009 and Population Census 2001 data.
Small Area Estimation of Latent Economic Well-being

Angelo Moretti, Natalie Shlomo, and Joseph Sakshaug

Social Statistics, School of Social Sciences, University of Manchester, United Kingdom

Abstract

Small area estimation (SAE) plays a crucial role in the social sciences due to the growing need for reliable and accurate estimates for small domains. In the study of well-being, for example, policy-makers need detailed information about the geographical distribution of a range of social indicators. We investigate data dimensionality reduction using factor analysis models and implement SAE on the factor scores under the empirical best linear unbiased prediction approach. We contrast this approach with the standard approach of providing a dashboard of indicators, or a weighted average of indicators at the local level. We demonstrate the approach in a simulation study and a real data application based on the European Union Statistics for Income and Living Conditions (EU-SILC) for the municipalites of Tuscany.

Keywords: Composite estimation; Direct estimation; EBLUP; Factor analysis; Factor scores; Model-based estimation.
3.1 Introduction

Measuring poverty and well-being is a key issue for policy makers requiring a detailed understanding of the geographical distribution of social indicators. This understanding is essential for the formulation of targeted policies that address the needs of people in specific geographical locations. Most large-scale social surveys can only provide reliable estimates at a national level. A relevant survey for analyzing well-being in the European Union (EU) is the European Union Statistics for Income and Living Conditions (EU-SILC). However, these data can only be used to produce reliable direct estimates at the NUTS (Nomenclature of Territorial Units for Statistics) 2 level (Giusti, Masserini and Pratesi, 2015) which are generally large regions within a country. For example, in Italy one such NUTS 2 region is Tuscany. Hence, if the goal is to measure poverty and well-being indicators at a sub-regional level, such as NUTS 3 or LAU (Local Administrative Units) 2, the indicators may not be directly estimated from EU-SILC. In fact, the domains corresponding to the regions under NUTS 2 are so-called unplanned domains where domain membership is not incorporated in the sampling design, and therefore the sample size in each domain is random (and may be large or small) and in many cases zero. In this case, indirect model-based estimation methods, in particular small area estimation approaches, can be used to predict target parameters for the small domains.

Small area estimation (SAE) is defined as a set of statistical procedures with the goal of producing efficient and precise estimates for small areas, as well as for domains with
zero sample size (Rao and Molina, 2015: Ch. 1). An area is defined as small, if the area-specific sample size is not large enough to provide precise and efficient direct design-based estimates. Small areas can also be defined by the cross-classification of geographical areas by social, economic or demographic characteristics.

SAE methods can be classified into two approaches: the unit-level and the area-level approach. The unit-level approach is used when covariates are available for each unit of the population, for example from census or administrative data, while the area-level approach is used when covariate information is known only at the area level. The use of the error-components model by Battese, Harter and Fuller (1988), also known as the Battese, Harter and Fuller (BHF) model, is commonly used for the unit-level SAE approach. In the SAE literature, estimation methods include empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). The EBLUP method can be used under linear mixed models, while the EB and HB methods can be used under generalized linear mixed models. For a review of these methodologies and their extensions we refer to Rao and Molina (2015).

A second important issue we consider in this paper is the multidimensionality of well-being indicators. Although it is generally agreed that well-being is a multidimensional phenomenon (OECD-JRC, 2008), there is continuing debate about the suitability of combining social indicators based on taking their average or using a dashboard of single indicators. On the one hand, Ravallion (2011) argues that a single multidimensional composite indicator in the context of poverty measurement leads to a loss of
information, and on the other hand, Yalonetsky (2012) points out that composite indicators are necessary when the aim is to measure multiple deprivations within the same unit (individual or household). For a theoretical review of statistical properties of multidimensional indicators obtained by multivariate statistical techniques and related problems we refer to Krishnakumar and Nagar (2008) and Bartholomew et al. (2008).

Taking this latter view, an approach to reducing data dimensionality is to consider the multidimensional phenomena as a latent variable construct measurable by a set of observed variables, and estimated using a factor analysis (FA) model. Factor scores are estimated from a FA model and are defined as a composite variable computed from more than one response variable. Indeed, factor scores provide details on each unit’s placement on the factor. When we have a substantive framework where a set of variables explains a latent construct, the confirmatory FA modeling approach can be used. This is also true when a measurement framework is provided by official statistics or international organizations. In the context of well-being measurement, the vector of unobserved variables represents a set of variables that jointly describe the underlying phenomenon (Sosa-Escudero, Caruso, and Svarc 2013). Other papers on the use of factor analysis in latent well-being measurement to reduce data dimensionality are Ferro Luzzi, Fluckiger, and Weber (2008) and Gasparini et al. (2011).

Once factor scores are estimated, they can be used to conduct other statistical analysis. For instance, they can be used as part of regression or predictive analyses to answer particular research questions. Kawashima and Shiomi (2007) use factor scores in order
to conduct an ANOVA analysis on high school students’ attitudes towards critical thinking and tested differences by grade level and gender. In addition, Bell, McCallum, and Cox (2003) investigated reading and writing skills where they extracted the factors and estimated factor scores before using them in a multiple regression analysis model. Skrondal and Laake (2001) note that using factor scores as dependent variables in regression modelling produces consistent estimates of model parameters since any measurement error from the FA model is absorbed into the prediction error and coefficients are not attenuated (see also Fuller, 1987). Also, as highlighted in Kaplan (2009), we can assume that the specific variances from the FA model are very small compared to the prediction error.

In the current literature on SAE of social indicators, there is a research gap on the estimation of multidimensional indicators. In particular, the use of factor scores and factor analysis in SAE models is an open area of research. This research area is important when we have to deal with data dimensionality in the estimation of social indicators at a local level. In this paper, we consider economic well-being as a latent variable construct with the aim of reducing the dimensionality of well-being indicators. We then implement the unit-level SAE approach on the factor scores in both a simulation study and on real data from EU-SILC for the region of Tuscany, Italy.

This paper is organized as follows. In section 3.2, we describe the FA model for reducing data dimensionality on a dashboard of economic well-being indicators. In section 3.3 we review the unit-level SAE approach according to the BHF model, and
present the point estimation of the EBLUP for small area means. In section 3.4, we show results of a simulation study considering factor scores to deal with data dimensionality reduction and contrast them to the approach of averaging single univariate EBLUPs on the original variables where we consider both a simple average and a weighted average where the weights are defined by the factor loadings. A bootstrap algorithm to estimate mean squared errors (MSE) of the EBLUP of factor score means is also evaluated. In section 3.5, we discuss multidimensional economic well-being in Italy considering indicators from the Italian framework BES (Equitable and Sustainable Well-being) 2015 (ISTAT 2015). Also, using real data from EU-SILC 2009 for the area of Tuscany, we apply the proposed method and compute small area EBLUPs for factor score means and their mean squared error (MSE) for each Tuscany municipality (LAU 2). Finally, in section 3.6, we conclude the work with some final remarks and a general discussion.

3.2 Using Factor Scores for Data Dimensionality Reduction

In this section, we provide a general discussion on the use of FA models to reduce data dimensionality and focus on the estimation of factor scores. Since the focus of the application in Section 3.5 is on measuring economic well-being based on a substantive framework and a small number of single indicators, we consider here a one-factor FA model. We acknowledge that in the presence of more complex multidimensional phenomena, one factor may not explain the total variability. Moretti, Shlomo and Sakshaug (2017) investigate the issue of multiple latent factors under a multivariate SAE approach.
3.2.1 Issues in Composite Indicators

Multivariate statistical methods aim to reduce the dimensionality of a multivariate random variable $Y$. Formally, considering a $R^K$ space, where $K$ denotes the number of observed variables, we want to represent the observations in a reduced space $R^M$ with $M \ll K$. Bartholomew et al. (2008) suggests several multivariate statistical techniques in order to deal with data dimensionality reduction in the social sciences (e.g. principal component analysis, factor analysis models, multiple correspondence analysis, etc.). In this work, we consider the linear one-factor model, where the factor can be interpreted as a latent characteristic of the individuals revealed by the original variables. This model allows for making inference on the population, since the observable variables are linked to the unobservable factor by a probabilistic model to develop a composite indicator (Bartholomew et al., 2008: p. 175).

There is an ongoing debate about how to construct indicators which are useful for decision makers to inform policies. Saisana and Tarantola (2002) and Nardo et al. (2005) emphasize that composite indicators are important when a summary of multidimensional phenomena is needed. In particular, the use of multivariate statistical models, e.g. factor analysis models, can be used to construct such composite indicators. Nardo et al. (2005) highlight that factor analysis models reduce the data dimensionality of a set of sub-indicators whilst keeping the maximum proportion of the total variability of the observed data.

In this work we focus on data dimensionality reduction starting from a well-established
multidimensional well-being framework, e.g. the BES framework for Italy (ISTAT, 2015). This means that single indicators have already been grouped into well-being dimensions, one example being the economic well-being dimension. Therefore, factor analysis models are used under a confirmatory approach only.

Factor scores are estimated from a factor analysis model and they can be defined as composite estimates providing details on a unit’s placement on the latent factor (DiStefano, Zhu and Mindrila, 2009). The factor scores, once estimated, are easy to interpret: they have the same economic interpretation of the observed responses as they are strongly linearly related to these via a linear model.

There have been some first attempts in SAE and data dimensionality reduction using FA (e.g. Smith et al., 2015). Here, the construction of the composite indicators was on the small area EBLUPs of the single indicators. In this paper, we first construct the composite indicator and then obtain small area estimates of the average factor score. We also focus on mean squared error (MSE) estimation for the estimates.

### 3.2.2 The Linear Single-factor Analysis Model

Let us consider a $K \times 1$ vector of observed variables $Y$ and we assume that they are linearly dependent on a factor $f$. Thus, we can write the following linking model (Kaplan, 2009):
\[ Y = \Lambda f + \epsilon, \quad (1) \]

where \( \epsilon \) denotes a vector \( K \times 1 \) containing both measurement and specific error due to selection of variables in the model, and \( \Lambda \) is a \( K \times 1 \) vector of factor loadings.

Following Kaplan (2009: 42) we assume that specific variances are small relative to measurement error.

It is assumed that:

i) \( E(\epsilon) = 0, \)

ii) \( Var(\epsilon) = \Theta, \)

iii) \( \epsilon \sim N(0, \Theta), \)

iv) \( \epsilon \)’s components are uncorrelated,

v) \( E(f) = 0, \)

vi) \( Cov(\epsilon, f) = 0. \)

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

\[ \Sigma = Cov(YY') = \Lambda \Phi \Lambda' + \Theta, \quad (2) \]

where \( \Phi \) denotes the factor variance, and \( \Theta \) is a \( K \times K \) diagonal matrix of specific variance.
The Maximum-Likelihood (ML) approach is used to estimate the model parameters. ML equations under FA models do not have closed solution, so iterative numerical algorithms are proposed in the literature (see e.g. Mardia, Kent and Bibby 1979). The log-likelihood function $\ell$ of the data $Y$ can be written as follows (Härdle and Simar, 2012: p. 316):

$$\ell(Y; \Lambda, \Theta) = -\frac{nK}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{n - 1}{2} tr(SS^{-1}),$$  \hspace{1cm} (3)

Where $S$ denotes the sample covariance matrix.

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 Bartlett’s method, the individual factor scores estimate for $i = 1, \ldots, n$ are given by (Bartholomew, Deary, and Lawn, 2009):

$$\hat{f}_i = \hat{\Lambda}'\hat{\Theta}^{-1}y_i.$$  \hspace{1cm} (4)

Where $\hat{f} = \hat{\Lambda}'\hat{\Theta}^{-1}\hat{\Lambda}$ and $y_i$ denotes a $K$-dimensional vector of observations of $K$ components of $Y$ for $i = 1, \ldots, n$.

Bartlett’s method produces unbiased estimates of the true factor scores (Hershberger,
In the application presented in section 3.5, we also have binary dependent variables. According to Muthén and Muthén (2012) logistic regression is employed for binary dependent variables where the following transformation is applied in a single-factor model for each observed variable $k$:

$$
\text{logit} \left[ \pi_k(f) \right] = \log \frac{\pi_k(f)}{1 - \pi_k(f)} = \lambda_k f, k = 1, ..., K.
$$

(5)

where $\pi_k(f)$ denotes the probability that the dependent variable is equal to one, and $\frac{\pi_k(f)}{1 - \pi_k(f)}$ the odds. We can then write the following expression:

$$
\pi_k(f) = \frac{\exp(\lambda_k f)}{1 + \exp(\lambda_k f)}.
$$

(6)

which is monotonic in $f$ and with domain in the interval $[0,1]$. 

In the presence of binary and continuous observed variables and under a maximum likelihood estimation approach, the factor scores may be estimated via the expected posterior method described in Muthén (2012) and applied in Mplus, Version 7.4.

### 3.3 Small Area Estimation using Empirical Best Linear Unbiased Prediction (EBLUP)

A class of models for SAE is the mixed effects models where we include random area-
specific effects in the models and take into account the between-area variation.

3.3.1 Notation

Let $d = 1, \ldots, D$ denote small areas for which we want to compute estimates of the target population parameter for each $d$, in our case the population mean $\bar{F}_d$ of the factor score. For a sample $s \subset \Omega$ of size $n$ drawn from the target population 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 $s = \bigcup_d s_d$. $r_d$ denotes the non-sampled units for the small area $d$ of $N_d - n_d$ dimension.

3.3.2 Model based prediction using EBLUP

We consider the small area estimation problem for the mean under the EBLUP approach in the BHF model. Focusing on the population parameter of factor score means $\bar{F}_d$, $d = 1, \ldots, D$, and as the population mean is a linear quantity, we can write the following decomposition:

$$
\bar{F}_d = N_d^{-1} \left( \sum_{i \in s_d} f_{di} + \sum_{i \in r_d} f_{di} \right). 
$$

(7)

where $f_{di}$ is the population factor score for unit $i$ within small area $d$ assuming that the factor model is implemented on the whole population.
When auxiliary variables are available at the unit level, the BHF model can be used in order to predict the out-of-sample units. Considering the data for unit \( i \) in area \( d \) being \((f_{di}, x_{di}^T)\) where \( x_{di}^T \) denotes a vector of \( p \) auxiliary variables, the nested error regression model is the following:

\[
f_{di} = x_{di}^T \beta + u_d + e_{di}, i = 1, ..., N_D, d = 1, ..., D
\]

\[
u_d \sim \text{iidN}(0, \sigma_u^2), e_{di} \sim \text{iidN}(0, \sigma_e^2), \text{independent.} \tag{8}
\]

In this model there are two error components, \( u_d \) and \( e_{di} \), the random effect and the residual error term, respectively.

According to Royall (1970), we can write the best linear unbiased predictor (BLUP) for the mean as follows:

\[
\bar{F}_{d,\text{BLUP}} = N_d^{-1} \left( \sum_{i \in s_d} f_{di} + \sum_{i \in r_d} \tilde{f}_{di} \right). \tag{9}
\]

Where \( \tilde{f}_{di} = x_{di}^T \tilde{\beta} + \tilde{u}_d \) is the BLUP of \( f_{di} \), and \( \tilde{u}_d = \gamma_d (\tilde{f}_{ds} - \bar{x}_{ds}^T \tilde{\beta}) \) the BLUP of \( u_d \).

Here, \( \tilde{f}_{ds} = n_d^{-1} \sum_{i \in s_d} f_{di}, \bar{x}_{ds} = n_d^{-1} \sum_{i \in s_d} x_{di} \), and \( \gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / n_d} \in (0,1) \). \( \gamma_d \) is the \textit{shrinkage estimator} measuring the unexplained between-area variability on the total variability.
Since in practice the variance components $\sigma^2_e$ and $\sigma^2_u$ are unknown, we replace these quantities by estimates, so we calculate the EBLUP of the mean:

$$\hat{F}_d^{EBLUP} = N_d^{-1} \left( \sum_{i \in S_d} f_{di} + \sum_{i \in R_d} \hat{f}_{di} \right).$$  \hspace{1cm} (10)$$

Where $\hat{f}_{di} = x_{di}^T \hat{\beta} + \hat{u}_d$ is the EBLUP of $f_{di}$. For details on $\hat{\beta}$ and $\hat{u}_d$ we refer to Rao and Molina (2015: Ch. 7). As showed in Rao and Molina (2015: Ch. 7), $\hat{F}_d^{EBLUP}$ can be also written as follows:

$$\hat{F}_d^{EBLUP} = \frac{n_d}{N_d} f_{ds} + \left( \bar{X}_d - \frac{n_d}{N_d} \bar{x}_{ds} \right)^T \hat{\beta} + \left( 1 - \frac{n_d}{N_d} \right) \hat{u}_d.$$ \hspace{1cm} (11)

$\bar{X}_d$ denotes the means of the auxiliary variable in the population for the $d^{th}$ area.

It is clear that if the sample size in a small area is zero, it holds that $\hat{F}_d^{EBLUP} = \bar{X}_d \hat{\beta} = \hat{F}_d^{Synthetic}$ where $\bar{X}_d$ denotes the means of the covariates in the population. We can see from (11) that linkage between the auxiliary data (e.g. Census) and sample data is not needed.

### 3.3.3 Mean Squared Error Estimation

The mean squared error (MSE) of (11) can be estimated via analytical approximations or resampling techniques. Prasad and Rao (1990) proposed an analytical approximation of
MSE and González-Manteiga et al. (2008a) proposed bootstrap techniques. Moreover, when large sample analytical approximations are available, the bootstrap might provide more accurate estimation alternatives to analytical approximations due to its second-order accuracy (González-Manteiga et al., 2008a). Here, we suggest the use of a bootstrap method to estimate the MSE of (11). The bootstrap method proposed by González-Manteiga et al. (2008a) has been adapted for the case of using factor score means as the dependent variable in the SAE models in order to take into account the variability arising from the FA models. The steps are provided in appendix A and we evaluate our proposed algorithm via an extension to the simulation in Section 3.4.4. Analytical approximations of the MSE estimation of (11) under factor analysis models are a subject for future work.

3.4 Simulation Study

The simulation study was designed to assess the behavior of the EBLUP estimation of factor score means under a FA model. We compare this approach with a weighted average of a dashboard of standardized univariate EBLUPs calculated from the original variables. We use a simple average and a weighted average where the weights are obtained by the factor loadings. We also assess the bootstrap MSE estimation for the EBLUP of factor score means which will be used in the Application in Section 3.5. The simulation is based on generating one population from a multivariate mixed-effect model where the parameters are taken from the real dataset in the application in Section 3.5. We then draw 500 simple random samples without replacement (SRSWOR) in
order to obtain unplanned domains within our small areas. This simulation belongs to the group of design-based under model data simulations (Münnich, 2014). One outcome of the multivariate mixed-effect model (Fuller and Harter, 1987) is used as fixed universe and the rest of the study is a design-based simulation. Although EU-SILC may have complex survey designs, one important feature in the Italian EU-SILC for Tuscany is that every household (and hence adult in the household) has an equal inclusion probability (EPSEM) design and hence the simulation results are in line with the real data application. It is common to find in the literature other examples of simulation studies where simple random sampling is used to obtain unplanned domains, for example, Giusti, et al. (2013) used this approach when investigating a range of estimators also based on the EU-SILC. The subject of complex survey designs in SAE is a topic of ongoing research.

3.4.1 Generating the population

A single population is generated from a multivariate mixed-effects model, the natural extension of the BHF model (Fuller and Harter, 1987) with $N = 20,000$, $D = 80$, and $130 \leq N_d \leq 420$. $N_d$ is generated from the discrete uniform distribution, $N_d \sim \mathcal{U}(a = 130, b = 420)$, with $\sum_{d=1}^{D} N_d = 20,000$ where the parameters are obtained from the Italian EU-SILC 2009 dataset used in the application in 3.5. Here the multivariate model that we use to generate the population for the original variables (observed variables $Y$) is:
\[ y_{di} = x_{di}^T \beta + u_d + e_{di}, \; i = 1, \ldots, N_D, d = 1, \ldots, D \]  
\[ u_d \sim \text{iidMVN}(0, \Sigma_u), \; e_{di} \sim \text{iidMVN}(0, \Sigma_e), \; \text{independent}. \]

where \( y_{di} \) denotes a \( 3 \times 1 \) vector of observed responses for unit \( i \) belonging to area \( d \).

Two uncorrelated covariates are generated from the Normal distribution:
\[ X_1 \sim N(9.93, 4.98^2), \; \quad X_2 \sim N(57.13, 17.07^2). \]

These parameters reflect two real variables in the Italian EU-SILC 2009 dataset: the years of education and age. We understand that by generating the covariates according to Normal distributions with those parameters we introduce negative values (years of education and age do not have negative values), however, this has been done for simplicity in the simulation. We selected \( K=3 \) response variables from the Italian EU-SILC 2009 data: the log of the income, squared meters of the house, and the number of rooms, and fit regression models using the covariates \( X_1 \) and \( X_2 \).

From these models, we estimate the beta coefficient matrix and standard errors to build the simulation population by the model in (12). The \( \beta (3 \times 3) \) matrix of coefficients is given by:

\[
\beta = \begin{bmatrix} 3.983 & 0.018 & 0.001 \\ 1.263 & 0.007 & 0.005 \\ 0.404 & 0.006 & 0.002 \end{bmatrix}
\]

The response vector was generated according to the following variance components, where the correlation was set at 0.5 as derived from the Italian EU-SILC 2009 data:
\[ \Sigma_e = \begin{bmatrix} 0.063 & 0.028 & 0.021 \\ 0.028 & 0.049 & 0.018 \\ 0.021 & 0.018 & 0.027 \end{bmatrix} \]

We control the intra-class correlation \( \rho \) defined as
\[ \rho_k = \frac{\sigma_{uy_k}^2}{\sigma_{uv_k}^2 + \sigma_{ey_k}^2}, \]
for the \( k \)th component of \( Y \) and obtain the variance-covariance matrices of the correlated random effects. We chose three levels of intra-class correlations: 0.1, 0.3 and 0.8, and obtain the following matrices:

\[ \Sigma_{0.1} = \begin{bmatrix} 0.00693 & 0.00306 & 0.00227 \\ 0.00306 & 0.00539 & 0.00200 \\ 0.00227 & 0.00200 & 0.00297 \end{bmatrix}, \]

\[ \Sigma_{0.3} = \begin{bmatrix} 0.02709 & 0.01195 & 0.00887 \\ 0.01195 & 0.02107 & 0.00782 \\ 0.00887 & 0.00782 & 0.01161 \end{bmatrix}, \]

\[ \Sigma_{0.8} = \begin{bmatrix} 0.25500 & 0.11112 & 0.08249 \\ 0.11112 & 0.19600 & 0.07275 \\ 0.08249 & 0.07275 & 0.10800 \end{bmatrix}. \]

In order to have *unplanned domains*, we select \( S = 1, \ldots, 500 \) simple random samples without replacement (SRSWOR) from the population, and therefore we incur zero sample size domains as well as small domains.

We estimate the FA model on the population to derive the population factor scores \( f_i, \)
\( i = 1, \ldots, N \) according to (4).

We note that although FA models have been developed for multilevel structures within
domains, it is not possible to use these models for *unplanned domains* given a random sample, due to small and zero sample size domains. Thus, the use of two-level factor analysis models in SAE is a subject for future work.

We first estimate an explanatory (unrestricted) factor analysis model (EFA) on the whole population, allowing for all possible factors. The EFA is estimated to check and identify the underlying relationships between observed variables (Norris and Lecavalier, 2009). The EFA results show that the first factor explains a large amount of the total variability. Table 1 shows the estimated eigenvalues under different scenarios and Figure 1 the scree plots. The eigenvalue represents the variance of factor $m$, and measures the variance in all the variables which is accounted for by that factor. With a large eigenvalue for the first factor, we then fit a one-factor confirmatory FA model (CFA) on the population and estimate the population-based factor scores. The CFA one-factor model provides good fit statistics: $RMSEA = 0$ and $CFI = 1, TLI = 1$ (Hu and Bentler, 1999).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\rho = 0.1$</th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.060</td>
<td>2.055</td>
<td>2.139</td>
</tr>
<tr>
<td>2</td>
<td>0.450</td>
<td>0.478</td>
<td>0.448</td>
</tr>
<tr>
<td>3</td>
<td>0.440</td>
<td>0.450</td>
<td>0.402</td>
</tr>
</tbody>
</table>

*Table 1* Eigenvalues from the EFA of the simulation population.
We now define the following ‘true’ values for each of the small areas $d$ from our simulated population for $i = 1, \ldots, N$, $d = 1, \ldots, D$, and $k = 1, \ldots, K$:

- the factor score means: $\overline{F}_d = N_d^{-1} \sum_i f_{di}$;

- simple average of the observed variable standardized means: $\overline{Y}_d^{S\text{Averages}} = \frac{\sum_{k=1}^{K} Y_{dk}^*}{K}$;

- weighted average of the observed variables standardized means using the FA loadings: $\overline{Y}_d^{W\text{Averages}} = \frac{\sum_{k=1}^{K} \hat{\lambda}_k Y_{dk}^*}{\sum_{k=1}^{K} \hat{\lambda}_k}$.

$Y_{dk}^*$ denotes the standardized (mean zero and unit variance) true mean of the $k=1,\ldots,3$ variable and $d=1,\ldots,D$. $\hat{\lambda}_k$ denotes the estimated loading related to the $k^{th}$ variable in the population.

We highlight again that under FA model assumptions, the factor scores are strongly...
linearly related to the observed variables and have the same economic interpretation as the observed variables.

### 3.4.2 Simulation steps

The simulation study consists of the following steps:

1. Draw $S = 500$ samples using simple random sampling without replacement;

2. Fit the one-factor confirmatory FA model and estimate the EBLUP of factor score means for each area $d$ in each sample. We also calculate Horvitz-Thompson (HT) (Horvitz and Thompson, 1952) direct estimates of the factor score means for those areas with a non-zero sample size. In addition, the EBLUP for each of the original variables is also estimated in order to construct a simple average of the standardized small area EBLUPs and a weighted average using the factor loadings;

3. As the true values are known from the simulation population, we are able to calculate the root mean squared error (RMSE) and the relative bias (RBIAS) for each area $d$ for the three types of estimates: EBLUPs of factor score means, and the simple and weighted average of EBLUPs. For example, for the EBLUPs of factor score means the RMSE is:

$$RMSE\left(\hat{F}_{dEBLUP}\right)_d = \sqrt{S^{-1} \sum_{s=1}^{S} \left(\hat{F}_{dEBLUP}^{s} - \bar{F}_d\right)^2},$$  \hspace{1cm} (13)

and the RBIAS is:
4. For the overall comparison across all areas, we rank the small areas according to the estimates averaged across the 500 samples and compare each to the ranking in the population. We also examine the average of the RMSE and RBIAS across all areas.

We estimate the EBLUP for each original variable separately on each of 500 samples, and then standardize them and construct weighted and simple averages. These are compared to the true values in the simulation population. The weighted mean after standardizing the EBLUP estimates estimated on each sample are given as follows:

\[
\hat{Y}_{d_{s\text{\_Averages}}}^{EBLUP\_W} = \frac{\sum_{k=1}^{K} \left( \hat{Y}_{d_{ks}^{EBLUP}} \hat{\lambda}_{ks} \right)}{\sum_{k=1}^{K} \hat{\lambda}_{ks}}, d = 1, ..., D, k = 1, ..., K,
\]

where \(k\) denotes the \(k^{th}\) variable and \(\hat{\lambda}_{ks}\) the factor loading estimated on the \(s^{th}\) sample for the \(k^{th}\) variable, and \(\hat{Y}_{d_{ks}^{EBLUP}} = (\hat{Y}_{d_{ks}}^{EBLUP} - M_{ks}^{EBLUP})/SD_{ks}^{EBLUP}\) denotes the standardized EBLUP of the mean denoted by \(\hat{Y}_{d_{k}}\) for the \(k^{th}\) variable in the \(s^{th}\) sample,

\[
M_{ks}^{EBLUP} = D^{-1} \sum_{d} \hat{Y}_{d_{ks}^{EBLUP}}, \quad \text{and} \quad SD_{ks}^{EBLUP} = \sqrt{D^{-1} \sum_{d} (\hat{Y}_{d_{ks}^{EBLUP}} - M_{ks}^{EBLUP})^2}.
\]

We refer to Njong and Ningaye (2008) for weighted averages using factor loadings in composite estimation.

In the following tables and figures we dropped the subscript \(d\) as we show the estimates averaged across the areas.
3.4.3 Results: factor scores versus weighted and simple averages of standardized EBLUPs

In this section we show the main results of the simulation study. Table 2 contains the average eigenvalues across 500 samples under the EFA model and can be compared to Table 1 obtained from the simulated population. We can see that we are able to obtain good estimates for the eigenvalues across the samples. In parentheses we show the ratios between the sample and population eigenvalues. Table 3 presents the intra-class correlation coefficients estimated from the SAE model (averaged across 500 samples) showing that we approximate the known intra-class correlation coefficients as defined in the simulation population.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( \rho = 0.1 )</th>
<th>( \rho = 0.3 )</th>
<th>( \rho = 0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.058 (0.999)</td>
<td>2.050 (0.998)</td>
<td>2.135 (0.998)</td>
</tr>
<tr>
<td>2</td>
<td>0.445 (0.989)</td>
<td>0.473 (0.990)</td>
<td>0.442 (0.987)</td>
</tr>
<tr>
<td>3</td>
<td>0.442 (1.005)</td>
<td>0.455 (1.011)</td>
<td>0.405 (1.007)</td>
</tr>
</tbody>
</table>

Table 2 Average eigenvalues across 500 samples from EFA model.

Parenthetical entries are ratios between the sample and population eigenvalues.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( \rho = 0.1 )</th>
<th>( \rho = 0.3 )</th>
<th>( \rho = 0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.108</td>
<td>0.325</td>
<td>0.795</td>
</tr>
</tbody>
</table>

Table 3 Average intra-class correlation \( \hat{\rho} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \) estimates across 500 samples.

For each of the three estimates in small area \( d \) averaged across the 500 samples, we
compare the ranking of the small area domain estimates with the true ranking based on true area means according to our population using a Spearman’s correlation coefficient. These are shown in Table 4. The EBLUPs of the factor score means show an improvement and higher correlation to the true means in the population compared to the averages of EBLUPs, especially for the case of $\rho = 0.1$.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\rho = 0.1$</th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\gamma}_{EBLUP:S,Averages}$</td>
<td>0.780</td>
<td>0.996</td>
<td>0.999</td>
</tr>
<tr>
<td>$\hat{\gamma}_{EBLUP:W,Averages}$</td>
<td>0.793</td>
<td>0.996</td>
<td>0.998</td>
</tr>
<tr>
<td>$\hat{\gamma}_{EBLUP}$</td>
<td>0.986</td>
<td>0.997</td>
<td>0.999</td>
</tr>
</tbody>
</table>

*Table 4* Spearman’s correlation estimates for the three approaches.

*Figure 2* RMSE for Direct estimates and EBLUP of factor score means for small areas with $n_d > 0$.

Figure 2 shows the individual RMSE of the small areas for those areas with non-zero
sample sizes. In line with the SAE literature the EBLUP approach produces estimates with lower variability than direct HT estimates. Table 5 shows the overall RMSE comparison defined in (13) across 500 samples for the EBLUPs of factor scores, and simple and weighted standardized EBLUPs. We do not show the overall relative bias across the samples and areas since the estimates are all unbiased.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\rho = 0.1$</th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y}<em>{EBLUP</em>{S\text{,Averages}}}$</td>
<td>1.432</td>
<td>0.336</td>
<td>0.119</td>
</tr>
<tr>
<td>$\hat{Y}<em>{EBLUP</em>{W\text{,Averages}}}$</td>
<td>0.793</td>
<td>0.334</td>
<td>0.118</td>
</tr>
<tr>
<td>$\hat{F}_{EBLUP}$</td>
<td>0.140</td>
<td>0.125</td>
<td>0.090</td>
</tr>
</tbody>
</table>

*Table 5 RMSE estimates: comparison across 500 samples for the three approaches.*

The overall RMSEs for the EBLUP factor score means are lower than in the case of the simple and weighted averages of the dashboard of single EBLUPs for all levels of intra-class correlations, even after taking into account the extra modeling step of estimating factor scores. Hence, applying the EBLUP method on factor score means provides more precise estimates whilst reducing the data dimensionality of multiple observed variables.

### 3.4.4 Bootstrap MSE Estimation

In the application, we will use the algorithm defined in Appendix A to estimate the MSE of the EBLUP of the factor score means which take into account the variability arising
from the FA model. We extend the simulation for the case of the intra-class correlation of 0.3 to assess the properties of our proposed MSE estimation.

We compare the bootstrap RMSE according to the algorithm in Appendix A with the empirical RMSE (ERMSE) obtained across the 500 samples calculated as

$$ERMSE\left( \hat{\mu}_{EBLUP} \right) = \sqrt{S^{-1} \sum_{s=1}^{S} (\hat{\mu}_{EBLUP}^{s} - \bar{\mu})^2}.$$  

This is the “true” MSE that we consider in the relative bias estimation of the bootstrap MSE estimator.

Figure 3 shows the ratio between the bootstrap RMSE averaged across the 500 samples and the empirical RMSE that was estimated under two approaches: treating the factor scores as fixed and accounting for the variability of the FA model in the steps of the bootstrap. It can be seen that the RMSE estimated via bootstrap without accounting for the factor model is underestimated with a relative bias of -34.6% across the small areas. However, the relative bias across the small areas when accounting for the variability in the FA model is negligible at 4.0%. If we apply the bootstrap procedure and treat the factor scores as fixed, this leads to an underestimation in the RMSE.
Figure 3 Ratios between bootstrap RMSE and empirical RMSE estimated via bootstrap taking into account the FA model variability (---) and bootstrap ignoring the FA model variability (—).

3.4.5 Final remarks on the simulation study

The use of factor scores provides better rankings to true values compared to weighted and simple averages of single variables, especially for the case of small intra-class correlations which are more typical in real settings. Furthermore, it can be seen that factor scores provide estimates with lower variability (in terms of RMSE) than weighted and simple averages of single variables for estimating multidimensional phenomena at the small area level. We also conclude that it is crucial to consider the variability arising from the FA model in the bootstrap MSE estimation; otherwise, the true MSE will be underestimated.
Based on these results, we will use the EBLUP of the factor score means approach to reduce the dimensionality of observed variables in a real application using the Italian 2009 EU-SILC data for the Tuscany region and the adapted bootstrap procedure for MSE calculations.

3.5 Economic Well-being in Tuscany: a Multidimensional Approach

The aim of this section is to demonstrate how we can provide estimates of an economic well-being indicator following the BES guidelines for Tuscany municipalities. In our application, we use data from the EU-SILC 2009 and the 2001 General Census of Population and Housing. We note that the EU-SILC 2009 data were collected several years after the census and this is a limitation of the study since we assume stationarity of growth between the periods. Obviously the economic and financial crisis occurring in 2008 violates this assumption and further studies are needed with more current covariates. Nevertheless, the application is useful to demonstrate how small area estimates can be calculated for a multidimensional indicator.

3.5.1 Data and variables

Income and economic resources can be seen as conditions by which an individual is able to have a sustainable standard of life. One of the dimensions in the Italian Equitable and Sustainable Well-being (BES) framework is dedicated to Economic Well-being (ISTAT 2015). It consists of ten single economic-related indicators (a dashboard of indicators).
In this work, we focus on a subset of these highly correlated variables:

- Severe material deprivation according to Eurostat;
- Equivalized disposable income;
- Housing ownership;
- Housing density.

Appendix B contains the variables nomenclature for the 2009 Tuscany EU-SILC dataset used in our study and descriptive statistics of these study variables which are explained in the next sections.

Material deprivation can be defined as the inability to afford some items considered to be desirable, or even necessary, to achieve an adequate standard of life. Indicators related to this are absolute measures useful to analyze and compare aspects of poverty in and across EU countries (Eurostat, 2012). According to Eurostat, material deprivation in the EU can be measured by the proportion of people whose living conditions are severely affected by a lack of basic resources. Technically, the severe material deprivation rate shows the proportion of people living in households that cannot afford at least four of the following nine items because of financial difficulty:

1. Mortgage or rent payments, utility bills, hire purchase installments or other loan payments;
2. One week holiday away from home;
3. A meal with meat, chicken, fish or vegetarian equivalent every second day;
4. Unexpected financial expenses;
5. A telephone (including mobile telephone);
6. A color TV;
7. A washing machine;
8. A car;
9. Heating to keep the home sufficiently warm.

It can be argued that some of these indicators (e.g. 5 and 6) are nowadays less relevant than in the past. Nevertheless, these indicators are still used to describe the difficulties that households face in achieving a standard of life considered to be sufficient by society. This index is described in Table B3 in Appendix B. Disposable household income is the sum of gross personal income components plus gross income components at the household level minus employer’s social insurance contributions, interest paid on mortgage, regular taxes on wealth, regular inter-household cash transfer paid and tax on income. In order to take into account differences in household size and composition, we consider disposable equivalized income $I_{DE}^D$ defined as follows:

$$I_{DE}^D = \frac{I_i^D}{n_i^E}, i = 1, ..., N,$$

(16)

where $i = 1, ..., n$ denotes households, $I_i^D$ is the disposable household income, $n_i^E$ is the equivalized household size calculated in the following way (Haagenars et al., 1994):

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

(17)

where $HM_{14+}$ and $HM_{13+}$ are the numbers of household members aged 14 and over and
13 or younger at the end of the income reference period, respectively. This so-called ‘OECD modified scaling’ procedure is crucial to taking into account the economy of scales in the household. Due to the skewness of the variable, we use the log transformation in the factor model and SAE. The histograms are in Figure B2 and descriptive statistics in Table B1 of Appendix B. Housing ownership is measured by a dichotomous variable (0,1) where 0 denotes that the property where the household lies is not owned. According to the 2009 Tuscany EU-SILC data, 73.96% of households own the property where they live. Overcrowding is one of the indicators that National Statistics Institutes include in their well-being measurement frameworks. A very simple indicator of housing density is given by the ratio between the number of rooms in the household (excluding kitchen, bathroom and rooms used for work purposes) and the household size:

\[ \bar{n}_i = \frac{R_i}{M_i} \]  

(18)

where \( i \) is the household, \( M_i \) denotes the number of people in the \( i^{th} \) household, and \( R_i \) denotes the number of rooms in the household. The histogram of this variable is in Figure B3 and descriptive statistics are in Table B2 of Appendix B.
EU-SILC is conducted yearly by ISTAT for Italy and coordinated by EUROSTAT at the EU level. The survey is designed to produce accurate estimates at the national and regional levels (NUTS-2).

Hence, for the Italian geography the survey is not representative of provinces, municipalities (NUTS-3 and LAU-2 levels, respectively), and lower geographical levels. The regional samples are based on a stratified two-stage sample design. 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 and SSUs are selected by systematic sampling in each selected PSU. The total number of households in the sample for Tuscany is 1,448.

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.

3.5.2 The construction of the factor scores

The one-factor analysis model described in section 3.2 is fitted, and according to the goodness-of-fit statistics estimated on the one-factor model solution, the Root Mean Squared Error of Approximation (RMSEA=0.047) and the Comparative Fit Index criteria (CFI=0.966), the model provides good fitness (Hu and Bentler 1999). This choice can be justified also substantively as our variables relate to economic well-being according to the BES framework, which is the phenomenon we want to measure.
The histogram and Q-Q plot of the factor scores are shown in Figure 4 as well as descriptive statistics in Table 6. We see evidence of a little bit of skewness in the factor scores likely due to discrete variables included in the FA model. One interesting thing to note based on Table B4 in Appendix B is that the estimated intra-class correlation (ICC) for the factor scores is 0.1987 which is considerably higher than the estimated ICC’s for the single study variables, thus as seen in the simulation study, we expect that the EBLUP of the factor scores will provide good rankings of the small areas compared to weighted and simple averages.

![Factor scores distribution graphs.](image)

**Figure 4** Factor scores distribution graphs.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max</th>
<th>S.d.</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.2630</td>
<td>-0.3712</td>
<td>0.1050</td>
<td>0.0034</td>
<td>0.4120</td>
<td>2.0940</td>
<td>0.6436</td>
<td>0.1987</td>
</tr>
</tbody>
</table>

**Table 6** Descriptive statistics of factor scores.
3.5.3 Small Area Estimates

In this application we treat municipalities as our small areas of interest. The municipalities within Tuscany are unplanned domains in EU-SILC and only 59 out of 287 were sampled. Sample sizes in municipalities range from 0 to 135 households.

First, we provide direct estimates for the small areas with \( n_d > 0 \). After this, we build a SAE model under the BHF approach where the response variable is the factor score interpreted as the latent economic well-being construct. The exploratory variables in the model relate to the head of the household and are those common to both the survey and Census data. In particular, after a preliminary analysis of the available data we chose gender, age, year of education, household size, size of the flat (in squared meters), and employment status as the explanatory variables.

The single EBLUPs of the dashboard indicators have been estimated to construct the simple and weighted averages, as was done in the simulation study. In the case of binary variables the following linear logistic mixed effects model was fitted (MacGibbon and Tomber 1989):

\[
\text{logit}(p_{di}) = \log\left(\frac{p_{di}}{1 - p_{di}}\right) = x_{di}^T \beta + u_d, \tag{19}
\]

where \( p_{di} \) is the probability that \( y_{di} = 1 \) and \( u_d \sim \text{iidN}(0, \sigma_u^2) \).
In Figure 5 we compare the relative root mean squared error (RRMSE) of the EBLUPs of factor score means with the direct estimates as we did in the simulation study in Figure 2. In this application, EU-SILC survey weights are used to provide the direct estimates, therefore the post-stratified estimator (see Rao and Molina, 2015) is used. Here, the estimates of the MSE for the predictions are obtained via the bootstrap with $B = 500$ bootstrap samples as described in Appendix A. We can see the gain in efficiency (in terms of reduction in the RRMSE) obtained by the EBLUP compared to the direct estimates and in particular the RRMSE’s are below 10% demonstrating reliable estimates.

Figure 5 RRMSE direct estimates (—) and EBLUPs (---) for small areas with $n_d > 0$ ordered by growing sample size.
To facilitate the interpretation and provide a comparison between the different economic well-being indicators obtained from the EBLUP factor score means and the simple and weighted averages of the dashboard of EBLUPs, we have normalized the EBLUPs using the ‘Min-Max’ method (OECD-JRC, 2008), with range [0,1]. For the factor score EBLUPs, the normalization (denoted with a ‘*’) is as follows:

\[
\hat{F}_{d}^{\text{EBLUP}*} = \frac{\hat{F}_{d}^{\text{EBLUP}} - \min(\hat{F}_{d}^{\text{EBLUP}})}{\max(\hat{F}_{d}^{\text{EBLUP}}) - \min(\hat{F}_{d}^{\text{EBLUP}})}, d = 1, \ldots, D,
\]

where \(\hat{F}_{d}^{\text{EBLUP}} = \text{col}_{1 \leq d \leq D}\hat{F}_{d}^{\text{EBLUP}}\); and similarly, for the simple and weighted averages of the dashboard of standardized EBLUPs.

Table 7 shows the percentiles for the latent economic well-being indicator based on the normalized EBLUP factor scores and the normalized averages of the dashboard of EBLUPs. Figure 6 and Figure 7 depict the maps of the quartiles of the EBLUPs under the different approaches for the Tuscany region.
<table>
<thead>
<tr>
<th>Percentile</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EBLUP</strong></td>
<td>0.0000</td>
<td>0.5110</td>
<td>0.5468</td>
<td>0.5819</td>
<td>1.0000</td>
</tr>
<tr>
<td><strong>Simple</strong></td>
<td>0.0000</td>
<td>0.4297</td>
<td>0.5297</td>
<td>0.6061</td>
<td>1.0000</td>
</tr>
<tr>
<td><strong>Weighted</strong></td>
<td>0.0000</td>
<td>0.4796</td>
<td>0.6006</td>
<td>0.7184</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 7 Percentiles for the transformed latent economic well-being indicator based on the EBLUP of factor score means and simple and weighted averages.

Figure 6 Latent economic well-being indicator based on transformed EBLUP of factor scores means (1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile).
Figure 7 Latent economic well-being indicator based on simple and weighted averages of single EBLUPs (1=1st quartile; 2=2nd quartile; 3=3rd quartile; 4=4th quartile).

In the maps of Figure 6 and Figure 7 a darker color denotes a better well-being phenomenon. Looking at these figures we can draw some interesting conclusions on economic well-being in the Tuscany region.

The municipalities located in the Massa-Carrara province, which is based in the North of Tuscany (i.e. Pontremoli and Zeri municipalities), and municipalities based in Grosseto province (south of Tuscany), are the poorest ones. The small areas based in the Florence province are wealthy municipalities, as well as the ones located in the center of the region (Siena province). The lowest point estimates of the latent economic well-being
indicator are estimated for Carrara and Seravezza municipalities, and the highest values for Firenze and Arezzo municipalities. Our results based on the EBLUPs of the factor scores in Figure 6 are more comparable with other SAE studies on welfare and poverty in Tuscany (Marchetti, Tzavidis, and Pratesi 2012; Giusti et al. 2015) compared to the averages of a dashboard of EBLUPs in Figure 7, though previous SAE studies consider only income variables rather than a composite indicator used here. This is not surprising given the low ICCs for each of the individual EBLUPs that form the dashboard which may result in more distortions on the rankings, particularly since some of the individual EBLUPs are based on discrete variables. This was highlighted in the simulation study as well.

3.5.4 Model diagnostics

We assess the fit of the model by analyzing the level-1 and level-2 standardized residuals. In particular, the Q-Q plots of the residuals, shown in Figure 8 and Figure 9 show the leverage measures versus standardized scaled residuals from the linear model. Both figures show a presence of outliers in the right tail, although the factor scores distribution is approximately symmetric. Figure 9 also shows the contour of the Cook’s distance which does not deviate much from zero and hence we can conclude that the outliers are not influential.
Figure 8 Q-Q plots for the level-1 and level-2 residuals of the BHF model fitting.

Figure 9 Standardized residuals versus leverage measure.

3.6 Conclusion and Discussion

In this paper we evaluated a method to estimate the mean of a latent economic well-being indicator at the local level for Tuscany using factor scores to reduce data dimensionality. We focused on the factor scores because they can be seen as a latent
economic well-being composite variable. The simulation study demonstrated that factor score means provide a better ranking of the small areas compared to the true population means as measured by the Spearman’s correlation coefficient, especially when intra-class correlations are small, which is typical in real settings. The simple and weighted averages of univariate standardized EBLUPs also provided good rankings for the higher intra-class correlations that were examined. In addition, the use of factor scores provided more precise estimates in terms of the MSE for an estimate of a multidimensional phenomenon compared to the averages of the EBLUPs. The use of factor analysis models and factor scores has important advantages and implications in data dimensionality reduction: it avoids arbitrary weighting of single indicators and it generates continuous composite scores, which can be modeled using model-based SAE methods. Since the factor scores are strongly linearly related to the multidimensional observed variables, this leads to easier interpretation.

Another important point studied in this paper, is the MSE estimation of EBLUPs of factor score means. In this work, we proposed a modification to the González-Manteiga et al. (2008a) bootstrap algorithm to account for the additional variability added to the small area estimates by using factor scores obtained from an FA model as the dependent variable. This has been tested via simulation and we showed that if the variability arising from the FA model is ignored, the MSE is underestimated and therefore biased. For more theoretical details on the bootstrap, we refer to González-Manteiga et al. (2008a). Analytical MSE approximations are left for future work.
There are several areas where this work could be extended. Future work might consider other geographical levels, such as SLL (Sistemi Locali del Lavoro – Labor Local System), by looking at the flow of daily travel home/work (commuting) detected during the General Census of Population and Housing. Further interesting applications would involve comparisons between Italian regions in the North, Central, and South.

Another worthwhile extension is accounting for more than one factor. When the goal is to reduce the dimensionality of the original data by identifying latent factors, one might face the issue of identifying multiple factors. Multiple latent factors can arise, particularly when we have many indicators referring to the same phenomenon which can be grouped substantively into subdomains. For example, if the goal is to study housing quality we may want to consider the following dimensions: type of dwelling and tenure status, housing affordability, and housing quality (e.g. overcrowding, housing deprivation, problems in the residential area). For multiple latent factors, we may have factor scores that are correlated and hence future research should explore the use of the multivariate mixed effects model (Fuller and Harter, 1987). Datta, Day, and Basawa (1999) showed that the use of the multivariate mixed effects model might lead to gains in efficiency in terms of MSE for the EBLUP compared to the BHF model. Therefore, the multivariate small area estimation method might provide better dashboard estimates and averages if the correlation between the single variables is taken into account. These extensions are currently being carried out in Moretti, Shlomo and Sakshaug (2017).
Acknowledgements

This analysis was carried out on confidential data released by ISTAT. Data were analyzed by respecting all of the Italian confidential restriction regulations (D.Lgs. 196/03 – Codice Privacy). 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. Also, we would like to thank Caterina Santi, researcher at Scuola Superiore Sant’Anna – Institute of Economics in Pisa for her kind suggestions on the software. This work was financially supported by the following grants: ESRC DTC Award and Advanced Quantitative Methods (also known as AQM).
Appendix A: Parametric bootstrap procedure for the EBLUPs of factor scores MSEs.

Here we show the bootstrap steps for the EBLUP’s MSE. The bootstrap procedure is the one proposed by González-Manteiga et al. (2008a) and we particularize the algorithm by taking into account the FA model variability (in step 1).

1. Draw $b = 1, \ldots, B$ simple random samples with replacement from the observed sample $S$ and estimate FA models to obtain factor scores. After this, the usual parametric bootstrap proposed by González-Manteiga et al. (2008a) is run for the $b = 1 \ldots, B$ bootstraps.

2. Fit the Battese, Harter and Fuller model to the sampled units $f_b = (f_{1b}', \ldots, f_{Db}')'$, and estimate the model parameters $\hat{\beta}, \hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$.

3. Generate $u_d^{*(b)} \sim \text{iidN}(0, \hat{\sigma}_u^2), d = 1, \ldots, D$, which are the bootstrap area effects.

4. Generate the bootstrap errors for the sample units $e_{dl}^{*(b)} \sim \text{iidN}(0, \hat{\sigma}_e^2)$, independently of the $u_d^{*(b)}$ and the error domain means $E_d^{*(b)} \sim \text{iidN}\left(0, \frac{\hat{\sigma}_e^2}{N_d}\right), d = 1, \ldots, D$.

5. Calculate the true means for each small area of the bootstrap population as follows:

$$F_d^{*(b)} = \bar{X}_d' \hat{\beta} + u_d^{*(b)} + E_d^{*(b)}, d = 1, \ldots, D,$$

where $\bar{X}_d'$ denotes the means of the from the population (auxiliary variables).

6. Generate the responses for the sample units by using the sample covariates vectors $x_{dl}, i \in s_d$:

$$F_{dl}^{*(b)} = x_{dl}' \hat{\beta} + u_d^{*(b)} + e_{dl}^{*(b)}, d = 1, \ldots, D.$$
7. Fit the nested errors model to the bootstrap sample data \( F_{d_i}^{*(b)} \) and obtain the bootstrap EBLUPs \( \hat{F}_d^{*(b)}, d = 1, \ldots, D \).

8. Replicate steps from 1 to 7 for \( b = 1, \ldots, B \). The Monte Carlo approximation of the bootstrap estimator of the EBLUP is given by:

\[
\text{mse}\left( \hat{F}_d^{EBLUP} \right) = \frac{1}{B} \sum_{b=1}^{B} \left( \hat{F}_d^{*(b)} - \bar{F}_d^{*(b)} \right)^2, \quad d = 1, \ldots, D.
\]

\( \bar{F}_d^{*(b)} \) denotes the true mean and \( \hat{F}_d^{*(b)} \) the EBLUP for the area \( d \) for replicate \( b \).

We run the bootstrap procedure with \( B=500 \) both in the simulation and application.
## Appendix B: EU-SILC data study variables

Here we describe the Italian EU-SILC 2009 data nomenclature and show some descriptive statistics on the study variables.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOM</td>
<td>Area code: comune (municipality)</td>
</tr>
<tr>
<td>HOUSEHOLD CROSS-SECTIONAL WEIGHT</td>
<td>Cross-sectional survey weight</td>
</tr>
<tr>
<td>TOTAL DISPOSABLE HOUSEHOLD INCOME</td>
<td>Total disposable household income</td>
</tr>
<tr>
<td>STANZE</td>
<td>Rooms in the flat (except: kitchen, toilet and bathroom, hallway, corridor, rooms used for work purposes).</td>
</tr>
<tr>
<td>GODAB_B</td>
<td>House ownership variable indicator</td>
</tr>
<tr>
<td><strong>Material deprivation variables</strong></td>
<td></td>
</tr>
<tr>
<td>IMPREV</td>
<td>Ability to deal with unexpected expenses of €1000</td>
</tr>
<tr>
<td>FERIE</td>
<td>Affordability of one week per year away from home</td>
</tr>
<tr>
<td>PASTO</td>
<td>Affordability of a meat or chicken, or fish (or equivalent vegetarian) every two days</td>
</tr>
<tr>
<td>RISADE</td>
<td>Capacity of heating the house properly</td>
</tr>
<tr>
<td>LAVATR</td>
<td>Washing machine ownership</td>
</tr>
<tr>
<td>TV</td>
<td>TV ownership</td>
</tr>
<tr>
<td>AUTO</td>
<td>Car ownership</td>
</tr>
<tr>
<td>CELL</td>
<td>Telephone ownership</td>
</tr>
<tr>
<td>PAGAFF</td>
<td>Difficulties in paying the rent</td>
</tr>
<tr>
<td>PAGBOL</td>
<td>Difficulties in paying bills</td>
</tr>
<tr>
<td>PAGALDEB</td>
<td>Difficulties in paying loans or something similar</td>
</tr>
<tr>
<td>PAGMUT</td>
<td>Difficulties in paying the mortgage</td>
</tr>
</tbody>
</table>

*Figure B1 Italian EU-SILC variables nomenclature.*
**Figure B2** Disposable equivalized income histogram.

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max</th>
<th>S.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-24,670</td>
<td>12,200</td>
<td>17,410</td>
<td>20,090</td>
<td>23,740</td>
<td>190,800</td>
<td>13,990.88</td>
</tr>
<tr>
<td>2.398</td>
<td>4.087</td>
<td>4.243</td>
<td>4.231</td>
<td>4.377</td>
<td>5.280</td>
<td>0.264</td>
</tr>
</tbody>
</table>

*Table B1 Equivalized disposable income and log equivalized disposable income descriptive statistics.*
Figure B3 Housing density.

Table B2 Descriptive statistics of the housing density.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max</th>
<th>S.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.250</td>
<td>1.000</td>
<td>1.600</td>
<td>1.989</td>
<td>2.500</td>
<td>8.000</td>
<td>1.239</td>
</tr>
</tbody>
</table>

Table B3 Frequencies of the binary variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material deprivation</td>
<td>3.94%</td>
</tr>
<tr>
<td>House ownership</td>
<td>73.96%</td>
</tr>
<tr>
<td>Variable</td>
<td>Estimated ICC</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Factor scores</td>
<td>0.1987</td>
</tr>
<tr>
<td>Disposable equivalized income</td>
<td>0.0019</td>
</tr>
<tr>
<td>Room average</td>
<td>0.0680</td>
</tr>
<tr>
<td>Material deprivation</td>
<td>0.0189</td>
</tr>
<tr>
<td>House ownership</td>
<td>0.0410</td>
</tr>
</tbody>
</table>

*Table B4 Estimation of the ICCs of the study variables and factor scores.*

<table>
<thead>
<tr>
<th>Factor</th>
<th>Eigenvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.791</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.727</td>
</tr>
<tr>
<td>4</td>
<td>0.566</td>
</tr>
</tbody>
</table>

*Table B5 Eigenvalues from exploratory FA model on Tuscany EU-SILC 2009.*
Appendix C: Specification of the main R functions and challenges in the computations

Here we describe the main R packages we used for the small area estimates. All the other analyses were programmed manually.

C.1 Estimation of small area means and MSE under EBLUP approach with the “sae” package (Molina and Marhuenda 2015)

Required packages: nlme, MASS

Functions: eblupBHF( ) and pbmseBHF( ).

Functions: fitsaemodel() and robpredict().


Functions: mplusObject( ), mplusModeler( ).

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

Functions: readShapePoly( ), spplot( )

As mentioned in the simulation sections, two-level factor analysis models may be helpful in multilevel data. However, due to our simulation setting (unplanned domains) the estimation algorithms may break in case of areas with $n_d < 55$ units. Future work will take into account for these problems in SAE. No other challenges were encountered in the computations. Also, algorithms for mixed-effect model may break in case of very small intra-class correlation, where these kinds of models are not appropriate.
4 Multivariate Small Area Estimation of Multidimensional Latent Economic Well-being Indicators

Introduction to the paper

This chapter is a paper submitted to the *International Statistical Review* and is currently under review. I am the lead author of the paper and responsible for the writing of the article and carrying out all of the analysis and simulation studies. All ideas and approaches are discussed through the normal supervision process. This paper was also submitted and accepted for an oral presentation at the *Joint Statistical Meetings* 2017 (Baltimore, U.S.).

In this paper we study the problem of multivariate SAE in data dimensionality reduction following the methods we evaluated in Chapter 3. This chapter is linked to the previous chapter in the sense that we have extended the factor analysis model as well as the small area estimation model to the case where multiple well-being dimensions arise. The case of two latent variables is considered. A simulation study and a real data application on Italian EU-SILC 2009 and Census 2001 are proposed in order to evaluate the method.
Multivariate Small Area Estimation of Multidimensional Latent Economic Well-being Indicators

Angelo Moretti, Natalie Shlomo and Joseph Sakshaug

Social Statistics, School of Social Sciences, University of Manchester, United Kingdom

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 well-being. We employ FA models and use multivariate EBLUP (MEBLUP) to predict a vector of means of factor scores representing well-being 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.

4.1 Introduction

The international scientific community, national statistical agencies, and international organisations have pointed out the multidimensional nature of well-being 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 well-being measurement frameworks. One example is the Italian Statistical Institute (ISTAT) and National Council for Economics and Labour (CNELE) “Equitable and Sustainable Well-being (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 well-being. 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 well-being) within the same unit (individual or household). In order to measure multidimensional well-being, 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 well-being measurement, there is a need to obtain reliable and accurate indicators at a local area level since well-being 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 well-being (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 well-being 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. (2017a) evaluate the use of Factor Analysis (FA) models in SAE in order to reduce data dimensionality for economic well-being indicators and show that they can provide good estimates of multidimensional well-being 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., 2017a). As mentioned, Moretti et al. (2017a) 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 4.2 we introduce the multivariate SAE approach for a mean vector and review the multivariate EBLUP (MEBLUP) under the mixed effects model. In section 4.3 we discuss the data dimensionality reduction problem via a Factor Analysis (FA) model. In section 4.4 we present a simulation study to evaluate our approach and address the comparisons (a) to (c) above. In section 4.5 we consider the multidimensionality issue of housing deprivation in Italy through an application using Italian EU-SILC data. We conclude our work in section 4.6 with a final discussion on the main findings and future work.
4.2 Multivariate Empirical Best Linear Unbiased Predictor (MEBLUP)

Let \( d = 1, \ldots, D \) 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 \( \Omega \) 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 \( y_{di} = (y_{d1i}, \ldots, y_{dKi})' \), which denotes the \( K \times 1 \) vector of interest for \( i = 1, \ldots, N_d, d = 1, \ldots, D \), we can write the target means vector as follows:

\[
\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}.
\]  

(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:

\[
\bar{Y}_d = N_d^{-1} \left( \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} y_{di} \right).
\]  

(2)

The quantity \( \sum_{i \in r_d} y_{di} \) 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.
4.2.1 Multivariate nested-error linear regression model

We assume that unit-specific auxiliary variables $x_{id}$ 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:

$$y_{di} = x_{di} \beta + u_d + e_{di}, \quad d = 1, \ldots, D, \ i = 1, \ldots, N_d,$$

where $y_{di}$ denotes the $K \times 1$ responses for the $i^{th}$ unit from the $d^{th}$ small area, $x_{di}$ is the $K \times p$ matrix of the auxiliary variables, $\beta$ is a $K \times p$ matrix of unknown regression coefficients, $u_d$ is the $K \times 1$ vector of the area effects, and $e_{di}$ is the $K \times 1$ vector of the individual effects. Here, the $K \times K$ positive-definite matrices $\Sigma_u, \Sigma_e$ are the covariance matrices of the area effects and individual effects respectively.

Under model (3) we can write the observed mean in the population of area $d$ as:

$$\bar{Y}_d = \bar{X}_d \beta + u_d$$

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

4.2.2 Estimation and prediction of unknown parameters

For simplicity we now make use of the following notation,
\[ y_d = \text{col}_{l \in s_d}(y_{di}), \quad Y = \text{col}_{1 \leq d \leq D}(y_d). \]

\[ x_d = \text{col}_{l \in s_d}(x_{di}), \quad X = \text{col}_{1 \leq d \leq D}(x_d) \]

Let us now denote the covariance matrix and expectation of \( Y \) by:

\[ \mathbf{V}(Y) = \text{block diag}(\mathbf{V}_{11}, \ldots, \mathbf{V}_{DD}), E(Y) = X\beta \] (5)

where \( \mathbf{V}_{dd} = (J_{nd} \otimes \Sigma_u) + (I_{n_d} \otimes \Sigma_e). J_{nd} \) is the \( n_d \times n_d \) matrix with every element equal to one and \( I_{n_d} \) is an identity matrix. The operator \( \otimes \) denotes the Kronecker product. The empirical best linear unbiased estimator of the matrix of regression coefficients is given by:

\[ \hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y. \] (6)

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

\[ \hat{u}_d = (\bar{Y}_d - \bar{x}_d\hat{\beta})[\Sigma_u + n_d^{-1}\Sigma_e]^{-1}\Sigma_u, d = 1, \ldots, D \] (7)

\( \Sigma_u \) and \( \Sigma_e \) are estimators of \( \Sigma_u \) and \( \Sigma_e \) (we refer to Schafer and Yucel (2002) for the algorithm and its implementation).

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

112
\[ \hat{\mathbf{Y}}_{d}^{MEBLUP} = \mathbf{X}'_{d} \hat{\mathbf{B}} + \hat{\mathbf{u}}_{d} = \mathbf{X}'_{d} \hat{\mathbf{B}} + (\mathbf{Y}_{d} - \overline{\mathbf{X}}_{d} \hat{\mathbf{B}})(\mathbf{\Sigma}_{u} + n_{d}^{-1}\mathbf{\Sigma}_{e})^{-1}\mathbf{\Sigma}_{u}, \quad d = 1, \ldots, D. \] (8)

where \( \overline{\mathbf{X}}_{d} \) denotes the sample auxiliary means for area \( d \). In case of areas with \( n_{d} = 0 \) it holds that \( \hat{\mathbf{Y}}_{d}^{MEBLUP} = \hat{\mathbf{Y}}_{d}^{Synthetic} = \overline{\mathbf{X}}_{d} \hat{\mathbf{B}}. \)

The mean squared error of (8) can be estimated via parametric bootstrap proposed by Moretti et al. (2017b). The mean squared error of \( \hat{\mathbf{Y}}_{d}^{Synthetic} \) 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: Ch. 7).

4.3 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-JRC, 2004). A vast literature on multivariate statistical analysis techniques is available; for a formal review on the main methods we refer to Härdle and Simar (2012). In this paper we assume that latent constructs exist for a well-being domain and use FA methods to reduce the data dimensionality from the original variables.
4.3.1 The factor analysis model

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

$$Y = \Lambda f + \epsilon$$

(9)

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

It is assumed that:

i) $E(\epsilon) = 0$,

ii) $E(f) = 0$,

iii) $Cov(\epsilon, f) = 0$.

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

$$\Sigma = Cov(YY') = \Lambda E(ff')\Lambda' + E(\epsilon \epsilon') = \Lambda \Phi \Lambda' + \Theta,$$

(10)

where $\Phi$ is a $M \times M$ matrix of factor variances and covariances, and $\Theta$ 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 do not have closed solution, so iterative numerical
algorithms are proposed in the literature (see e.g. Mardia et al., 1979). The log-likelihood function $\ell$ of the data $Y$ can be written as follows (Härdle and Simar, 2012: p. 316):

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

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

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 $i = 1, \ldots, n$ are given by (Härdle and Simar, 2012: p. 323) where $\tilde{\Lambda}$ denotes the estimator of $\Lambda$:

$$\tilde{f}_i = \tilde{\Lambda}'\tilde{\Sigma}^{-1}y_i.$$  

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 (2012) and applied in Mplus7.4 (Muthén and Muthén, 2012).

### 4.4 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.

This simulation belongs to the group of design-based under model data simulations
(Münnich, 2014). One outcome of the multivariate mixed-effect model (Fuller and
Harter, 1987) is used as fixed universe.

The overall results of the simulation study are evaluated via the empirical root mean
squared error (RMSE) described in Section 4.4.3.

4.4.1 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
\( \sum_{d=1}^{D} N_d = 20,000 \). \( y_{di} \) observations are generated according to the multivariate mixed
effects model shown in formula (3). The simulation parameters \( \Sigma_e \) and \( \beta \) 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 \( \Sigma_e \):

\[
\Sigma_e = \begin{bmatrix}
0.386 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\
\sigma_{21} & 0.414 & \sigma_{23} & \sigma_{24} \\
\sigma_{31} & \sigma_{32} & 0.213 & \sigma_{34} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & 0.301
\end{bmatrix}
\]

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

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 (9): 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 10 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 $f_i, i = 1, ..., N$ from (12).

![Diagram showing relationship between factors and observed variables](image)

*Figure 10 Relationship between the factors and observed variables two-factor CFA.*

As mentioned in Moretti et al. (2017a), 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 \( k = 1, \ldots, K \) that contribute to the \( m \)th factor is given by:

\[
\hat{\theta}_{EBLUP, W_Averages}^{Y_{dm}} = \frac{\sum_{k=1}^{K} (\hat{\tau}_{dk}^{standard, EBLUP} \hat{\lambda}_{km})}{\sum_{k=1}^{K} \hat{\lambda}_{km}}, d = 1, \ldots, D, m = 1, \ldots, M
\]  

(13)

where \( \hat{\lambda}_{k} \) is the estimated factor loading for variable \( k \) related to factor \( m \) (refer to Njong and Ningaye (2008) for weighted averages using factor loadings).

### 4.4.2 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:
\[ RMSE\left(\hat{\phi}_{dk}^{EBLUP}\right) = \sqrt{S-1} \sum_{s=1}^{S} \left(\hat{\phi}_{dks}^{EBLUP} - \bar{Y}_{dk}^{TRUE}\right)^2 \]

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

4.4.3 Results

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

Table 8 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:

\[ \Delta_{dk} = \frac{RMSE(\hat{\phi}_{MEBLUP}^{dk}) - RMSE(\hat{\phi}_{EBLUP}^{dk})}{RMSE(\hat{\phi}_{EBLUP}^{dk})} \cdot 100, \quad k = 1, ..., K, \ d=1,...,D. \]

\( \Delta_{dk} \) estimates are then averaged across the areas to provide summary statistics for each variable \( k \):

\[ \Delta_k = \frac{1}{D} \sum_d \Delta_{dk}. \]
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 $\Sigma_e$ 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.

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

Table 9 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.
Table 9 $\hat{r}_e$, $\hat{r}_u$, and $\hat{ICC}$ of factor scores under multivariate MEBLUP averaged across samples.

Table 10 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 11.
<table>
<thead>
<tr>
<th>( \text{Scenario} )</th>
<th>( I C C_k )</th>
<th>Factor scores 1</th>
<th>Factor scores 2</th>
<th>( \Delta k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_e = 0.7, r_u = 0.2 )</td>
<td>0.05</td>
<td>-2.44</td>
<td>-2.50</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>-2.56</td>
<td>-3.13</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>-4.48</td>
<td>-5.56</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*Table 10 Percentage relative reduction (%) in terms of RMSE of MEBLUP over EBLUP (\( \Delta k \)), two-factor CFA model.*

**4.4.3.3 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 11 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 12 shows the percentage relative reduction in RMSE for the simple and weighted averages of standardised MEBLUPs over EBLUPs shown in Table 11.
**Scenario**

<table>
<thead>
<tr>
<th>ICC&lt;sub&gt;k&lt;/sub&gt;</th>
<th>( r_e = 0.7, r_u = 0.2 )</th>
<th>( r_e = 0.7, r_u = -0.2 )</th>
<th>( r_e = 0.2, r_u = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EBLUP</td>
<td>MEBLUP</td>
<td>EBLUP</td>
</tr>
<tr>
<td><strong>0.05</strong> Factor scores</td>
<td>0.081</td>
<td>-</td>
<td>0.080</td>
</tr>
<tr>
<td>simple</td>
<td>0.267</td>
<td>0.244</td>
<td>0.231</td>
</tr>
<tr>
<td>weighted</td>
<td>0.230</td>
<td>0.220</td>
<td>0.207</td>
</tr>
<tr>
<td><strong>0.1</strong> Factor scores</td>
<td>0.070</td>
<td>-</td>
<td>0.061</td>
</tr>
<tr>
<td>simple</td>
<td>0.246</td>
<td>0.225</td>
<td>0.250</td>
</tr>
<tr>
<td>weighted</td>
<td>0.180</td>
<td>0.190</td>
<td>0.224</td>
</tr>
<tr>
<td><strong>0.3</strong> Factor scores</td>
<td>0.065</td>
<td>-</td>
<td>0.039</td>
</tr>
<tr>
<td>simple</td>
<td>0.200</td>
<td>0.177</td>
<td>0.181</td>
</tr>
<tr>
<td>weighted</td>
<td>0.175</td>
<td>0.157</td>
<td>0.163</td>
</tr>
</tbody>
</table>

Table 11 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).
<table>
<thead>
<tr>
<th>Scenario</th>
<th>$ICC_k$</th>
<th>$r_e = 0.7, r_u = 0.2$</th>
<th>$r_e = 0.7, r_u = -0.2$</th>
<th>$r_e = 0.2, r_u = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>simple</td>
<td>-8.61</td>
<td>-21.65</td>
<td>-0.87</td>
</tr>
<tr>
<td></td>
<td>weighted</td>
<td>-4.35</td>
<td>-20.77</td>
<td>-0.54</td>
</tr>
<tr>
<td>0.1</td>
<td>simple</td>
<td>-8.54</td>
<td>-28.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>weighted</td>
<td>-5.56</td>
<td>-27.68</td>
<td>0.00</td>
</tr>
<tr>
<td>0.3</td>
<td>simple</td>
<td>-11.50</td>
<td>-11.60</td>
<td>-0.51</td>
</tr>
<tr>
<td></td>
<td>weighted</td>
<td>-10.29</td>
<td>-11.66</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 12 Percentage relative reduction (%) in terms of RMSE of simple and weighted averages of standardised MEBLUP over EBLUP ($\Delta_k$).

Looking at Table 11, 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. (2017a), 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 12. 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 13 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 14 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 13. 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 10 and discussed in Section 4.4.3.2.
<table>
<thead>
<tr>
<th>Scenario</th>
<th>(<em>IC_k</em> = 0.05, r_e = 0.7, r_u = 0.2)</th>
<th>(<em>IC_k</em> = 0.7, r_u = -0.2)</th>
<th>(<em>IC_k</em> = 0.2, r_u = 0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>EBLUP</strong></td>
<td><strong>MEBLUP</strong></td>
<td><strong>EBLUP</strong></td>
</tr>
<tr>
<td><strong>0.05</strong></td>
<td>(0.082)</td>
<td>(0.080)</td>
<td>(0.080)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.082)</td>
<td>(0.080)</td>
<td>(0.080)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.380)</td>
<td>(0.360)</td>
<td>(0.360)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.378)</td>
<td>(0.358)</td>
<td>(0.353)</td>
</tr>
<tr>
<td><strong>0.1</strong></td>
<td>(0.078)</td>
<td>(0.076)</td>
<td>(0.064)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.078)</td>
<td>(0.076)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.450)</td>
<td>(0.410)</td>
<td>(0.450)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.430)</td>
<td>(0.390)</td>
<td>(0.440)</td>
</tr>
<tr>
<td><strong>0.3</strong></td>
<td>(0.067)</td>
<td>(0.064)</td>
<td>(0.036)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.067)</td>
<td>(0.064)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.600)</td>
<td>(0.530)</td>
<td>(0.600)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.589)</td>
<td>(0.519)</td>
<td>(0.530)</td>
</tr>
<tr>
<td><strong>0.05</strong></td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.039)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.487)</td>
<td>(0.468)</td>
<td>(0.443)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.485)</td>
<td>(0.462)</td>
<td>(0.440)</td>
</tr>
<tr>
<td><strong>0.1</strong></td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.035)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.400)</td>
<td>(0.364)</td>
<td>(0.470)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.388)</td>
<td>(0.345)</td>
<td>(0.465)</td>
</tr>
<tr>
<td><strong>0.3</strong></td>
<td>(0.036)</td>
<td>(0.034)</td>
<td>(0.030)</td>
</tr>
<tr>
<td><strong>Factor scores</strong></td>
<td>(0.036)</td>
<td>(0.034)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>simple</td>
<td>(0.360)</td>
<td>(0.310)</td>
<td>(0.312)</td>
</tr>
<tr>
<td>weighted</td>
<td>(0.350)</td>
<td>(0.305)</td>
<td>(0.305)</td>
</tr>
</tbody>
</table>

*Table 13 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).*
### Table 14

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( ICC_k )</th>
<th>( r_e = 0.7, r_u = 0.2 )</th>
<th>( r_e = 0.7, r_u = -0.2 )</th>
<th>( r_e = 0.2, r_u = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple</td>
<td>Weighted</td>
<td>Simple</td>
<td>Weighted</td>
</tr>
<tr>
<td>0.05</td>
<td>Factor 1</td>
<td>-5.26</td>
<td>-5.29</td>
<td>-5.56</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>-3.90</td>
<td>-4.74</td>
<td>-20.99</td>
</tr>
<tr>
<td>0.1</td>
<td>Factor 1</td>
<td>-8.89</td>
<td>-9.30</td>
<td>-26.67</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>-9.00</td>
<td>-11.08</td>
<td>-25.53</td>
</tr>
<tr>
<td>0.3</td>
<td>Factor 1</td>
<td>-11.67</td>
<td>-11.88</td>
<td>-16.67</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>-13.89</td>
<td>-12.86</td>
<td>-19.87</td>
</tr>
</tbody>
</table>

Table 14 Percentage relative reduction (%) in terms of RMSE for simple and weighted averages of variables associated to each of the factors of MEBLUP over EBLUP, \( \Delta_k \) two-factors CFA model.

Table 13 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 14, 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.
4.4.4 Final remarks on the 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.

4.5 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 Well-being” of the BES framework. Housing quality is also an important determinant of well-being 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.

4.5.1 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 well-being, although monetary measurement is not always exhaustive for measuring poverty and well-being 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 $I_{DE}^E$, which is calculated as follows (Atkinson et al., 2002):

$$I_{DE}^E = \frac{I^D_i}{n^E_i}, i = 1, \ldots, N,$$

where $i = 1, \ldots, N$ denotes households, $I^D_i$ is the disposable household income, and $n^E_i$ is the equivalised household size calculated in the following way:

$$n^E_i = 1 + 0.5 \cdot (HM_{14+} - 1) + 0.3 \cdot HM_{13-},$$

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.

4.5.2 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 15 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 11. Figure 12 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).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe material deprivation</td>
<td>0.010</td>
<td>0.733</td>
</tr>
<tr>
<td>Smog</td>
<td>0.757</td>
<td>0.025</td>
</tr>
<tr>
<td>Noise</td>
<td>0.617</td>
<td>0.154</td>
</tr>
<tr>
<td>Crime</td>
<td>0.659</td>
<td>0.130</td>
</tr>
<tr>
<td>Housing ownership</td>
<td>0.096</td>
<td>-0.589</td>
</tr>
<tr>
<td>Presence of humidity</td>
<td>0.010</td>
<td>0.596</td>
</tr>
<tr>
<td>Darkness inside the house</td>
<td>-0.002</td>
<td>0.551</td>
</tr>
<tr>
<td>Absence of rubbish in the street</td>
<td>-0.843</td>
<td>0.084</td>
</tr>
<tr>
<td>Absence of damages in public buildings</td>
<td>-0.810</td>
<td>0.012</td>
</tr>
<tr>
<td>Log equivalised disposable income</td>
<td>0.139</td>
<td>-0.398</td>
</tr>
</tbody>
</table>

*Table 15 Factor structure for two latent factors using EFA.*
Figure 11 Housing quality sub-dimensions.

Figure 12 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 13 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 $\delta$ (Box and Cox, 1964) in order to approximate the normal distribution assumption needed for the SAE models. For Factor 1 we used $\delta=3.2$ and for Factor 2 we used $\delta = 3.0$.

![Figure 13 Factor scores histograms from CFA two-factor model after transformations.](image)

4.5.3 **Small area estimates and model diagnostic**

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. (2017a). The MSEs of the MEBLUPs are estimated as in Moretti et al. (2017b), taking into account for the variability arising from the CFA model as proposed in Moretti et al. (2017a). This method has been evaluated in the simulation study in comparison (b). We are able to estimate the multivariate EBLUPs because this approach in case of responses measured on different scales has not been studied in SAE yet.

In the case of areas where \( n_d = 0 \) the direct estimates and model random effects cannot be estimated. Therefore, considering formula (3) it holds that:

\[
\hat{f}_{dm}^{EBLUP} = \hat{f}_{dm}^{Synthetic} = \bar{X}'_d \hat{\beta}, m = 1,2 \\
\hat{f}_{d}^{MEBLUP} = \hat{f}_{d}^{Synthetic} = \bar{X}'_d \hat{\beta}
\] (17)

where \( \hat{f}_{dm}^{EBLUP} \) and \( \hat{f}_{dm}^{MEBLUP} \) 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. This is due to the fact that the random effect of a zero sample size area cannot be estimated along with the direct estimate. Therefore, the EBLUP and MEBLUP coincide with the synthetic model estimator (Rao and Molina, 2015: Ch. 7). The quantity \( \hat{f}_{dm}^{Synthetic} = \bar{X}'_d \hat{\beta} \) is estimated from the univariate Battese, Harter and Fuller model (Battese et al,
The final EBLUP and MEBLUP factor score means are then transformed for enabling interpretation and mapping using the ‘Min-Max’ criterion (OECD-JRC, 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:

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

where $\hat{f}_{dm}^{EBLUP}$ 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 $d=1,...,D$. $\hat{f}_{dm}^{MEBLUP}$ are rescaled with the same formula but considering the multivariate EBLUP estimates. The rescaled estimates are denoted by $\hat{f}_{dm}^{MEBLUP*}$.

We proceed with the MEBLUP of factor score means and interpret our findings. Table 16 shows the percentiles for the transformed latent housing quality indicators based on MEBLUP of factor score means. Figure 14 shows the maps of residential area deprivation and housing material deprivation, respectively.
Table 16 Percentiles for transformed latent housing quality indicators based on MEBLUP of factor score means.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential area deprivation</td>
<td>0.000</td>
<td>0.261</td>
<td>0.266</td>
<td>0.270</td>
<td>1.000</td>
</tr>
<tr>
<td>Housing material deprivation</td>
<td>0.000</td>
<td>0.418</td>
<td>0.457</td>
<td>0.502</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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 and this means that even in the economic well-being sub-dimensions (housing quality) we need to disentangle the indicators. These differences can be seen in the maps. In order to help the reader, who may not be familiar
with Tuscany geography, we added a map of Tuscany provinces (NUTS-2 level) in Appendix C (Figure C1). Looking at residential area deprivation estimates (Figure 14; 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 14, 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.

The application 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 15 Root Mean Squared Error (RMSE) of MEBLUP (___) and direct estimates (---) of residential area deprivation small areas with $n_d > 0$.

Figure 16 Root Mean Squared Error (RMSE) of MEBLUP (___) and direct estimates (---) of housing material deprivation small areas with $n_d > 0$. 
Figure 15 and Figure 16 show the Root Mean Squared Error (RMSE) of MEBLUP and direct estimates calculated via the post-stratified estimator (since we are using the EU-SILC survey weights) (Rao and Molina, 2015) for those small areas with $n_d > 0$ for residential area deprivation and housing material deprivation, respectively. Figure 17 and Figure 18 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.

![RMSE residential area deprivation](image)

*Figure 17 Root Mean Squared Error (RMSE) of MEBLUP (___) and EBLUP (---) of residential area deprivation small areas with $n_d > 0$.  

142
Figure 18 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. These results are in line with the simulation study, where we show that the use of MEBLUP does not provide very large improvements in terms of MSE compared to the EBLUP.

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

$$\hat{\sigma}_{e,f_1}^2 = 0.086, \sigma_{u,f_1}^2 = 0.023,$$
\[
\hat{\sigma}_{e, f_2}^2 = 0.170, \sigma_{u, f_2}^2 = 0.017, \\
\Sigma_e = \begin{bmatrix} 0.086 & 0.012 \\ 0.012 & 0.169 \end{bmatrix} \text{with } \hat{r}_e = 0.10, \\
\Sigma_u = \begin{bmatrix} 0.023 & 0.015 \\ 0.015 & 0.016 \end{bmatrix} \text{with } \hat{r}_u = 0.78.
\]

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

Figure 19 and Figure 20 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 19 Q-Q plots of the residuals estimated from the univariate BHF model.*
4.6 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

<table>
<thead>
<tr>
<th>Correlation structure</th>
<th>One-factor model</th>
<th>Two-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$ICC_k$</td>
<td>SRMR</td>
</tr>
<tr>
<td>$r_e = 0.2, r_u = 0.2$</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td>$r_e = 0.7, r_u = 0.2$</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td>$r_e = 0.7, r_u = -0.2$</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>

*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 4.5

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

*Table B1 Descriptive statistics of the observed variables (EU-SILC, Tuscany 2009).*
## Access to public services

<table>
<thead>
<tr>
<th></th>
<th>Absolute frequency</th>
<th>Relative frequency %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very difficult</strong></td>
<td>133</td>
<td>9.19</td>
</tr>
<tr>
<td><strong>Some difficulties</strong></td>
<td>249</td>
<td>17.20</td>
</tr>
<tr>
<td><strong>Easy</strong></td>
<td>631</td>
<td><strong>43.58</strong></td>
</tr>
<tr>
<td><strong>Very easy</strong></td>
<td>290</td>
<td>20.03</td>
</tr>
<tr>
<td><strong>Not needed</strong></td>
<td>145</td>
<td>10.01</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1448</td>
<td>100.00</td>
</tr>
</tbody>
</table>

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

## Perception of damages to public buildings

<table>
<thead>
<tr>
<th></th>
<th>Absolute frequency</th>
<th>Relative frequency %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Always</strong></td>
<td>65</td>
<td>4.49</td>
</tr>
<tr>
<td><strong>Often</strong></td>
<td>83</td>
<td>5.73</td>
</tr>
<tr>
<td><strong>Sometime</strong></td>
<td>294</td>
<td>20.30</td>
</tr>
<tr>
<td><strong>Never</strong></td>
<td>1006</td>
<td><strong>69.48</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1448</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*Table B3 Frequency distribution of damages to public buildings (EU-SILC, Tuscany 2009).*
### Table B5 Descriptive statistics of the auxiliary variables (EU-SILC, Tuscany 2009).

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

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

<table>
<thead>
<tr>
<th>Perception of rubbish in the street</th>
<th>Absolute frequency</th>
<th>Relative frequency %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always</td>
<td>75</td>
<td>5.18</td>
</tr>
<tr>
<td>Often</td>
<td>82</td>
<td>5.66</td>
</tr>
<tr>
<td>Sometimes</td>
<td>308</td>
<td>21.27</td>
</tr>
<tr>
<td>Never</td>
<td>983</td>
<td>67.89</td>
</tr>
<tr>
<td>Total</td>
<td>1448</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Appendix C: Tuscany region map

Figure C1 Tuscany provinces.
Appendix D: Specification of the R functions used and issues in computation

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.

For the analysis carried out in this paper we did not have a very powerful computer (Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz, RAM 8.00GB (7.88GB usable)), so in the multivariate mixed effect model fitting took some time. The simulation study took around 2 weeks for running. This is due to the CFA model and multivariate mixed effect model fitting.
5 Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP

Introduction to the paper

This chapter is a paper submitted to *Communications in Statistics – Simulation and Computation* and is currently under review. I am the lead author of the paper and responsible for the writing of the article and carrying out all of the analysis and simulation studies. All ideas and approaches are discussed through the normal supervision process.

This paper is strongly linked to Chapter 4. Here, we developed a parametric bootstrap algorithm for MSE approximation of a MEBLUP. This method is applied and particularised to the case of factor scores in Chapter 4 following the main results on MSE estimation of EBLUP of factor scores means evaluated in Chapter 3. We provide an application using the well-known corn and soybean data (LANDSAT data). We understand that this thesis is about measuring multidimensional well-being indicators, however, since the focus of the chapter is on developing new methodology, we use the well-known LANDSAT data that is commonly used in the SAE literature. Since this is a methodological paper, the application can be seen as an example only. For a meaningful application in terms of well-being we refer to Chapter 4.
Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP

Angelo Moretti, Natalie Shlomo and Joseph Sakshaug

Social Statistics, School of Social Sciences, University of Manchester, United Kingdom

Abstract

This article deals with mean squared error (MSE) estimation of a multivariate empirical best linear predictor (MEBLUP) under the unit-level multivariate nested-errors regression model for small area estimation via parametric bootstrap. A maximum likelihood estimation approach for estimating model parameters is used. A simulation study is designed in order to evaluate the performance of our algorithm and compare it with the parametric bootstrap for the univariate empirical best linear unbiased predictor (EBLUP) which has been shown to be consistent to the true MSE. The simulation shows that, in line with the literature, MEBLUP provides estimates with lower MSE than EBLUP. Further, the MSE estimates provided by the parametric bootstrap are unbiased. We also provide a short empirical analysis based on actual survey and satellite data collected from the U.S. Department of Agriculture. In this paper we study the case of a vector of means; however, the proposed method can be extended to other types of parameters, linear and non-linear.

Keywords: Multivariate empirical best linear unbiased predictor; Model-based
Regional policies need to base their funding allocation on reliable statistical information. However, large-scale social sample surveys are typically not designed to be representative at a low geographical level. Thus, small area estimation (SAE) methods based on models might provide more accurate estimates than direct estimators (Rao and Molina, 2015). Mixed effects linear regression models are traditionally used in order to provide more accurate estimates than design-based estimation techniques. These kinds of models have been used extensively in the literature, and for a detailed review of these in SAE we refer to Rao and Molina (2015). Estimating the precision of small area estimates is a crucial and challenging exercise (Marchetti et al, 2012).

As Molina (2009) points out, when the target of inferential interest is a random vector, multivariate regression models might be a natural model setting. Indeed, multivariate models take into account the correlation structure among the vector components; hence, it is possible to improve the precision of the estimates over the univariate case Datta et al. (1999). Fuller and Harter (1987) develop a multivariate mixed effects model to predict a vector of means of multiple characteristics of a finite population. Datta et al. (1999) propose a multivariate empirical best linear unbiased predictor (MEBLUP) and empirical bayes (EB) approach for small area mean vectors along with an approximation
for the mean squared error (MSE). Some recent work in the literature are Molina (2009) and Baillo and Molina (2009). Molina (2009) deals with the multivariate mixed effects model with logarithmic transformation, and Baillo and Molina (2009) study a particular case of the multivariate nested error regression model with correlated sampling errors. Both articles provide analytical approximations for the MSE.

The best linear unbiased predictor (BLUP) depends on unknown quantities (variance components). When these quantities are estimated using suitable estimation techniques, we obtain the empirical BLUP (EBLUP). Unfortunately, the exact MSE of an EBLUP cannot be obtained in closed form; therefore, approximations have been proposed in the literature (González-Manteiga et al., 2008a). Kackar and Harville (1981) propose an approximation of the MSE assuming normality of the errors and random effects. Prasad and Rao (1990) obtain an MSE approximation for models with block-diagonal covariance matrices. Datta and Lahiri (2000) provide analytical approximations for general models with a block-diagonal structure when variance components are estimated by maximum likelihood (ML) or restricted maximum likelihood (REML). Das et al. (2004) deals with approximations for a wider class of models. In multivariate SAE, Datta et al. (1999) propose a second-order unbiased analytical approximation for the MSE of a multivariate EBLUP following Datta and Lahiri (2000).

When the MSE exact analytical estimator cannot be computed, an alternative way to approximate the MSE is via bootstrap techniques. It is important to highlight that, even when large sample approximations are available, the bootstrap may provide more accurate estimation alternatives due to its second-order accuracy (González-Manteiga et
This property is not achieved by the majority of asymptotic methods. We refer to Efron and Tibshirani (1993) and Hall (1992) for a broader discussion of this property.

In this article, we assume that the values of the target vector in the units of a finite population are realizations of a random multivariate variable following the Fuller and Harter multivariate mixed effects model (Fuller and Harter, 1987). We propose a maximum likelihood (ML)-based parametric bootstrap procedure designed for estimating a vector of MSEs for a vector of means when the auxiliary variables are available at the unit-level.

This paper is organised as follows. In section 2 the multivariate mixed effects model is reviewed along with the multivariate EBLUP. In section 3 we discuss the MSE estimation via bootstrapping. In section 4 we study the behaviour of our bootstrap MSE in a simulation study and compare with the univariate case. In section 5 we present an example based on survey data on corn and soy bean production. In section 6 we conclude the work with some final remarks.

5.2 Multivariate Small Area Estimation of a Means Vector

Let \( d = 1, \ldots, D \) denote the small areas for which we want to compute the estimates and let us consider a sample \( s \subset \Omega \) of size \( n \) drawn from the target finite population \( \Omega \) of size \( N \). The non-sampled units, \( N - n \) are denoted by \( r \), hence, \( s_d = s \cap \Omega_d \) is the subsample 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 $N_d - n_d$ dimension.

Considering $y_{di} = (y_{d1i}, ..., y_{dKi})'$, which denotes the $K \times 1$ vector of interest for $i = 1, ..., N_d$, $d = 1, ..., D$, we can write the target mean vector as follows:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}. \quad (1)$$

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

$$\bar{Y}_d = N_d^{-1} \left( \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} y_{di} \right). \quad (2)$$

The quantity $\sum_{i \in r_d} y_{di}$ is not observed, so it needs to be predicted. In this work we make use of the multivariate mixed effects model advocated in SAE by Fuller and Harter (1987).

### 5.2.1 Multivariate nested-error linear regression model

We assume that unit-specific auxiliary variables $x_{id}$ are available for all the population elements in each small area $d$. Also, we assume that the following linear model relates the response variables to the covariates as follows:
\[ \xi: \mathbf{y}_{di} = \mathbf{x}_{di}\beta + \mathbf{u}_d + \mathbf{e}_{di}, \quad d = 1, \ldots, D, \quad i = 1, \ldots, N_d, \]  
\[ \mathbf{u}_d \sim N_K(\mathbf{0}, \Sigma_u), \quad \mathbf{e}_{di} \sim N_K(\mathbf{0}, \Sigma_e) \]

where, \( \mathbf{y}_{di} \) is the response vector for the \( i^{th} \) unit from the \( d^{th} \) small area, \( \mathbf{x}_{di} = (1, \mathbf{x}_{1di}, \mathbf{x}_{2di})' \) is the \( K \times p \) matrix of the \( p \) auxiliary variables, \( \beta \) is a \( K \times p \) matrix of regression unknown coefficients, \( \mathbf{u}_d \) is the \( K \times 1 \) vector of the area effects, and \( \mathbf{e}_{di} \) is the \( K \times 1 \) vector of the individual effects; \( \mathbf{u}_d \) and \( \mathbf{e}_{di} \) are assumed independent. Here, the \( K \times K \) positive-definite matrices \( \Sigma_u \) and \( \Sigma_e \) 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:

\[ \bar{Y}_d = \bar{X}_d\beta + \mathbf{u}_d \]

where \( \bar{X}_d \) denotes the known population covariates means for area \( d \).

### 5.2.2 Estimation and prediction of unknown parameters

For simplicity we now make use of the following notation:

\[ \mathbf{y}_d = col_{i \in s_d}(\mathbf{y}_{di}), \quad \mathbf{Y} = col_{1 \leq d \leq D}(\mathbf{y}_d). \]
\[ x_d = col_{i \in S_d}(x_{di}), \quad X = col_{1 \leq d \leq D}(x_d) \]

\(Y\)' is the \(NK\) vector of observations and \(X\) the \(NK \times pK\) matrix of covariates. The operator \(\otimes\) denotes the Kronecker product.

Let us now denote the covariance matrix and expectation of \(Y\) by

\[ V(Y) = \text{block diag}(V_{11}, \ldots, V_{DD}), E(Y) = X\beta, \tag{5} \]

where \(V_{dd} = (J_{dd} \otimes \Sigma_u) + (I_{nd} \otimes \Sigma_e)\). \(J_{dd}\) is the \(n_d \times n_d\) matrix with every element equal to one, and \(I_{nd}\) is an identity matrix. The empirical best linear unbiased estimator of the matrix of regression coefficients is given by:

\[ \hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y. \tag{6} \]

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

\[ \hat{u}_d = (\bar{Y}_d - \bar{x}'d\hat{\beta})[(\hat{\Sigma}_u + n_d^{-1}\hat{\Sigma}_e)^{-1}\hat{\Sigma}_u, d = 1, \ldots, D \tag{7} \]

\(\hat{\Sigma}_u\) and \(\hat{\Sigma}_e\) are estimators of \(\Sigma_u\) and \(\Sigma_e\). We refer to Schafer et al. (2002) for the estimation algorithm where the Maximum Likelihood (ML) approach is used.
The Multivariate Empirical Best Linear Unbiased Predictor (MEBLUP) of \( \bar{Y}_d \) is given by Fuller and Harter (1987):

\[
\hat{Y}_d^{MEBLUP} = \bar{X}_d' \hat{\beta} + \hat{u}_d = \bar{X}_d' \hat{\beta} + (\bar{Y}_d - \bar{X}_d' \hat{\beta})[(\hat{\Sigma}_u + n_d^{-1} \hat{\Sigma}_e)^{-1}] \hat{\Sigma}_u], d = 1, \ldots, D .
\] (8)

5.3 Parametric bootstrap

This section introduces a bootstrap algorithm approximation of the MSE of \( \bar{Y}_d^{MEBLUP} \) denoted by \( MSE(\bar{Y}_d^{MEBLUP}) \) and given by the following (Kackar and Harville, 1984):

\[
MSE(\bar{Y}_d^{MEBLUP}) = E \left[ \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d \right) \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d \right)' \right] =
\] (9)

\[
= MSE(\bar{Y}_d^{MBLUP}) + E \left[ \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d^{MBLUP} \right) \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d^{MBLUP} \right)' \right] + \\
+ E \left[ \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d \right) \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d \right)' \right] \\
+ E \left[ \left( \bar{Y}_d^{MBLUP} - \bar{Y}_d \right) \left( \bar{Y}_d^{MEBLUP} - \bar{Y}_d^{MBLUP} \right)' \right]
\]

where we denote the Multivariate Best Linear Unbiased Predictor of \( \bar{Y}_d \) (assuming known covariance matrices) by \( \bar{Y}_d^{MBLUP} \). It can be shown that the last two terms of equation (9) are equal to zero for any unbiased and translation invariant estimator of the variance components (Kackar and Harville, 1984).
The term $E \left[ \left( \hat{Y}^{MEBLUP}_d - \bar{Y}^{MEBLUP}_d \right) \left( \hat{Y}^{MMLUP}_d - \bar{Y}^{MMLUP}_d \right)' \right]$ accounts for the estimation of the variance components.

We propose to use the parametric bootstrap procedure proposed by González-Manteiga et al. (2008a) extended to the multivariate mixed effects model used in this article. Let $\Omega$ be a finite population of dimension $N$ generated by the superpopulation model given by (3), and let $\bar{Y}_d = N^{-1}_d \sum_{i=1}^N Y_{di}$ be the linear vector of target parameters of $\Omega$. Let $S$ be a random sample drawn from $\Omega$ of dimension $n$, using a specific sampling design.

We list the steps of the algorithm as follows:

1. Fit the multivariate model (3) to the sample $S$, $y_S = (y_{1S}', ..., y_{DS}')'$, and obtain the estimates of the model parameters: let us denote the estimates as $\hat{\beta}, \hat{\Sigma}_u$, and $\hat{\Sigma}_e$.

2. Generate the bootstrap area effects $u_{d}^{*(b)} \sim iid \mathcal{N}_K(0, \hat{\Sigma}_u), d = 1, ..., D$. We use the symbol * for the bootstrap quantities, while (b) refers to the index of the $b^{th}$ bootstrap replication, $b=1,...,B$.

3. Generate the bootstrap errors for the sample units $e_{di}^{*(b)} \sim iid \mathcal{N}_K(0, \hat{\Sigma}_e), i \in s_d$ independently of the $u_{d}^{*(b)}$, $d = 1, ..., D$.

4. Calculate the true means vectors for each small area of the bootstrap population as follows:
\[ y_{d_i}^{\ast(b)} = x_{di}' \hat{\beta} + u_{d_i}^{\ast(b)} + e_{d_i}^{\ast(b)}, \quad d = 1, \ldots, D, \]

where \( \bar{X}_d \) denotes the known population covariates means.

5. Generate the responses for the sample units by using the sample covariates vectors 
\( x_{d_i}, \quad i \in S_d \):

\[ \xi^*: \quad y_{d_i}^{\ast(b)} = x_{di}' \hat{\beta} + u_{d_i}^{\ast(b)} + e_{d_i}^{\ast(b)}, \quad d = 1, \ldots, D, \]

The bootstrap sample data vector is denoted by 
\( y_{S}^{\ast(b)} = [ (y_{1S}^{\ast(b)})', \ldots, (y_{DS}^{\ast(b)})' ]' \). Under model \( \xi^* \), given \( S \), the MSE of \( \bar{Y}_d^{\text{MEBLUP}^*} \) is denoted by \( \text{MSE}_*(\bar{Y}_d^{\text{MEBLUP}^*}) \). Hence, for estimating the MSE of \( \bar{Y}_d^{\text{MEBLUP}} \) given in (9), we propose to use the bootstrap MSE.

6. Fit model (3) to the bootstrap sample data \( y_{S}^{\ast(b)} \) and obtain the bootstrap MEBLUPs 
\( \bar{Y}_d^{\ast(b)}, \quad d = 1, \ldots, D \).

7. Replicate steps (2) through (6) for \( b = 1, \ldots, B \). The Monte Carlo approximation of the bootstrap estimator \( \text{MSE}_*(\bar{Y}_d^{\text{MEBLUP}^*}) \) is given by:

\[ \text{mse}_*(\bar{Y}_d^{\text{MEBLUP}^*}) = \frac{1}{B} \sum_{b=1}^{B} (\bar{Y}_d^{\ast(b)} - \bar{Y}_d^{\ast(b)})(\bar{Y}_d^{\ast(b)} - \bar{Y}_d^{\ast(b)})', \quad d = 1, \ldots, D. \]
We note that when \( B \to \infty \), \( \text{mse}_e(\hat{V}^{MEBLUP}_d) \) is a consistent estimator of \( \text{MSE}_e(\hat{V}^{MEBLUP}_d) \) (Rao and Molina, 2015: p. 183).

The parametric bootstrap procedure has been proven to be consistent as an estimator of the true MSE under the univariate unit-level model (González-Manteiga et al., 2008a) and the Fay-Harriot model (González-Manteiga et al., 2008b). In general, the proofs in these papers have been based on the fact that the final estimate of the MSE obtained by the bootstrap procedure is consistent if the model parameter estimates are consistent. Since we are using the Maximum Likelihood estimators for estimating the model parameters in the multivariate SAE approach, which have well-known consistency properties as shown in Sweeting (1980) and Mardia and Marshall (1984), we can prove the consistency of our proposed parametric bootstrap algorithm to the true MSE by the method of imitation (see Shao and Tu, 1995, p.76).

### 5.4 Simulation study

This simulation is designed to study the performance of the bootstrap MSE estimator in Section 5.3 under a multivariate mixed effects model when the target vector parameter is a vector of means. The results are compared with the “truth” as described in Section 5.4.2 and the aim is to show that a multivariate bootstrap procedure will be appropriate in the case of multivariate SAE. The bias is also studied. In the case of the univariate SAE, the bootstrap MSEs are compared with the Prasad-Rao analytical approximation.
of MSE Prasad and Rao (1990). This simulation belongs to the group of design-based under model data simulations (Münnich, 2014). One outcome of the multivariate mixed-effect model (Fuller and Harter, 1987) is used as fixed universe.

5.4.1 Generating the population

We generate an unbalanced population using parameters with \( N = 20,000, D = 80, \) and \( 130 \leq N_d \leq 420. \) \( N_d, d = 1, \ldots, D \) is generated from the discrete uniform distribution, \( N_d \sim \mathcal{U}(a = 130, b = 420), \) with \( \sum_{d=1}^{D} N_d = 20,000. \) The simulation modelling parameters (\( \beta \) and variances in \( \Sigma_e \) and \( \Sigma_u \)) have been chosen by fitting two mixed effect models on the survey and satellite data for corn and soy beans in 12 Iowa counties, obtained from the 1978 June survey of the U.S. Department of Agriculture and from land observatory satellites, also known as LANDSAT during the 1978 growing season. These data were also used by Datta et al. (1999) where a large simulation study was conducted in order to evaluate the performances of a multivariate EBLUP under the multivariate mixed-effect model.

\( y_{dl} \) observations are generated according to the multivariate mixed effects model (3) described in section 5.2. Two uncorrelated covariates are generated from the discrete uniform distribution as follows:

\[
X_1 \sim \text{dUnif}(20000,145,459), \quad X_2 \sim \text{dUnif}(20000,55,345).
\]

The \( \beta(2 \times 3) \) matrix of regression coefficients is given by:
\[
\beta = \begin{bmatrix}
17.97 & 0.36 & -0.03 \\
-16.35 & 0.02 & 0.50
\end{bmatrix}
\]

The response vector \( Y_{(2 \times 1)} = (Y_1, Y_2)' \) was generated according to the following variance components:

\[
\Sigma_e = \begin{bmatrix}
297.71 & -150.82 \\
-150.82 & 170.29
\end{bmatrix}
\]

\[
\Sigma_u = \begin{bmatrix}
63.31 & 35.35 \\
35.35 & 219.32
\end{bmatrix}
\]

with associated correlation coefficients \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \). The intra-class correlations are 0.2 and 0.6 for the first and second components, respectively; these have been chosen according to the LANDSAT data. We also studied the case where \( \rho_e \) and \( \rho_u \) have the same signs i.e. \( \rho_e = 0.7 \) and \( \rho_u = 0.3 \).

For computational reasons, we did not perform a simulation varying many \( \rho_e, \rho_u \) and intra-class correlation coefficients. For more details on the role of \( \rho_e \) and \( \rho_u \) in multivariate SAE we refer to Datta et al. (1999). In particular, Datta et al. (1999) conduct a large simulation experiment based on LANDSAT data showing that the use of multivariate modelling in SAE provides more efficient estimates than the univariate case. The gains in efficiency depend on \( \rho_e, \rho_u \) and intra-class correlation coefficient.

This work shows that when \( \rho_e \) and \( \rho_u \) have opposite signs the multivariate model performs much better than the univariate model in terms of MSE. Of course, when \( \rho_e \) and \( \rho_u \) are small (theoretically tending to zero), we are close to the independence case, where the univariate modelling performs identically to the multivariate modelling Datta
et al. (1999). They also provide an analytical approximation of MSE. We propose a bootstrap method since more attention has been given to resampling techniques in SAE due to their statistical properties.

The steps of the simulation are as follows (for \( d = 1, \ldots, D \) and \( s = 1, \ldots, S \)):

1. Draw \( S = 500 \) simple random samples without replacement of size \( n = 1,000 \) from the simulated population;

2. Fit the univariate Battese, Harter and Fuller model (Battese et al., 1988) on each sample \( s \) and obtain the estimates of the model parameters: \( \hat{\sigma}^2_e(s) \), \( \hat{\sigma}^2_u(s) \), \( \hat{\beta}(s) \), thus the univariate EBLUPs are estimated: \( \hat{Y}_{d(s),1}^{EBLUP} \) and \( \hat{Y}_{d(s),2}^{EBLUP} \);

3. Estimate the MSEs of \( \hat{Y}_{d(s),1}^{EBLUP} \) and \( \hat{Y}_{d(s),2}^{EBLUP} \) on each sample \( s \) via parametric bootstrap (González-Manteiga et al., 2008a) with \( B = 500 \) replications and Prasad-Rao analytical approximation Prasad and Rao (1999); only for the \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) case;

4. Fit the multivariate mixed effects model in (3) and estimate the model parameters: \( \hat{\Sigma}_e^{(s)} \), \( \hat{\Sigma}_u^{(s)} \), \( \hat{\beta}(s) \), and the multivariate EBLUP: \( \hat{Y}_{d(s)}^{MEBLUP} \) for \( \rho_e = -0.7 \) and \( \rho_u = 0.3 \) and \( \rho_e = 0.7 \) and \( \rho_u = 0.3 \).

5. Estimate the vector of MSEs of \( \hat{Y}_{d(s)}^{MEBLUP} \) on each sample \( s \) via the parametric bootstrap proposed in section 5.3 with \( B = 500 \) replications.

The results are evaluated via the empirical MSE (EMSE), which is considered to be the “truth”, the averages of the bootstrap MSE across the \( S=500 \) samples, and the relative
bias (RBIAS) for each small area $d$. These quantities are respectively defined by the following estimators:

$$EMSE(\hat{Y}_d^{MEBLUP}) = S^{-1} \sum_{s=1}^{S} (\hat{Y}_{ds}^{MEBLUP} - \bar{Y}_d)(\hat{Y}_{ds}^{MEBLUP} - \bar{Y}_d)' ,$$

$$mse^B(\hat{Y}_d^{MEBLUP*}) = S^{-1} \sum_{s=1}^{S} mse_s(\hat{Y}_{ds}^{MEBLUP*}) ,$$

where we denote the bootstrap MSE of sample $s$ in area $d$ by: $mse_s(\hat{Y}_{ds}^{MEBLUP*})$.

$$RBIAS[mse, (\hat{Y}_d^{MEBLUP*})] = S^{-1} \sum_{s=1}^{S} [mse_s(\hat{Y}_{ds}^{MEBLUP*}) - EMSE(\hat{Y}_d^{MEBLUP})] / EMSE(\hat{Y}_d^{MEBLUP}) .$$

$EMSE(\hat{Y}_d^{MEBLUP})$ denotes the empirical mean squared error of $\hat{Y}_d^{MEBLUP}$, where the true mean vector is given by $\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}$. $mse^B(\hat{Y}_d^{MEBLUP*})$ denotes the average of the bootstrap MSEs (based on B=500 replicates) across the S=500 samples drawn in the simulation, and $RBIAS[mse, (\hat{Y}_d^{MEBLUP*})]$ denotes its relative bias.

These measures of performance are then averaged across the areas $d$ in order to provide summary statistics: $EMSE(\hat{Y}^{MEBLUP}), mse^B(\hat{Y}^{MEBLUP*})$ and $RBIAS[mse, (\hat{Y}^{MEBLUP*})]$.

The same measures can be written for the univariate case both for the bootstrap and Prasad-Rao (PR) approximation. We do not review the Prasad-Rao analytical
approximation in this paper, thus we refer to Prasad and Rao (1999) for theoretical
details.

5.4.2 Results

Here we compare first the MSE estimates obtained via the Prasad-Rao analytical
approximation with the MSE estimates obtained by parametric bootstrap for the
univariate case. Table 17 shows the descriptive statistics and bias of the Prasad-Rao and
bootstrap estimators for univariate EBLUP across the small areas. It can be seen that the
Prasad-Rao MSEs analytical approximations are slightly more biased than the bootstrap
MSEs (by comparing the EMSE with the its mean across the S=500 samples) under our
scenario. Figure 21 and Figure 22 show the relative bias of the MSEs; these show that
the Prasad-Rao MSE approximation overestimates the true MSE for some areas: the
relative bias of the Prasad-Rao MSE approximation is positive and larger than the
bootstrap case. This is particularly true for $y_1$. These results are also summarized in
Table 17. For more details on the Prasad-Rao approximation compared to the bootstrap
for univariate EBLUP we refer to (González-Manteiga et al., 2008a).
<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean</th>
<th>Median</th>
<th>IQR</th>
<th>SD</th>
<th>Rbias</th>
</tr>
</thead>
<tbody>
<tr>
<td>$mse^{PR}(mse(\hat{\bar{Y}}_{1}^{EBLUP}))$</td>
<td>19.00</td>
<td>18.97</td>
<td>0.23</td>
<td>0.18</td>
<td>0.20</td>
</tr>
<tr>
<td>$mse^{B}(mse(\hat{\bar{Y}}_{1}^{EBLUP}))$</td>
<td>18.53</td>
<td>18.68</td>
<td>0.31</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>$mse^{PR}(mse(\hat{\bar{Y}}_{2}^{EBLUP}))$</td>
<td>14.03</td>
<td>14.01</td>
<td>0.24</td>
<td>0.18</td>
<td>0.07</td>
</tr>
<tr>
<td>$mse^{B}(mse(\hat{\bar{Y}}_{2}^{EBLUP}))$</td>
<td>13.69</td>
<td>13.95</td>
<td>0.27</td>
<td>0.20</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*Table 17 Descriptive statistics and relative bias of the Prasad-Rao and bootstrap estimators for univariate EBLUP MSE across small areas, $EMSE(\hat{\bar{Y}}_{1}^{EBLUP}) = 18.35$, $EMSE(\hat{\bar{Y}}_{2}^{EBLUP}) = 13.42$, for $\rho_e = -0.7$ and $\rho_u = 0.3$. 
Figure 21 Relative bias of univariate EBLUPs’ MSEs of $y_1$ estimated via Prasad-Rao approximation and parametric bootstrap for $\rho_e = -0.7$ and $\rho_u = 0.3$.

Figure 22 Relative bias of univariate EBLUPs’ MSEs of $y_2$ estimated via the Prasad-Rao approximation and parametric bootstrap for the $\rho_e = -0.7$ and $\rho_u = 0.3$. 

171
In Table 18 we compare the results of the univariate with the multivariate bootstrap MSE estimation in terms of reduction in MSE and bias. We calculate the relative percentages of reduction in terms of EMSE (and bootstrap MSE) as follows: $\Delta_k = \frac{EMSE(\hat{y}_{\text{MEBLUP}}^k) - EMSE(\hat{y}_{\text{EBLUP}}^k)}{EMSE(\hat{y}_{\text{EBLUP}}^k)} \cdot 100$, where $k = 1, 2$ denotes the index of the $k^{th}$ component of the MEBLUP means vector. These are shown in brackets ( ). Figure 23 and Figure 24 show the comparisons of the bootstrap MSEs estimated for the EBLUPs and MEBLUPs for the opposite signs case only. It can be seen that, in line with the EMSEs, the multivariate bootstrap procedure provides predictions with lower variability than the univariate approach, and the MSE estimates show no noticeable bias across the small areas. The percentages of reduction in terms of MSE are smaller in the case of opposite signs in the covariance matrices.
<table>
<thead>
<tr>
<th>Correlation structure</th>
<th>Performance measure</th>
<th>EBLUP</th>
<th>MEBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_1$ $\gamma_2$</td>
<td>$\gamma_1$ $\gamma_2$</td>
<td></td>
</tr>
<tr>
<td>$\rho_e = -0.7, \rho_u = 0.3$</td>
<td>EMSE</td>
<td>18.35 13.42</td>
<td>16.82 (-8.34) 11.30 (-15.80)</td>
</tr>
<tr>
<td></td>
<td>$mse^B$</td>
<td>18.53 13.69</td>
<td>16.81 (-9.28) 11.42 (-16.58)</td>
</tr>
<tr>
<td></td>
<td>$RBIAS$</td>
<td>0.16 0.06</td>
<td>0.17 0.06</td>
</tr>
<tr>
<td>$\rho_e = 0.7, \rho_u = 0.3$</td>
<td>EMSE</td>
<td>17.31 13.23</td>
<td>17.00 (-1.79) 11.87 (-10.28)</td>
</tr>
<tr>
<td></td>
<td>$mse^B$</td>
<td>17.53 13.86</td>
<td>17.40 (-1.00) 12.16 (-12.27)</td>
</tr>
<tr>
<td></td>
<td>$RBIAS$</td>
<td>0.14 0.08</td>
<td>0.13 0.07</td>
</tr>
</tbody>
</table>

Table 18 Empirical mean squared error, bootstrap MSE, relative bias across the small areas: EBLUP and MEBLUP estimates – parametric bootstrap. ($\Delta_k$ shown in parenthesis).
Figure 23 Bootstrap MSEs $y_1$: comparison between EBLUP and MEBLUP $\rho_e = -0.7$ and $\rho_u = 0.3$.

Figure 24 Bootstrap MSEs $y_2$: comparison between EBLUP and MEBLUP $\rho_e = -0.7$ and $\rho_u = 0.3$. 
We also present the results of the coverage rates of the MSE estimates Figure 25 and Figure 26. The coverage rates are calculated as percentages of confidence intervals (at 95%) out of the total number of simulations that include the true mean obtained from the simulated population. For reasons of space, we focus on the case of $\rho_e = -0.7$ and $\rho_u = 0.3$ only. The results for the same-sign case are similar.

![Coverage rates y1](chart.png)

**Figure 25** Coverage rates (%) comparisons: MSEs of MEBLUP and EBLUP estimated via bootstrap and via Prasad-Rao approximation (PR) for EBLUP case: $y1$ for $\rho_e = -0.7$ and $\rho_u = 0.3$. 
Figure 26 Coverage rates (%) comparisons: MSEs of MEBLUP and EBLUP estimated via bootstrap and via Prasad-Rao approximation (PR) for EBLUP case: y2 for $\rho_e = -0.7$ and $\rho_u = 0.3$.

Looking at both Figure 25 and Figure 26, it can be seen that the bootstrap algorithm provides good coverage rates both for univariate and multivariate EBLUP cases. In case of $y_1$, the multivariate bootstrap returns slightly better coverage. Here, Prasad-Rao approximation performs slightly poorly. For $y_2$, there are not large differences between the Prasad-Rao analytical approximation and univariate bootstrap. The multivariate bootstrap shows larger coverage rates. Overall, coverage rates are very good for bootstrap MSE approximation.
5.4.3 Final remarks on the simulation study

The percentage of reductions in terms of MSE (and EMSE) may depend on the magnitude and sign of \( \rho_e \) and \( \rho_u \) as well as the intra-class correlation coefficient. As Datta et al. (1999) points out, when \( \rho_e \) and \( \rho_u \) have opposite signs, the multivariate model performs better in terms of MSE than the univariate modelling case. It can be seen that when the signs in the variance-covariance matrices are the same the percentages of reduction in terms of MSE of the multivariate EBLUP over the univariate one are smaller than in the case of opposite signs.

In order to give some intuition on this, we consider an argument which can be found in Datta et al (1999). Here, we refer to \( y_1 \) only. The same results are valid for \( y_2 \) under the same model assumptions. Let us define the following quantities: \( MSE(\hat{Y}_{d1}^{MBLUP}) \) and \( MSE(\hat{Y}_{d1}^{BLUP}) \) denoting the first component of \( MSE(\hat{Y}_d^{MBLUP}) \) (multivariate case) and MSE of the BLUP of \( \tilde{Y}_{d1} \) (univariate case), respectively. These quantities are dominating terms in \( MSE(\hat{Y}_d^{MBLUP}) \) and \( MSE(\hat{Y}_{d1}^{EBLUP}) \) (Datta et al, 1999). Datta et al. (1999) shows the following expression denoting the percentage of reduction of \( MSE(\hat{Y}_{d1}^{MBLUP}) \) over \( MSE(\hat{Y}_{d1}^{BLUP}) \):  

\[
P_{d1} = \frac{MSE(\hat{Y}_{d1}^{MBLUP}) - MSE(\hat{Y}_{d1}^{BLUP})}{MSE(\hat{Y}_{d1}^{BLUP})} \times 100
\]

\[
= \frac{n_d(\rho_u\sqrt{r_2} - \rho_e\sqrt{r_1})^2}{(1 + n_d r_1)[(1 - \rho_e^2) + n_d r_2(1 - \rho_u^2)] + n_d(\rho_u\sqrt{r_2} - \rho_e\sqrt{r_1})^2} \times 100,
\]
Where \( r_1 = \frac{\sigma_{u11}^2}{\sigma_{e11}^2} \) with \( \sigma_{u11}^2 \) and \( \sigma_{e11}^2 \) diagonal elements of \( \Sigma_u \) and \( \Sigma_e \), respectively (\( r_2 = \frac{\sigma_{u22}^2}{\sigma_{e22}^2} \) refers to \( y_2 \)). It can be noted that \( P_{d1} \) gets larger when \( \rho_u \) and \( \rho_e \) have opposite signs, as can be seen in the numerator of (17). Therefore, due to the dominance property, it holds multivariate analysis perform better than the univariate analysis in case of \( \rho_u \) and \( \rho_e \) having opposite signs.

Our bootstrap procedure performs well under the model assumptions, and we can see appreciable gains in efficiency in terms of MSE over the univariate modelling. Also, we note that there is no bias in the estimates of the MSE. Looking at the coverage rates we can see that the “true means” fall into the confidence intervals with large probabilities. Here, the multivariate bootstrap approach provides larger rates of coverage.

5.5 Application to Corn and Soy Bean Data

We apply our multivariate bootstrap method to the well-known corn and soy bean data of the LANDSAT data that was used in Battese et al. (1988) comparing the multivariate and univariate models. LANDSAT comprises survey and satellite data for corn and soy beans for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and from land observatory satellites during the 1978 growing season. The data file consists of \( n = 37 \) observations, \( D = 12 \) areas, and the following variables:

- CornHec: hectares of corn (\( y_1 \));
• SoyBeansHec: hectares of soy beans \((y_2)\);

• CornPix: number of pixels of corn in sample segment within county \((X_1)\);

• SoyBeansPix: number of pixels of soy beans in sample segment within county \((X_2)\).

As shown, the county means of number of pixels per segment of corn and soy beans, from satellite data, for 12 counties in Iowa is also used where we have the population size, sample size, and means of these auxiliary variables. These data files can be downloaded from Molina and Marhuenda (2015). In order to provide better modelling fit we applied Box-Cox family transformations (Box and Cox, 1964) to the response variables.
Figure 27 Bootstrap RMSEs $y_1$ – corns: comparison between EBLUP (---) and MEBLUP (___).

Figure 28 Bootstrap RMSEs $y_2$ – soy beans: comparison between EBLUP (---) and MEBLUP (___).
Figure 27 and Figure 28 show the RMSE of the univariate and multivariate EBLUPs where the small areas are ordered by growing sample sizes. It can be seen that the multivariate bootstrap algorithm provides estimates with smaller variability than the univariate case as was confirmed in the simulation study. The model diagnostics show good model fitting in both cases.

5.6 Conclusion

In this paper we proposed the use of parametric bootstrap for estimating MSEs vectors for means vectors of small domains under a multivariate mixed effects model for unit-level SAE. The multivariate SAE is more appropriate than the univariate SAE in the case of correlated responses. Indeed, in this case, multivariate mixed effects models may lead to more reliable estimates than the univariate BHF model. This, of course, needs to be taken into account when estimating the MSE and hence we have proposed the parametric bootstrap for the MEBLUP. In the simulation study we assessed empirically the behaviour of our approach for estimating the MSE and in particular the bias. Our results are in line with the literature and no bias is shown.

Although this work focuses on vectors of means as the target inferential parameter, this bootstrap procedure can be extended to other quantities in a multivariate setting. Non-parametric bootstrap procedures could be studied in future work, and comparisons between the two methodologies would be useful for practitioners. Normality assumptions can be relaxed according to Hall and Maiti (2006). Furthermore, hybrid
bootstrap MSE estimators should be considered. González-Manteiga et al. (2008a) studied hybrid bootstrap MSE estimators which are second-order unbiased. Other interesting extensions to this work may involve the study of robustness to non-normality.

Acknowledgments

This work was financially supported by the following grants: ESRC DTC Award and Advanced Quantitative Methods.
Appendix A Issues in computations

The bootstrap algorithm along with the algorithm used for the multivariate mixed effect model fitting are computationally intensive.

The simulation study took around 10 days to run completely. However, we did not have a very powerful computer available: (Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz, RAM 8.00GB (7.88GB usable)). Regarding the application to real data we did not encounter particular problems in the computation. This shows that the method can be easily applied even with a not very powerful computer.
6 Discussion

6.1 Summary of this work and relationship to the literature

This thesis investigates the issue of multivariate small area estimation in data dimensionality reduction. The focus has been placed on the measurement of multidimensional well-being indicators. The empirical best linear unbiased prediction (EBLUP) under linear mixed effects models is the estimation approach we followed in unit-level SAE. We discuss the need for multidimensional indicators in this field and the problem of multivariate small area estimation. Unplanned domains often arise in large-scale sample surveys; therefore, direct estimators, using only survey weights, may provide unreliable estimates for small domains (Rao and Molina, 2015). We will now briefly discuss our main findings considering our research questions.

In Chapter 3 we answered research questions 1 and 2 mentioned in the introduction. We studied a particular SAE problem: the estimation of a latent well-being variable mean under a univariate EBLUP approach. Here, we compared the use of simple and weighted averages of standardized EBLUPs of single indicators to the use of EBLUPs of factor score means. We found that the use of factor scores provides more efficient estimates, in terms of variability, in data dimensionality reduction than weighted and simple averages. Also, factor scores show good ranking of estimates particularly when the intra-class
correlation is small. An application on economic well-being using EU-SILC Italian data is proposed: we provided estimates for Tuscany municipalities with associated bootstrap mean squared errors. In this chapter, we proposed a bootstrap procedure for factor score means MSE estimation. In the simulation study, we showed that when the variability arising from the factor analysis model is not taken into account in the MSE estimation, the MSE estimates may be biased (underestimated).

Chapter 4 deals with research question number 3. Here, we propose the use of the Harter and Fuller (1987) multivariate mixed effects model in data dimensionality reduction. The case of more than one latent factor is studied here. We compared the use of simple and weighted averages of standardised EBLUPs and multivariate EBLUPs, with the use of EBLUP and MEBLUP of factor scores. In this chapter, we show interesting results in data dimensionality reduction and SAE. This study demonstrates that if the correlation in the data is taken into account before small area estimates are computed (by using the latent variables and factor scores) then multivariate EBLUP does not provide large percentages of reduction in terms of MSE over the univariate EBLUP approach. Large gains in efficiency can be seen in weighted and simple averages of standardized MEBLUPs over the EBLUPs approach. The use of factor scores in data dimensionality reduction at the local level has an important advantage. This was showed in an application using the EU-SILC data that was used in Chapter 3. In fact, when a dashboard of single indicators needs to be studied, we may have variables measured on different scales; thus, multivariate SAE requires joint mixed effects models. These models have not been studied in this field.
In chapter 5 we proposed a parametric bootstrap algorithm for the mean squared error estimation of a multivariate EBLUP under a multivariate mixed-effects model and we answered research question 4. Our bootstrap procedure is developed following González-Manteiga et al. (2008a), where the univariate case is studied. We demonstrate the consistency of the MSE estimation by the imitation method as in González-Manteiga et al. (2008a). In their paper, the parametric bootstrap procedure has been proven to be consistent as an estimator of the true MSE in the univariate unit-level modelling setting. The reader may want to note that as we used ML estimators, the consistency of our proposed parametric bootstrap algorithm to the true MSE can be proven following the theorems developed in González-Manteiga et al. (2008a).

6.2 Future work and limitations

As it is highlighted in Pratesi (2016), local governments have a crucial role in poverty alleviation policies; indeed, they provide a large variety of social services. Therefore, they need accurate and reliable estimates of well-being indicators. Our work provided a variety of methodologies in the field of multidimensional well-being indicators measured at the local level. In order to support policy makers and Official Statistics, we paid particular attention to the estimation of MSE as well.

This work has a number of possible limitations. First, we proposed SAE methods in the unit-level approach. If the auxiliary variables are not available for the units, area-level
models can be used.

We did not address the issue of outliers in the distribution because we did not encounter this problem in our data. In the case of distributions with influential outliers, robust SAE methods can be considered. This is particularly true in the case of MSE estimation. Simulation studies on the performance of the EBLUP under outliers in $e_{di}$ and/or $u_d$ are available in the univariate SAE literature and are important in multivariate SAE as well. We refer to Sinha and Rao (2009) for the problems of outliers in EBLUP estimation and to González-Manteiga, et al. (2008a) for evaluations on bootstrap MSE approximations.

In this work we considered the case of continuous variables in multivariate EBLUP estimation. However, in well-being measurement there are many variables which are measured on a different scale, such as nominal scales. Therefore, it is important to focus attention on the case of non-continuous variables in multivariate SAE as well. We did not study the case of weighted and simple averages of standardized MEBLUP components where variables are measured on different scales. The use of joint multivariate mixed effects models may be interesting to evaluate here. Of course, the issue of MSE estimation will have to be addressed carefully. This is a topic for future research. In our work we overcame this issue by using factor analysis models for different types of response variables in Chapter 4 and factor scores are close to normally distributed.

Further investigations in our bootstrap algorithm, such as non-parametric bootstrap and
hybrid bootstrap are interesting and important methods to be evaluated in multivariate SAE. Also, the case of non-normal $e_{dl}$ and/or $u_d$ and heteroscedasiticy issues are topics we are currently aiming to study. Since our bootstrap was a first attempt in multivariate SAE under the model we considered, more research in this area is surely needed.

In this thesis we used one data source only. However, we are aware that well-being frameworks use various sources. Our methodology can be applied within the same data source since it is based on a unit-level SAE approach. In case of multiple sources, area-level SAE may be used as well as different weighting systems. This is a topic of future research. Of course, the statistical quality of the single indicators may differ; e.g. some indicators might have large variability. This problem, however, is not discussed in this thesis because it refers to the theoretical framework development stage. In case of indicators with a poor quality, further weighting systems can be developed in order to attenuate the impact of these in the composite estimates.

6.3 Final remarks

Local governments in charge of policies about poverty alleviation should base their decisions on precise and accurate estimates at the local level. Moreover, in the presence of multidimensional phenomena, large frameworks of well-being phenomena may be available. We believe that survey methodologists need to address this in developing SAE methods, as these frameworks are becoming more frequent and prominent in measuring well-being.
References


and practices for composite indicator development. EUR 20408 EN.


