Linköping Studies in Science and Technology.

Dissertations No. 1855

Contributions to Small Area Estimation

Using Random Effects Growth Curve Model

Innocent Ngaruye

Department of Mathematics, Division of Mathematical Statistics
Linköping University, SE-581 83 Linköping, Sweden

Linköping 2017

Linköping Studies in Science and Technology. Dissertations No. 1855

Contributions to SAE - Using Random Effects GC model

Copyright © Innocent Ngaruye, 2017

Division of Mathematical Statistics

Department of Mathematics

Linköping University

SE-581 83, Linköping, Sweden

`innocent.ngaruye@liu.se`

`www.liu.se/mai/ms`

Typeset by the author in L^AT_EX2e documentation system.

ISSN 0345-7524

ISBN 978-91-7685-516-4

Printed by LiU-Tryck, Linköping, Sweden 2017

Dedication

To my wife Christine,
My sons Bertin, Fabrice and Prince,
My daughters Angelique and Gentille:
For your love and support during my long stay away from home,
This thesis is affectionately dedicated.

Abstract

This dissertation considers Small Area Estimation with a main focus on estimation and prediction for repeated measures data. The demand of small area statistics is for both cross-sectional and repeated measures data. For instance, small area estimates for repeated measures data may be useful for public policy makers for different purposes such as funds allocation, new educational or health programs, etc, where decision makers might be interested in the trend of estimates for a specific characteristic of interest for a given category of the target population as a basis of their planning.

It has been shown that the multivariate approach for model-based methods in small area estimation may achieve substantial improvement over the usual univariate approach. In this work, we consider repeated surveys taken on the same subjects at different time points. The population from which a sample has been drawn is partitioned into several non-overlapping subpopulations and within all subpopulations there is the same number of group units. The aim is to propose a model that borrows strength across small areas and over time with a particular interest of growth profiles over time. The model accounts for repeated surveys, group individuals and random effects variations.

Firstly, a multivariate linear model for repeated measures data is formulated under small area estimation settings. The estimation of model parameters is discussed within a likelihood based approach, the prediction of random effects and the prediction of small area means across time points, per group units and for all time points are obtained. In particular, as an application of the proposed model, an empirical study is conducted to produce district level estimates of beans in Rwanda during agricultural seasons 2014 which comprise two varieties, bush beans and climbing beans.

Secondly, the thesis develops the properties of the proposed estimators and discusses the computation of their first and second moments. Through a method based on parametric bootstrap, these moments are used to estimate the mean-squared errors for the predicted small area means. Finally, a particular case of incomplete multivariate repeated measures data that follow a monotonic sample pattern for small area estimation is studied. By using a conditional likelihood based approach, the estimators of model parameters are derived. The prediction of random effects and predicted small area means are also produced.

Populärvetenskaplig sammanfattning

I den här avhandling diskuteras *small area estimation* (SAE) med fokus på skattningar av parametrarna samt prediktion för slumpvariabler givet en modell för upprepade mätningar. Efterfrågan på statistisk inferens för undergrupper av en population (*small area statistics*) har ökat markant. Till exempel kan sådan inferens för upprepade mätningar användas av offentliga beslutsfattare vid tilldelning av ekonomiska resurser, planering av nya utbildnings- eller hälsoprogram där det i vissa fall kan vara av intresse att studera en specifik undergrupp av befolkningen.

Det har visat sig att den multivariata framställningen för modellbaserade metoder vid SAE kan uppnå betydande bättre resultat jämfört med det vanliga endimensionella modellantagandet. I den här avhandlingen betraktar vi upprepade undersökningar på samma individer vid olika tidpunkter. Populationen från vilket ett urval har tagits är uppdelat i flera delpopulationer och inom alla undergrupper finns det samma antal observationer. Under dessa förutsättningar, är en multivariat linjär regressionsmodell formulerad. Syftet med den föreslagna modellen är att låna egenskaper mellan undergrupperna och över tid. I modellen ingår det upprepade undersökningar, grupper av individer samt slumpmässiga effekter.

Parametrarna i den föreslagna modellen skattas med hjälp av en likelihoodbaserad metod samt prediktion av slumpmässiga effekter och prediktion av de förväntade värdena i undergrupperna beräknas. Egenskaper, så som första och andra moment, för de likelihoodbaserade skattningarna härleds och de används via en parametrisk bootstrap för att uppskatta medelkvadratfelet för prediktionen av de förväntade värdena i undergrupperna. Vidare så tillämpar vi den föreslagna modellen på produktion av två bönsorter, *bush* - och *climbing beans*, i Rwanda på distriktsnivå och under två jordbrukssäsonger 2014.

Ofta så har man tyvärr ofullständig data med saknade observationer. I den här avhandlingen så diskuterar vi även hur man kan analysera sådan ofullständig data givet den föreslagna modellen för upprepade mätningar.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisors Dr Martin Singull and Professor Dietrich von Rosen. You have introduced me to the field of Multivariate Statistics. Thank you for your friendly advises, your patience, encouragement and guidance. I am grateful for your constructive discussions and valuable comments, I was lucky to work with you. Thank you.

I want to express my deep thanks to my assistant advisor Associate Professor Alexandre Lyambabaje, I greatly appreciate your excellent advises.

My sincere thanks go to Bengt Ove Turesson, Björn Textorius, Theresia, Theresa and Meaza, your assistance and support are highly appreciated. I would also like to thank current and former PhD students as well as other staff of the Department of Mathematics for friendly working environment.

Special thanks go to my wonderful family, my wife Christine Nyirabatware and our children. From the bottom of my heart, I thank you for your patience during my long absence from home.

Finally, I take this opportunity to acknowledge the Swedish International Development Cooperation Agency, SIDA, in collaboration with the University of Rwanda for financial support of my studies.

Innocent Ngaruye
Linköping, April 26, 2017

List of Papers

The thesis is based on the following four papers, which are referred to in the text by alphabetical order.

- A. I. Ngaruye**, J. Nzabanita, D. von Rosen and M. Singull, (2016a). Small Area Estimation under a Multivariate Linear Model for Repeated measures Data. *Communications in Statistics - Theory and Methods*. Accepted for publication.
<http://dx.doi.org/10.1080/03610926.2016.1248784>.
- B. I. Ngaruye**, D. von Rosen and M. Singull, (2016b). Crop yield estimation at district level for agricultural seasons 2014 in Rwanda. *African Journal of Applied Statistics*, 3(1):69-90.
- C. I. Ngaruye**, D. von Rosen and M. Singull, (2017). Mean-squared errors of small area estimators under a multivariate linear model for repeated measures data. *Linköping University Electronic Press*, LiTH-MAT-R-2017/05-SE.
- D. I. Ngaruye**, D. von Rosen and M. Singull, (2017). Small area estimation under a multivariate linear model for incomplete repeated measures data. *Linköping University Electronic Press*, LiTH-MAT-R-2017/06-SE.

Material of this thesis have been presented at the following international conferences

1. **The First Asian ISI Satellite Meeting on Small Area Estimation (SAE)**, Chulalongkorn University, Bangkok, Thailand, September 1-4th, 2013
2. **The International Conference on Trends and Perspectives (LinStat2014)**, Linköping University, Linköping, Sweden, August 24 - 28th, 2014
3. **The 25th International Workshop on Matrices and Statistics (IWMS'2016)**, Madeira, Portugal, June 6-9th, 2016

Contents

Dedication	i
Abstract	iii
Populärvetenskaplig sammanfattning	v
Acknowledgments	vii
List of Papers	ix
1 Introduction	1
1.1 Background and problem formulation	1
1.2 Example: Regional estimation of malnutrition	3
1.3 Aims of the thesis	4
1.4 Thesis outline	4
1.5 Summary of papers	4
1.5.1 Contributions of co-authors	6
2 Methods in Small Area Estimation	9
2.1 Design-based methods	9
2.1.1 Direct estimators	10
2.1.2 Indirect estimators	11
2.2 Model-based methods	12
2.2.1 Prediction in finite populations	12
2.2.2 Area level model	14
2.2.3 Unit level model	15
2.3 Estimation of Mean-squared errors of model-based estimators	16
2.3.1 Estimation of MSE for unit-level model	18
2.3.2 Estimation of MSE for area-level model	19
2.3.3 Estimation of MSE for unit and area level models by a parametric Bootstrap	19
2.4 Models used in this thesis	21
2.4.1 Growth Curve model	22
2.4.2 Random Effects Growth Curve models	24
2.4.3 The main model of the thesis and main results	26

3	Summary of contributions, concluding remarks and further research	35
3.1	Summary of contributions	35
3.2	Concluding remarks	36
3.3	Further research	37
	Bibliography	39

INCLUDED PAPERS

A.	Small Area Estimation under a Multivariate Linear Model for Repeated measures Data	45
1	Introduction	47
2	Model formulation	49
3	Decomposition of the model into sub-models	51
4	Estimation of the mean and covariance	53
5	Prediction of random effects and target small area means	57
6	Parametric Bootstrap of Mean Squared Error (MSE)	59
7	Simulation study	60
8	Concluding remarks	64
9	References	64
10	Appendix	67
B.	Crop yield estimation at district level for agricultural seasons 2014 in Rwanda	69
1	Introduction	72
2	Description of the data	73
	2.1 Seasonal Agricultural Survey (SAS)	74
	2.2 Statistics of covariates	75
3	Estimation of population mean and sampling variance	76
4	SAE with a multivariate linear model for repeated measures data	77
	4.1 Estimation of mean and covariance	79
	4.2 Prediction of random effects and target small area means	81
5	Main results	82
6	Discussion and assessment of results	83
7	Concluding remarks	83
8	Appendix	85
9	References	93

C. Mean-squared errors of small area estimators under a multivariate linear model for repeated measures data 95

1 Introduction 98

2 Description of the model 99

3 Estimation and prediction 100

3.1 Prediction of small area means 103

4 Preparations for moments calculation 104

5 Moments of proposed estimators 105

6 Mean-squared errors of predicted small area means 109

6.1 Derivation of $MSE(\hat{\tau}_{ig})$ 109

6.2 Estimation of $MSE(\hat{\tau}_{ig})$ 112

7 Appendix 113

8 References 113

D. Small area estimation under a multivariate linear model for incomplete repeated measures data 117

1 Introduction 119

2 Motivating example 121

3 Multivariate linear model for repeated measures data 121

4 Incomplete repeated measures data 122

4.1 Model formulation 123

5 Estimation of parameters and prediction of small area means 126

5.1 Estimation 126

5.2 Prediction 128

6 References 129

Introduction

THIS dissertation focuses on the problem of *Small Area Estimation* (SAE). Mainly, the problem is about how to produce reliable estimates of characteristics of interest (totals, means, proportions, quantiles, etc.) for small areas or domains based on few samples or even when no samples are taken from these areas. The second related problem is how to assess the estimation or prediction error.

Sampling surveys are preferred cost-effective ways of obtaining statistical information about a finite population rather than complete enumerations or censuses. These surveys are often originally designed to provide efficient estimates of parameters of interest for large populations and not for subpopulations or small domains and hence these domains are poorly represented in the sample. Thus, the surveys often provide very little information on a small area level and direct survey estimates on a target small area are not reliable due to a small sample size connected to this area.

The aim of this work is to develop methods within SAE with a focus put on estimation and prediction of small area characteristics of interest for repeated measures data. We are interested in small area means.

1.1 Background and problem formulation

The history of small area estimation goes back to eleventh century in England and seventeenth century in Canada (Brackstone, 1987; Marshall, 1991). Small area statistics have long been used for example in prevalence of disease mapping. In recent years, SAE methods have received much attention due to their usefulness in both public and private sectors and their demand has greatly increased worldwide (Rao, 2003). Both frequentist and Bayesian approaches as well as new developments in small area estimation have been investigated by different authors for example, Pfeffermann (2002, 2013), Rao (2003) and Chambers and Clark (2012). The demand for small area statistics has increased due to their use in formulating social and economic policies, allocation of government funds, regional planning,

business decision making etc. SAE has been used in a wide range of applications such as unemployment rates, poverty mapping, disease mapping, demography etc. For example, SAE has been used in health-related characteristics such as proportions of malnutrition, disability and mortality rates of diseases. SAE has also been used in agriculture to produce county estimates of crop production and other agricultural statistics. One may refer to Ghosh and Rao (1994); Rao (2003); Rao and Isabel (2015) for some examples and case studies in SAE.

Following the definition given by Rao (2003), the term “*small area*” or “*small domain*” is referred to a subpopulation for which the domain-specific sample is not large enough to produce direct estimates with reliable precision. This subpopulation can be a small geographical area (county, state, district, etc.), a demographic group within a geographical region (specific sex-age group, etc.) or any subdivision of the population. One possible solution to improve direct estimates is to “*borrow strength*” from other related data sets by using the data either from similar areas, or using relevant “*auxiliary information*” obtained from census or some other administrative records and thus increase the effective sample size.

Repeated measures data which refer to response outcomes taken on the same experimental unit at different time points have been widely used in research. The analysis of repeated measures data allows us to study trends over time. The demand for small area statistics is for both cross-sectional (i.e., a specific point in time) and for repeated measures data. For instance, small area estimates for repeated measures data may be useful for public policy makers for different purposes such as funds allocation, new educational or health programs, etc. In some cases, decision makers might be interested in the trend of estimates for a specific characteristic of interest for a given category of the target population as a basis of their planning.

The main focus of model-based methods in small area estimation which are methods based on the use of explicit linking models, that use linear mixed models has been put on cross-sectional data. There are also some studies dealing with small area estimation problems for longitudinal surveys which have been carried out by various authors, for example The EURAREA Consortium (2004), Nissinen (2009) and Singh and Sisodia (2011) who have used a design-based approach for repeated measures data in small area estimation which consists of using survey weights and base the inference on the probability distribution induced by associated sampling design. Singh and Sisodia (2011) have developed direct, synthetic and composite estimators for small area means for repeated measures data at a given time point when the population units contain non-overlapping groups. Datta et al. (1999) have shown that the multivariate approach for model-based methods in small area estimation may achieve substantial improvement over the usual univariate approach. Fuller and Harter (1987); Ghosh et al. (1991); Datta et al. (1999), among other authors, have worked on multivariate models in small area estimation such as the multivariate extensions of the

univariate Fay-Herriot model originally proposed by Fay and Herriot (1979) and nested linear regression model introduced by Battese et al. (1988). Recently, Benavent and Morales (2016) proposed multivariate Fay-Herriot models for estimating small area indicators. In their study, Benavent and Morales (2016) considered multiple response characteristics of interest for cross-sectional data and derived empirical best predictor of the vector of area means as well as estimators of mean-squared errors.

To the best of our knowledge, although multivariate models have been adopted in small area estimation, a random effects growth curve model useful to study the treatment mean and the treatment change over time has not yet been considered so far. In this work, we consider repeated surveys for the variable of interest on the same subjects at different time points. We assume that the target population from which the sample has been drawn is partitioned into several subpopulations and within all subpopulations, there is the same categories of units. We assume that the sampling design is not planned to estimate subpopulation level. Hence, direct estimates at this level are not of any high precision. The purpose with this work is to develop models which allow to provide reliable estimates for a given subpopulation which we call “small area” by prediction of target small area means.

1.2 Example: Regional estimation of malnutrition

In order to make the formulation of the problem clear, we give a realistic motivating example. In a country, say C, having 8 sub-national regions, surveys about child nutrition were carried out every 3 months during the whole year on the same subjects with intention of providing nutritional status of male and female children under 5 years and the factors affecting it. With a simple random sampling without replacement scheme, a random sample of children was drawn, and their height and weight were measured. A child is considered as stunting or underweight if she/he has a height or weight which is below specific reference values. Therefore, the variable of interest y is height or weight and direct estimates are available for sampled units. These surveys were not designed for small regions so that these direct estimates have large standard errors and hence are not reliable. We also assume that from census, some auxiliary information $\mathbf{x}$ are available for all units of the target population. The auxiliary information can be the household living conditions such as monthly expenditure of the household, employment of the head of the household. The aim is to develop the yearly proportions of malnutrition for each sub-region, firstly when all children remain in the sample for the whole year and secondly when there are some drop outs during the study.

1.3 Aims of the thesis

The aim of the thesis is to contribute to the development of new methods to be used in small area estimation statistics. Since the demand for small area statistics is not only for cross-sectional data, but it is also for repeated measures data, this thesis proposes a model which borrows strength both across areas and over time. The specific objectives of this work are

- (i) to formulate a mathematical model in the context of small area estimation for repeated measures data and propose a likelihood based inference approach used for estimation of unknown parameters, prediction of random effects and small area means;
- (iii) to apply the obtained results of the proposed model to a specific case study with relevant empirical data;
- (iv) to assess the precision of the proposed estimators through the estimation of the corresponding mean-squared errors;
- (v) to investigate the estimation and prediction methods for a particular case of incomplete data.

1.4 Thesis outline

This thesis consists of two parts. The first part comprises three chapters where we present necessary key concepts for further reading of the papers in the second part of the thesis. The first chapter is the Introduction which consists of the background, problem formulation, motivation example, aims of the thesis and summary of the papers. Chapter 2 gives a brief review of methods used in SAE with main focus on prediction of small area means and estimation of mean-squared errors and an overview of some multivariate linear models considered later in the thesis. Chapter 3 presents the summary contributions of the papers, concluding remarks and further research.

1.5 Summary of papers

This thesis consists of four papers, of which one has been published in an international peer reviewed journal, another has been accepted for publication and two others are available as technical reports. The summary of the papers is presented here below.

Paper A. Small Area Estimation under a Multivariate Linear Model for Repeated measures Data

Ngaruye, I., Nzabanita, J., von Rosen, D. and Singull, M. (2016). Small Area Estimation under a Multivariate Linear Model for Repeated measures Data. *Communications in Statistics - Theory and Methods*. Accepted for publication. <http://dx.doi.org/10.1080/03610926.2016.1248784>.

This paper develops a multivariate linear model for repeated measures data in small area estimation settings. A random effects growth curve model with covariates is formulated for repeated measures data on the response variable with the aim of estimating population characteristics of interest such as population totals or means at sub-population level with small sample sizes. The purpose of the considered model is to get a model which borrows strength both across small areas and over time. The model accounts for repeated surveys, grouped response units and random effects variations. Through a likelihood based approach, model parameters are estimated and random area effects are predicted. The prediction of small area means across time points and per group units are derived. A parametric bootstrap method is proposed for estimating the mean squared error of the predicted small area means and a simulation study is conducted.

Paper B. Crop yield estimation at District Level for Agricultural Seasons 2014 in Rwanda

Ngaruye, I., von Rosen, D., and Singull, M. (2016). Crop yield estimation at district level for agricultural seasons 2014 in Rwanda. *African Journal of Applied Statistics*, 3(1):69-90.

This paper discusses an application of Small Area Estimation techniques under the proposed multivariate linear regression model for repeated measures data in Paper A. The aim is to produce district level estimates of crop yield for beans which comprise two varieties, bush beans and climbing beans in Rwanda during agricultural seasons 2014 by using the micro data of National Institute of Statistics of Rwanda (NISR) obtained from the Seasonal Agricultural Survey (SAS) 2014 combined with the data from the Crop Assessment Survey 2013. The techniques of estimation provided to handle the proposed model are applied to the data and efficient estimates of beans yield for both climbing and bush beans at district level are derived. Statistical diagnostics are carried out to compare direct survey estimates and model-based estimates which show that model-based estimates produced at district level are reliable and representative of the corresponding districts. A cluster map showing the

distribution of average beans yield model-based estimates during season A and B, year 2014 per district is produced.

Paper C. Mean-squared errors of small area estimators under a multivariate linear model for repeated measures data

Ngaruye, I., von Rosen, D., and Singull, M. (2017). Mean-squared errors of small area estimators under a multivariate linear model for repeated measures data. *Linköping University Electronic Press*, LiTH-MAT-R-2017/05-SE.

In this paper, mean-squared errors (MSE) of predicted small area estimators under the multivariate linear model for repeated measures data developed in Paper A are considered. Firstly, the paper develops the properties of proposed estimators, discusses the computation of the moments for the obtained estimators and an approximation of the MSE for the predicted small area means. The parametric bootstrap approach is used for bias correction and for prediction error that reflects the uncertainty when the unknown covariance is replaced by its suitable estimator.

Paper D. Small area estimation under a multivariate linear model: model selection and incomplete repeated measures data

Ngaruye, I., von Rosen, D., and Singull, M. (2017). Small area estimation under a multivariate linear model: model selection and incomplete repeated measures data. *Linköping University Electronic Press*, LiTH-MAT-R-2017/06-SE.

The last paper addresses the issue of small area estimation analysis for multivariate repeated measures data that follow a monotonic missing sample data pattern. Random effects growth curve models with covariates for both complete and incomplete data are formulated. The estimation and prediction of model parameters are discussed within a conditional likelihood based approach. Further, the prediction of random effects and predicted small area means are also discussed. The proposed techniques may be useful for small area estimation under longitudinal surveys with grouped response units and drop outs.

1.5.1 Contributions of co-authors

I would like to claim that in all aforementioned works, I have contributed by doing all detailed calculations, deriving the results, conducting coding and simulations and writing

the papers. For paper A, the research problem was not my proper idea as it was for the rest of the 3 papers. The papers were the product of many discussions between me and co-authors.

Methods in Small Area Estimation

METHODS used in Small Area Estimation are mainly divided into “design-based” and “model-based” methods. For the first category, the inference is fully based on the used sampling design and for the latter methods, also called “model-dependent”, the inference is involved with statistical methods based on the frequentist or Bayesian approaches or the combination of the two (Pfeffermann 2002; Rahman 2008). Both methods use auxiliary information to “borrow strength” from related neighboring areas, from censuses, surveys or registers. Throughout this chapter, we suppose that the sampling design for the population as a whole does not correspond well to the divisions in the population. Therefore, the samples sizes for subpopulations may become small which leads to the small area estimation problem.

2.1 Design-based methods

Design-based estimation or the *randomization* approach in SAE is based on traditional probability sampling theory. One can refer to Fuller (2009) for probability sampling theory in statistics. Under this estimation approach, the randomness is only induced by the sampling design used to select the sample with population measurements regarded as fixed (e.g., Lehtonen and Veijanen 2009; Pfeffermann 2002). These methods make use of survey weights and associated statistical inferences are based on the sample selection probabilities. Two types of estimators are derived according to the related data sources used; either direct estimators when only domain-specific data are used or indirect estimators if the estimation procedure “borrows strength” from related areas. A common feature of design-based estimators is that there is no explicit model assumptions used for their derivation, and the variance and bias are calculated under the randomization distribution.

2.1.1 Direct estimators

The direct estimators of small area quantities under a design-based approach is often based on estimation of population characteristics of interest such as population means/totals in classical probability theory by using for example Horvitz-Thompson form of direct estimator (see Cochran, 1977; Särndal et al., 1992). The direct estimator is obtained by using only the sample data from the corresponding domain.

Let U be a finite population of size N consisting on m disjoint subpopulations called domains, small areas or simply areas U_i each with population size N_i , ($i = 1, \dots, m$) such that $U = \bigcup_{i=1}^m U_i$ and $N = \sum_{i=1}^m N_i$.

We assume that a sample s of size n is selected from the whole population U with selection probabilities π_j of unit j and sampling weights $w_j = \frac{1}{\pi_j}$; $j \in s$.

Let Y denote the characteristic of interest, y_{ij} the outcome value of the j^{th} population unit coming from the small area i (with $i = 1, \dots, m$; $j = 1, \dots, N_i$) and s_i be the corresponding sub-sample of size n_i taken from the small area i such that $s = s_1 \cup \dots \cup s_m$ and $n = \sum_{i=1}^m n_i$.

Then, if there is no auxiliary information available, the true area mean $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$ can be estimated using Hajek type of direct estimator

$$\hat{\bar{Y}}_i^{\text{Hajek}} = \frac{\sum_{j=1}^{n_i} w_j}{\sum_{j=1}^{n_i} w_j y_{ij}},$$

or by Horvitz-Thompson type of direct estimator if the population size N_i is known

$$\hat{\bar{Y}}_i^{HT} = \frac{1}{N_i} \sum_{j=1}^{n_i} w_j y_{ij}.$$

As an example, consider simple random sampling without replacement. Then the selection probability equals $\pi_{ij} = \frac{n_i}{N_i}$ and the sampling weight is $w_{ij} = \frac{N_i}{n_i}$. A direct estimator of means for small area i is given by

$$\hat{\bar{Y}}_i = \frac{1}{N_i} \sum_{j=1}^{n_i} w_{ij} y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_i.$$

An unbiased variance of this estimator is given by

$$\text{Var}[\hat{\bar{Y}}_i] = (1 - f_i) \frac{S_i^2}{n_i}, \quad f_i = \frac{n_i}{N_i},$$

where $(1 - f_i)$ is the finite population correction factor and

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2, \quad N_i \geq 2.$$

Note that for unknown N_i , f_i is replaced by $f = n/N$ and the corresponding unbiased variance estimator is given by

$$\text{Var}[\widehat{Y}_i] = (1 - f) \frac{s_i^2}{n_i},$$

where $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$, $n_i \geq 2$. It follows that for small n_i , the variance will be larger unless the variability of the y -values is sufficiently small. In order to reduce the variance, we suppose in addition that we have r auxiliary variables known for every sample units. Denote by $\mathbf{x}_{ij}$, the r -vector of covariables for the j^{th} unit in area i and $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ the corresponding sample mean. We assume that the population mean $\bar{\mathbf{X}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ are also known. Then, a more efficient design-based estimator is the *regression estimator* given by

$$\bar{Y}_i^{reg} = \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \boldsymbol{\beta}_i, \quad (2.1)$$

which variance equals

$$\text{Var}[\bar{Y}_i^{reg}] = (1 - \rho_i^2)(1 - f_i) \frac{S_i^2}{n_i} = (1 - \rho_i^2) \text{Var}[\widehat{Y}_i],$$

where $\boldsymbol{\beta}_i$ and ρ_i are respectively the vector of regression coefficients and the multiple correlation between the survey variable Y and auxiliary variables $\mathbf{x}_{ij}$ in area i .

We note that the use of auxiliary information reduces the variance by the factor $(1 - \rho_i^2)$, the reason why it increases the prediction power in SAE. However, in practice the regression coefficient $\boldsymbol{\beta}_i$ is replaced by its ordinary least squares estimator from the sample s_i which may not be effective because of a small sample size.

2.1.2 Indirect estimators

In order to increase the level of precision of small area estimates, practitioners often use so called “indirect” estimators that “borrow strength” by using observation values from related areas and/or from the other time periods and thus increase the effective sample size.

By assuming that all small areas are similar with respect to the quantity to be estimated, an estimator obtained from a large domain covering several small areas can be used to derive an indirect estimator for separate areas comprising that domain (Gonzalez, 1973). This type of estimator is called *regression synthetic estimator*. It is equivalent to put an assumption on $\boldsymbol{\beta}_i$ and on intercepts $(\bar{Y}_i - \bar{\mathbf{X}}_i' \boldsymbol{\beta}_i)$ to be similar across the small areas. An effective regression synthetic estimator is obtained by

$$\bar{Y}_{syn,i}^{reg} = \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \widehat{\boldsymbol{\beta}}, \quad (2.2)$$

where $\hat{\beta}$ is the full sample estimator computed using data from all samples s_i . The difference between the direct regression estimator (2.1) and the synthetic regression estimator (2.2) is that the vectors of regression coefficients β_i and β are calculated using the data from the target area i and using the data from entire areas, respectively. The synthetic estimator is more efficient if the assumption of homogeneity within the larger domain holds. However, even though it reduces the variance, it can lead to severe biases if there is a strong individual effect on the regression coefficient.

The bias reduction is then improved by the use of a composite estimator which is a weighted sum of the area direct estimator. Let θ_i be the small area characteristic of interest, the composite estimator has the general form

$$\hat{\theta}_{com,i} = \phi_i \hat{\theta}_i + (1 - \phi_i) \hat{\theta}_{syn,i},$$

where $\hat{\theta}_i$ is the direct estimator, $\hat{\theta}_{syn,i}$ is the synthetic estimator and $\phi_i = f_i - n_i/N_i$ is a suitable weight chosen to minimize mean-squared errors (MSE).

A composite estimator has small or no bias but large variance. The more weight is given to the direct unbiased estimator as the sampling fraction f_i increases. However, the sampling fractions are usually very small and thus the use of this weight in practice implies the use of the synthetic estimator.

2.2 Model-based methods

In recent years, model-based approaches in SAE have received much attention due to their usefulness for estimating small area characteristics. For example model-based methods are now being extensively used to find indirect estimates for small area means. As stated previously, even though the design-based estimators are simple to implement and can provide efficient estimators which do not depend on an assumed model, they can lead to severe biases if the assumption of homogeneity within larger domains is violated. Furthermore, they are relying on assumption of similarity of all areas with respect to the variable of interest and do not account for area specific variability. The model-based methods are based on the use of explicitly linking of a model for incorporating random area effects to overcome these underlying problems. However, model-based methods can lead to severe bias if the assumed model is not correct.

2.2.1 Prediction in finite populations

Model-based sampling theory applied to finite population treats the problem of estimating finite population characteristics of interest as a prediction problem. In this approach, the

finite population is treated itself as a random sample from a larger population (Bolfarine and Zacks, 1992). The population under study is assumed to be a random realization from a superpopulation characterized by a suitable model. By using this model, the predictive distribution of the values for non sampled units given the realized values of sampled units is obtained which constitutes the basis for drawing inference on any function of the values of the characteristic of interest. For example, the estimation of a finite population mean from the sample returns to the prediction of a mean of non-sampled values. The observed population vector $\mathbf{y} = (y_1, \dots, y_N)'$ is considered as a realization of a random variable characterized by a model, say φ . Under the superpopulation approach, the finite quantities θ to be estimated are random and then we do not estimate them, but predict them.

By denoting s and r the sampled and remainder of the finite population U so that $U = s \cup r$ and $\mathbf{y} = (\mathbf{y}'_s, \mathbf{y}'_r)'$, we can express the quantities of interest $\theta = \mathbf{l}'\mathbf{y}$ by

$$\theta = \theta_s + \theta_r = \mathbf{l}'_s \mathbf{y}_s + \mathbf{l}'_r \mathbf{y}_r,$$

where $\mathbf{l} = (\mathbf{l}'_s, \mathbf{l}'_r)'$. Therefore, the predictor $\hat{\theta}$ of θ is given by

$$\hat{\theta} = \theta_s + \hat{\theta}_r,$$

where $\hat{\theta}_r$ is the predictor of θ_r .

The predictor $\hat{\theta}$ is φ -unbiased if

$$E_\varphi[\hat{\theta} - \theta] = 0, \text{ for the given model } \varphi,$$

where $E_\varphi[\cdot]$ denotes the expectation with respect to the superpopulation model. The MSE of $\hat{\theta}$ is given by

$$E_\varphi[\hat{\theta} - \theta]^2 = \text{Var}_\varphi[\hat{\theta} - \theta] + \left(E_\varphi[\hat{\theta} - \theta]\right)^2,$$

and if $\hat{\theta}$ is unbiased with respect to φ , then

$$\text{Var}_\varphi[\hat{\theta} - \theta] = E_\varphi[\hat{\theta} - \theta]^2,$$

where $\text{Var}_\varphi[\cdot]$ stands for the variance with respect to the superpopulation model.

Model-based methods in SAE largely use linear mixed models involving suitable auxiliary variables in the fixed part of the model and random area-specific effects. These area-specific effects account for between-variations in data that are not explained by auxiliary variables. Commonly used models in SAE are area-level models and unit-level models (Rao, 2003). In the following two subsections, we present a general estimation and prediction overview about area-level and unit-level models and keep the same population settings as defined in the beginning of this chapter.

2.2.2 Area level model

The area level model is a widely used small area model originally presented by Fay and Herriot (1979) for prediction of mean per capital income in small geographical areas within counties in United States. It is used when the values $\mathbf{x}_i$ for auxiliary variables are only available at the area level. This model consists of two models, one is the sampling model comprising the direct estimates and the sampling error, the other one is the linking model relating the population value to some auxiliary variables with unknown random area effects for small area $i = 1, \dots, m$. Suppose that we are interested in prediction of unknown small area means of variable of interest. Denote by $\tilde{\theta}_i$ the survey estimates of true small area means μ_i . The sampling model and linking model are respectively defined as

$$\begin{aligned}\tilde{\theta}_i &= \mu_i + e_i, \\ \mu_i &= \mathbf{x}_i' \boldsymbol{\beta} + u_i, \quad i = 1, \dots, m,\end{aligned}\tag{2.3}$$

where e_i are sampling errors of direct estimator assumed to be independent and normally distributed with $\mathbb{E}(e_i | \mu_i) = 0$ and known $\text{Var}(e_i | \mu_i) = \sigma_{e_i}^2$. The u_i 's are uncorrelated area-specific random effects assumed to be iid normal with $\mathbb{E}[u_i] = 0$ and $\text{Var}(u_i) = \sigma_u^2$. Further, u_i and e_i are assumed to be independent. Combining the two models in (2.3) leads to the area level linear mixed model

$$\tilde{\theta}_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i + e_i, \quad i = 1, \dots, m.\tag{2.4}$$

From the prediction of random effects in linear mixed model under normality, the *Best Linear Unbiased Predictor* (BLUP) of u_i is given by

$$\begin{aligned}\hat{u}_i &= \mathbb{E}[u_i | \tilde{\theta}_i] = \mathbb{E}[u_i] + \text{Cov}[\tilde{\theta}_i, u_i] (\text{Cov}[\tilde{\theta}_i])^{-1} (\tilde{\theta}_i - \mathbb{E}[\tilde{\theta}_i]) \\ &= \frac{\sigma_u^2}{\sigma_{e_i}^2 + \sigma_u^2} (\tilde{\theta}_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}).\end{aligned}$$

Put $\gamma_i = \frac{\sigma_u^2}{\sigma_{e_i}^2 + \sigma_u^2}$ and $V_i = \sigma_u^2 + \sigma_{e_i}^2$, for known $\sigma_{e_i}^2$ and σ_u^2 , the BLUP for μ_i is then given by

$$\hat{\mu}_i^{BLUP} = \mathbf{x}_i' \tilde{\boldsymbol{\beta}} + \hat{u}_i = \mathbf{x}_i' \tilde{\boldsymbol{\beta}} + \gamma_i (\tilde{\theta}_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}) = \gamma_i \tilde{\theta}_i + (1 - \gamma_i) \mathbf{x}_i' \tilde{\boldsymbol{\beta}},$$

where $\tilde{\boldsymbol{\beta}}$ is the *Generalized Least Squares* (GLS) estimator of $\boldsymbol{\beta}$ given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^m V_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m V_i^{-1} \mathbf{x}_i \tilde{\theta}_i \right) = \left(\sum_{i=1}^m \gamma_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \gamma_i \mathbf{x}_i \tilde{\theta}_i \right).$$

The factor γ_i is called *shrinkage factor*. In practice the variance σ_u^2 is unknown and replaced by its sample estimates $\hat{\sigma}_u^2$ which yields the corresponding *Empirical Best Linear Unbiased*

Predictor (EBLUP). The $\hat{\sigma}_u^2$ can be estimated by the method of moments, *Maximum Likelihood* (ML) or *Restricted Maximum Likelihood* (RML) and the EBLUP for μ_i is then given by

$$\hat{\mu}_i^{EBLUP} = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{u}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\tilde{\theta}_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}), \quad (2.5)$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^m \hat{V}_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \hat{V}_i^{-1} \mathbf{x}_i \tilde{\theta}_i \right) = \left(\sum_{i=1}^m \hat{\gamma}_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \hat{\gamma}_i \mathbf{x}_i \tilde{\theta}_i \right), \\ \hat{\gamma}_i &= \frac{\hat{\sigma}_u^2}{\sigma_{ei}^2 + \sigma_u^2}, \quad \hat{V}_i = \hat{\sigma}_u^2 + \sigma_{ei}^2. \end{aligned}$$

2.2.3 Unit level model

Unit level Model also called "*Nested error unit regression model*" which was originally proposed by Battese et al. (1988) is used when the values $\mathbf{x}_{ij}$ of auxiliary variables are known for every unit j in the sample from small area i and the true population area means $\bar{\mathbf{X}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ are also known. We also assume that all areas are sampled. The model is defined as

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + u_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, N_i, \quad (2.6)$$

where y_{ij} denote the values of variable of interest Y ; the random effect u_i are the effect of area characteristic that are not accounted for by the auxiliary variables and ε_{ij} are the sampling errors. The two error terms u_i and ε_{ij} are assumed to be mutually independent with zero means and variances σ_u^2 and σ_e^2 , respectively. Normality of u_i and ε_{ij} is often assumed to hold.

Under this model, for known σ_e^2 and σ_u^2 , the BLUP for the small area means $\bar{Y}_i$ is expressed as

$$\begin{aligned} \hat{\mu}_i^{BLUP} &= \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \gamma_i (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}) \\ &= (1 - \gamma_i) \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \gamma_i \left[\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \hat{\boldsymbol{\beta}} \right], \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the GLS estimator of $\boldsymbol{\beta}$, $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_i}$ and $\bar{y}_i, \bar{\mathbf{x}}_i$ are the sample means of y_{ij} and $\mathbf{x}_{ij}$ for area i , respectively. The BLUP of u_i is similarly obtained following the prediction of random effects in mixed linear models (Henderson, 1975). The n_i is the sample size selected

from N_i units in the i -th area. For unknown σ_e^2 and σ_u^2 , they are replaced by their estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ which leads to the EBLUP of small area means $\bar{Y}_i$ given by

$$\hat{\mu}_i^{EBLUP} = \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}), \quad (2.7)$$

where

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}.$$

It should be pointed out that for small areas with no samples ($n_i = 0$), the shrinkage factor tends to zero and the the EBLUP of small area means are reduced to $\hat{\mu}_i = \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}$. Furthermore, when the sampling fraction $f_i = \frac{n_i}{N_i}$ is non-negligible, Battese et al. (1988) showed that the EBLUP for the means in small area i is given by

$$\hat{\mu}_i^{EBLUP} = \frac{1}{N_i} \left\{ \sum_{s_i} y_{ij} + \sum_{r_i} (\bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \hat{u}_i) \right\} = f_i + (1 - f_i) \left\{ \bar{\mathbf{X}}_{ri}' \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}) \right\},$$

where s_i and r_i denote the sampled and non-sampled units, respectively while

$$\bar{\mathbf{X}}_{ri}' = \frac{1}{N_i - n_i} (N_i \bar{\mathbf{X}}_i - n_i \bar{\mathbf{x}}_i)$$

is the mean of auxiliary values for $(N_i - n_i)$ non-sampled units from small area i .

2.3 Estimation of Mean-squared errors of model-based estimators

As stated in the Introduction, the SAE problem is twofold. Firstly, the problem is how to produce population characteristics of interest of subpopulations for which reduced sample sizes connected to these subpopulations lead to design-based estimators without high precision. Secondly, it is the assessment of the prediction error. One of the methods used to ensure the precision of model-based estimators is the assessment of their mean-squared errors (MSE). For comprehensive review of MSE in SAE, one can refer to Kackar and Harville (1984); Prasad and Rao (1990); Das et al. (2004); Datta et al. (2005) among others.

In this section we investigate the methods of estimation for MSE in mixed linear models and present the theory of estimation of MSE for area level and unit level models and the estimation of MSE by a parametric bootstrap.

Consider the linear mixed model defined by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ is an N -vector of response variable, $\mathbf{X} : N \times p$ and $\mathbf{Z} : N \times M$ are known design matrices for fixed effects and random effects, respectively, $\boldsymbol{\beta}$ is a p -vector of unknown regression coefficients, $\mathbf{u}$ is an M -vector of the random effects and $\mathbf{e}$ is an N -vector of the random errors. Moreover, $\mathbf{u}$ and $\mathbf{e}$ are mutually independently distributed as $\mathbf{u} \sim \mathcal{N}_M(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim \mathcal{N}_N(\mathbf{0}, \mathbf{R})$, respectively. It follows that $\mathbf{y} \sim \mathcal{N}_N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\sigma}))$, where $\boldsymbol{\Sigma}(\boldsymbol{\sigma}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_q)$ denotes a q -dimensional vector of unknown parameters.

For known variance component $\boldsymbol{\sigma}$, Henderson (1975) has developed the derivation of the BLUP for the linear combination of $\boldsymbol{\beta}$ and realized $\mathbf{u}$, say $\mu = \mathbf{l}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u}$ which is given by

$$\mu(\boldsymbol{\sigma}) = \mathbf{l}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\tilde{\mathbf{u}} = \mathbf{l}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}),$$

where

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \tilde{\boldsymbol{\beta}}(\boldsymbol{\sigma}) = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}, \\ \tilde{\mathbf{u}} &= \tilde{\mathbf{u}}(\boldsymbol{\sigma}) = \mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}).\end{aligned}$$

The MSE of $\mu(\boldsymbol{\sigma})$ is given by

$$MSE[\mu(\boldsymbol{\sigma})] = \mathbb{E}[\{\mu(\boldsymbol{\sigma}) - \mu\}^2] = \begin{bmatrix} \mathbf{l}' & \mathbf{m}' \end{bmatrix} \text{cov} \begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} \begin{bmatrix} \mathbf{l} \\ \mathbf{m} \end{bmatrix} = g_1(\boldsymbol{\sigma}) + g_2(\boldsymbol{\sigma}),$$

where

$$\begin{aligned}g_1(\boldsymbol{\sigma}) &= \mathbf{m}'(\mathbf{G} - \mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{G})\mathbf{m} \\ g_2(\boldsymbol{\sigma}) &= (\mathbf{l} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m})'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{l} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}).\end{aligned}$$

Since in practice the variance component vector $\boldsymbol{\sigma}$ in the BLUP estimator $\mu(\boldsymbol{\sigma})$ is unknown, it is replaced by its estimator $\hat{\boldsymbol{\sigma}}$ which yields the EBLUP

$$\mu(\hat{\boldsymbol{\sigma}}) = \mathbf{l}'\hat{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

It should be pointed out that the naive estimator of $MSE[\mu(\hat{\boldsymbol{\sigma}})]$ given by

$$\widehat{MSE}[\mu(\hat{\boldsymbol{\sigma}})] = g_1(\hat{\boldsymbol{\sigma}}) + g_2(\hat{\boldsymbol{\sigma}})$$

underestimates the $MSE[\mu(\hat{\boldsymbol{\sigma}})]$ since it ignores the error of the estimator associated with $\hat{\boldsymbol{\sigma}}$.

Kackar and Harville (1984) showed that if $\hat{\boldsymbol{\sigma}}$ is translation invariant, that is if $\hat{\boldsymbol{\sigma}}(\mathbf{y}) = \hat{\boldsymbol{\sigma}}(-\mathbf{y})$ and $\hat{\boldsymbol{\sigma}}(\mathbf{y} + \mathbf{X}\boldsymbol{\beta}) = \hat{\boldsymbol{\sigma}}(\mathbf{y})$, for all $\mathbf{y}$ and $\boldsymbol{\beta}$, the $MSE[\mu(\hat{\boldsymbol{\sigma}})]$ can be expressed by

$$MSE[\mu(\hat{\boldsymbol{\sigma}})] = MSE[\mu(\boldsymbol{\sigma})] + \mathbb{E}[\mu(\hat{\boldsymbol{\sigma}}) - \mu(\boldsymbol{\sigma})]^2.$$

Note that Kackar and Harville (1981) proved that maximum likelihood and restricted maximum likelihood estimators are translation invariant.

2.3.1 Estimation of MSE for unit-level model

Consider the nested error unit regression model (2.6)

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, N_i,$$

and the corresponding EBLUP for small area means given by (2.7)

$$\widehat{\mu}_i^{EBLUP} = \overline{\mathbf{X}}_i' \widehat{\boldsymbol{\beta}} + \widehat{\gamma}_i(\bar{y}_i - \overline{\mathbf{x}}_i' \widehat{\boldsymbol{\beta}}).$$

Let $\boldsymbol{\sigma} = (\sigma_u^2, \sigma_e^2)'$ and denote by Σ_{uu} , Σ_{ee} the asymptotic variances of the estimators $\widehat{\sigma}_u^2$, $\widehat{\sigma}_e^2$, respectively and by Σ_{ue} the asymptotic covariance between $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$. Prasad and Rao (1990) obtained the second order approximation of $MSE[\widehat{\mu}_i]$ which is split into three components

$$MSE[\widehat{\mu}_i] = g_{1i}(\boldsymbol{\sigma}) + g_{2i}(\boldsymbol{\sigma}) + g_{3i}(\boldsymbol{\sigma}), \quad (2.8)$$

where

$$\begin{aligned} g_{1i}(\boldsymbol{\sigma}) &= (1 - \gamma_i)\sigma_u^2 = \gamma_i\sigma_e^2/n_i, \\ g_{2i}(\boldsymbol{\sigma}) &= (\overline{\mathbf{X}}_i - \gamma_i\overline{\mathbf{x}}_i)' \left(\sum_{i=1}^m V_i \right)^{-1} (\overline{\mathbf{X}}_i - \gamma_i\overline{\mathbf{x}}_i), \\ g_{3i}(\boldsymbol{\sigma}) &= n_i(n_i\sigma_u^2 + \sigma_e^2)^{-3} h(\boldsymbol{\sigma}), \end{aligned}$$

with

$$\begin{aligned} V_i &= \sigma_u^{-2} \sum_{j=1}^{n_i} (\mathbf{x}_{ij}\mathbf{x}'_{ij} - \gamma_i\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i'), \\ h(\boldsymbol{\sigma}) &= \sigma_e^4 \Sigma_{uu} + \sigma_u^4 \Sigma_{ee} - 2\sigma_u^2 \sigma_e^2 \Sigma_{ue}. \end{aligned}$$

The three components above $g_{1i}(\boldsymbol{\sigma})$, $g_{2i}(\boldsymbol{\sigma})$ and $g_{3i}(\boldsymbol{\sigma})$ measure the uncertainty when all parameters are known, the uncertainty about estimation of fixed effects $\boldsymbol{\beta}$ and the uncertainty due to the estimation of γ_i , respectively. Prasad and Rao (1990) showed that the term $g_{3i}(\boldsymbol{\sigma})$ is of order $o(m^{-1})$ provided that m is large while the leading term are $g_{1i}(\boldsymbol{\sigma})$ and $g_{2i}(\boldsymbol{\sigma})$ which are of order $o(1)$ and $o(m^{-1})$, respectively. Prasad and Rao (1990) also proposed an estimator of $MSE[\widehat{\mu}_i]$ expressed by

$$\widehat{MSE}[\widehat{\mu}_i] = g_{1i}(\widehat{\boldsymbol{\sigma}}) + g_{2i}(\widehat{\boldsymbol{\sigma}}) + 2g_{3i}(\widehat{\boldsymbol{\sigma}}),$$

where $g_{1i}(\widehat{\boldsymbol{\sigma}})$, $g_{2i}(\widehat{\boldsymbol{\sigma}})$ and $g_{3i}(\widehat{\boldsymbol{\sigma}})$ are the same as $g_{1i}(\boldsymbol{\sigma})$, $g_{2i}(\boldsymbol{\sigma})$ and $g_{3i}(\boldsymbol{\sigma})$ given by expression (2.8) with σ_u^2 and σ_e^2 replaced by their estimators $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$.

2.3.2 Estimation of MSE for area-level model

Consider the area level model (2.4)

$$\tilde{\theta}_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i + e_i, \quad i = 1, \dots, m,$$

and the corresponding EBLUP for small area means given in (2.5) by

$$\hat{\mu}_i^{EBLUP} = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{u}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\tilde{\theta}_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}).$$

In a similar way identical to the unit-level model, following Prasad and Rao (1990) the second order approximation of $MSE[\hat{\mu}_i]$ can be expressed by

$$MSE[\hat{\mu}_i] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2), \quad (2.9)$$

where

$$\begin{aligned} g_{1i}(\sigma_u^2) &= \sigma_u^2 \sigma_{ei}^2 (\sigma_u^2 + \sigma_{ei}^2)^{-1}, \\ g_{2i}(\sigma_u^2) &= \sigma_{ei}^4 (\sigma_u^2 + \sigma_{ei}^2)^{-2} \mathbf{x}'_i \left(\sum_{i=1}^m (\sigma_{ei}^2 + \sigma_u^2)^{-1} \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \mathbf{x}_i, \\ g_{3i}(\sigma_u^2) &= \sigma_{ei}^4 (\sigma_u^2 + \sigma_{ei}^2)^{-3} \Sigma_{uu}. \end{aligned}$$

Note that the derivation of the asymptotic variance Σ_{uu} depends on the estimation method used. For different expressions of $g_{3i}(\sigma_u^2)$ and the relation between them, one can refer for example to Datta et al. (2005). An estimator of $MSE[\hat{\mu}_i]$ is then given by

$$\widehat{MSE}[\hat{\mu}_i] = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2),$$

where σ_u^2 is replaced by its estimator $\hat{\sigma}_u^2$.

2.3.3 Estimation of MSE for unit and area level models by a parametric Bootstrap

In some cases, small area methods do not have an analytic expressions of MSE for point estimates. Particularly, the closed forms of $g_{3i}(\boldsymbol{\sigma})$ and $g_{3i}(\sigma_u^2)$ presented in (2.8) and (2.9) can not be obtained for some cases and require approximation. When explicit exact formula of mean-squared errors cannot be obtained, an alternative approach used by practitioners in SAE that avoids Taylor linearizations and further approximations is re-sampling methods. Some of the methods of re-sampling are parametric and non-parametric bootstrap approaches. It should be pointed out that re-sampling techniques are nowadays competing

with asymptotic analytic approximations in SAE for estimation of MSE (González-Manteiga et al., 2008b). The parametric bootstrap which is a re-sampling based method consists of assessing the variability between generated replicates obtained by re-sampling and re-fitting the model to each replicate sample. For further references on this topic one can see Hall and Maiti (2006); Lahiri et al. (2007); González-Manteiga et al. (2008a); Pereira and Coelho (2010); Jiongo et al. (2013).

Following González-Manteiga et al. (2008a), the following steps describe the algorithm used for estimating the MSE under unit-level model defined in (2.6) by

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, N_i.$$

Step 1 From the sample data, calculate estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ for $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 , respectively;

Step 2 Generate bootstrap random area effects u_i^* , $i = 1, \dots, m$ from $u_i^* \sim \mathcal{N}_p(0, \hat{\sigma}_u^2)$;

Step 3 Generate normal error term ε_{ij}^* from $\varepsilon_{ij}^* \sim \mathcal{N}_p(0, \hat{\sigma}_e^2)$ independently of u_i^* ;

Step 4 Construct the bootstrap population from the model

$$y_{ij}^* = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + u_i^* + \varepsilon_{ij}^*, \quad i = 1, \dots, m, \quad j = 1, \dots, N_i,$$

where auxiliary values $\mathbf{x}_{ij}$ are known for each unit in the population;

Step 5 From the generated bootstrap population from *Step 4*, calculate the true population means for each small area i , $i = 1, \dots, m$ by $\hat{\mu}_i = \overline{X}'_i\hat{\boldsymbol{\beta}} + u_i^*$;

Step 6 Fit the model from *Step 4* to bootstrap sample, calculate estimators $\hat{\boldsymbol{\beta}}^*$, $\hat{\sigma}_u^{2*}$ and $\hat{\sigma}_e^{2*}$ and calculate the EBLUP $\hat{\mu}_i^* = \overline{X}'_i\hat{\boldsymbol{\beta}}^* + \hat{u}_i^*$;

Step 7 Repeat steps (2-6) R times (say, 1000 times) obtaining $\hat{\mu}_i(r)$ and $\hat{\mu}_i^*(r)$ for each r replicate, $r = 1, \dots, R$;

Step 8 The bootstrap estimator of MSE is then given by

$$MSE(\hat{\mu}_i^*) = \frac{1}{R} \sum_{r=1}^R \left(\hat{\mu}_i^*(r) - \hat{\mu}_i(r) \right)^2.$$

Similarly, the algorithm used for estimating the MSE under area level model defined in (2.4)

$$\tilde{\theta}_i = \mathbf{x}'_i\boldsymbol{\beta} + u_i + e_i, \quad i = 1, \dots, m,$$

is described by the steps bellow

Step 1 Using the sample data, calculate estimators $\widehat{\beta}$, and $\widehat{\sigma}_u^2$ for β and σ_u^2 , respectively;

Step 2 Generate bootstrap random area effects u_i^* , $i = 1, \dots, m$ from $u_i^* \sim \mathcal{N}_p(0, \widehat{\sigma}_u^2)$;

Step 3 Independently of u_i^* , generate sampling errors e_i^* from $e_i^* \sim \mathcal{N}_p(0, \sigma_{ei}^2)$, $i = 1, \dots, m$;

Step 4 Generate bootstrap true population area means through the bootstrap linking model

$$\mu_i^* = \mathbf{x}_i' \widehat{\beta} + u_i^*, \quad i = 1, \dots, m;$$

Step 5 Generate bootstrap direct estimators through the bootstrap sampling model

$$\widetilde{\theta}_i^* = \mu_i^* + e_i^*, \quad i = 1, \dots, m;$$

Step 6 Fit the model to bootstrap data $\{\widetilde{\theta}_i^*, \mathbf{x}_i\}$, calculate estimators $\widehat{\beta}^*$, $\widehat{\sigma}_u^{2*}$ and the EBLUP

$$\widehat{\mu}_i^* = \mathbf{x}_i' \widehat{\beta}^* + \widehat{u}_i^* = \mathbf{x}_i' \widehat{\beta}^* + \widehat{\gamma}_i^* (\widetilde{\theta}_i^* - \mathbf{x}_i' \widehat{\beta}^*),$$

$$\text{where } \widehat{\gamma}_i^* = \widehat{\sigma}_u^{2*} (\widehat{\sigma}_u^{2*} + \sigma_{ei}^2)^{-1};$$

Step 7 Repeat steps (2-6) R times (say, 1000 times) obtaining $\mu_i^*(r)$ and $\widehat{\mu}_i^*(r)$ for each r replicate, $r = 1, \dots, R$;

Step 8 The bootstrap estimator of MSE is then given by

$$MSE(\widehat{\mu}_i^*) = \frac{1}{R} \sum_{r=1}^R \left(\widehat{\mu}_i^*(r) - \mu_i^*(r) \right)^2.$$

2.4 Models used in this thesis

In this section, we discuss briefly some multivariate linear models used in the thesis. The aim is to present definitions, some properties and main results that are referred to throughout this thesis to facilitate a further reading of the thesis, especially the reading of the included papers.

Multivariate normal distribution of a random vector can be generalized to matrix variate normal distribution when sampling from multivariate normal population. Hereafter, we give the definition of matrix normal distribution and a set of models that have been used or are connected to the models used in the thesis.

Definition 2.4.1 (Matrix normal distribution). Let $\Sigma : p \times p$ and $\Psi : n \times n$ be such that $\Sigma = \Gamma\Gamma'$ and $\Psi = \Upsilon\Upsilon'$. A random matrix $Y : p \times n$ is said to be matrix normally distributed with parameters $M : p \times n$, Σ and Ψ , we write $Y \sim \mathcal{N}_{p,n}(M, \Sigma, \Psi)$ if

$$Y \stackrel{d}{=} M + \Gamma X \Upsilon',$$

where $X : p \times n$ consists of pn i.i.d. standard normal distributions x_{ij} , $i = 1, 2, \dots, p$, $j = 1, 2, \dots, n$. If Σ and Ψ are positive definite, then the corresponding density function is given by

$$f(Y) = (2\pi)^{-\frac{pn}{2}} |\Sigma|^{-\frac{n}{2}} |\Psi|^{-\frac{p}{2}} e^{-\frac{1}{2} \text{tr}\{\Sigma^{-1}(Y-M)\Psi^{-1}(Y-M)'\}}.$$

It follows that $Y \sim \mathcal{N}_{p,n}(M, \Sigma, \Psi)$ if and only if

$$\text{vec}Y \sim \mathcal{N}_{pn}(\text{vec}M, \Psi \otimes \Sigma).$$

As result

$$\mathbb{E}[Y] = M, \quad \mathbb{D}[Y] = \mathbb{D}[\text{vec}Y] = \Psi \otimes \Sigma,$$

where $\text{vec}(\cdot)$ stands for the vectorization operator and $\otimes$ is the usual Kronecker product. The notation “ $\stackrel{d}{=}$ ” means “equality in distribution”, $|\cdot|$ and tr denote the determinant and the trace of a matrix, respectively.

It is important to note that if

$$X_1 \sim \mathcal{N}_{p,n}(M_1, \Sigma_1, \Psi_1),$$

and

$$X_2 \sim \mathcal{N}_{p,n}(M_2, \Sigma_2, \Psi_2),$$

then $X_1 + X_2$ is normally distributed but not necessarily matrix normally distributed. This is because usually it does not hold true that there exist matrices Υ and Φ such that

$$\Psi_1 \otimes \Sigma_1 + \Psi_2 \otimes \Sigma_2 = \Upsilon \otimes \Phi.$$

See Kollo and von Rosen (2005) for more details.

2.4.1 Growth Curve model

In the following, we use the notation A° for any matrix of full rank spanning $\mathcal{C}(A)^\perp$, i.e., $\mathcal{C}(A^\circ) = \mathcal{C}(A)^\perp$, where $\mathcal{C}(A)$ denotes the column vector space generated by the columns of

the matrix $\mathbf{A}$ and $\mathcal{C}(\mathbf{A})^\perp$ is its orthogonal complement. Moreover, $\mathbf{A}^-$ denotes an arbitrary generalized inverse of the matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$. We also denote by $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ and $\mathbf{Q}_\mathbf{A} = \mathbf{I} - \mathbf{P}_\mathbf{A}$ the orthogonal projection matrices onto the column space $\mathcal{C}(\mathbf{A})$ and onto its orthogonal complement $\mathcal{C}(\mathbf{A})^\perp$, respectively. For positive definite matrix $\mathbf{S}$, we have projections which are denoted by $\mathbf{P}_{\mathbf{A},\mathbf{S}} = \mathbf{A}(\mathbf{A}'\mathbf{S}\mathbf{A})^-\mathbf{A}'\mathbf{S}$ and $\mathbf{Q}_{\mathbf{A},\mathbf{S}} = \mathbf{I} - \mathbf{P}_{\mathbf{A},\mathbf{S}}$.

The classical Multivariate Analysis of Variance (MANOVA) model is defined by

$$\mathbf{Y} = \mathbf{B}\mathbf{C} + \mathbf{E},$$

where $\mathbf{Y} : p \times n$, $\mathbf{B} : p \times k$, $\mathbf{C} : k \times n$ are the observation, unknown parameter and design matrices, respectively. It is assumed that $\mathbf{E} \sim \mathcal{N}_{p,n}(\mathbf{0}, \mathbf{\Sigma}_e, \mathbf{I})$ with $\mathbf{\Sigma}_e$ supposed to be positive definite and $\mathcal{N}_{p,n}(\mathbf{0}, \mathbf{\Sigma}_e, \mathbf{I})$ denotes the matrix normal distribution where the columns of $\mathbf{E}$ are independent and the dispersion of each column equals $\mathbf{\Sigma}_e$. When $n \geq \text{rank}(\mathbf{C}) + p$ and $\mathbf{C}$ assumed to have full rank, the MLEs for $\mathbf{B}$ and $\mathbf{\Sigma}_e$ are given by

$$\begin{aligned}\hat{\mathbf{B}} &= \mathbf{Y}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}, \\ \hat{\mathbf{\Sigma}}_e &= \frac{1}{n}\mathbf{Y}\mathbf{Q}_{\mathbf{C}'}\mathbf{Y}'.\end{aligned}$$

The *Growth Curve* (GC) model is an extension of the MANOVA model and also known as *Generalized Multivariate Analysis of Variance* (GMANOVA) model or *Multivariate Linear Normal Model with mean ABC* (written as *MLNM(ABC)*). The GC model is given by

$$\mathbf{Y} = \mathbf{A}\mathbf{B}\mathbf{C} + \mathbf{E}, \quad (2.10)$$

where $\mathbf{Y} : p \times n$ is the observation matrix, $\mathbf{B} : q \times k$ is the parameter matrix, $\mathbf{A} : p \times q$ is the within individual design matrix indicating the time dependency within the individuals, $\mathbf{C} : k \times n$ with $\text{rank}(\mathbf{C}) + p \leq n$ is the between individual design matrix accounting for group effects and $\mathbf{E} : p \times n$ is the error matrix with $\mathbf{E} \sim \mathcal{N}_{p,n}(\mathbf{0}, \mathbf{\Sigma}_e, \mathbf{I})$. The estimation, testing and model diagnostic problems for GC model and its extensions have been investigated by several authors, see for example, Potthoff and Roy (1964); von Rosen (1989); Khatri (1973); Rao (1958); Nzabanita et al. (2012); Ohlson and Srivastava (2010).

The maximum likelihood estimators of $\mathbf{B}$ and $\mathbf{\Sigma}_e$ are respectively given by

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{Y}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}, \\ n\hat{\mathbf{\Sigma}}_e &= (\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{C})(\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{C})',\end{aligned}$$

where

$$\mathbf{S} = \mathbf{Y}\mathbf{Q}_{\mathbf{C}'}\mathbf{Y}'$$

and the design matrices $\mathbf{A}$ and $\mathbf{C}$ are assumed to have full rank. More details can be found for example in Kollo and von Rosen (2005).

The classical GC model defined in (2.10) relies on the assumption of the same profile of different individuals. If this does not hold, a more general model, sometimes called *sum of profiles* can be introduced. That is an *Extended Growth Curve* (EGC) model. The following EGC model is used in the research paper A and is an extension of the GC model discussed by Filipiak and von Rosen (2012) where also the MLEs can be found.

The Extended Growth Curve Model with nested subspace condition $\mathcal{C}(\mathbf{A}_i) \subseteq \mathcal{C}(\mathbf{A}_{i-1})$, $i = 2, 3, \dots, m$, is given by

$$\mathbf{Y} = \sum_{i=1}^m \mathbf{A}_i \mathbf{B}_i \mathbf{C}_i + \mathbf{E},$$

where $\mathbf{Y} : p \times n$ is a data matrix, $\mathbf{A}_i : p \times q_i$, $\mathbf{C}_i : k_i \times n$ with $\text{rank}(\mathbf{C}_1) + p \leq n$, $i = 1, 2, \dots, m$ are design matrices, $\mathbf{B}_i : q_i \times k_i$ are unknown parameters and $\mathbf{E} \sim \mathcal{N}_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}_e, \mathbf{I})$.

In summary,

$$\mathbf{Y} \sim N_{p,n} \left(\sum_{i=1}^m \mathbf{A}_i \mathbf{B}_i \mathbf{C}_i, \boldsymbol{\Sigma}_e, \mathbf{I}_n \right),$$

and when $m = 1$, the model reduces to the classical Growth Curve model.

2.4.2 Random Effects Growth Curve models

There are two versions of Random Effects Growth Curve models. First, we present some results for the most well known case which includes the random coefficient regression model. The beauty of this model is that we obtain unweighted maximum likelihood estimators of the mean parameter. Then after, we introduce another kind of Random Effects Growth Curve model, which is used in the work of the thesis, but unfortunately there does not exist explicit maximum likelihood estimators.

Consider the Growth Curve model given in equation (2.10) and suppose that there exist additional individual random effects $\mathbf{U} : r \times n$. The first version of Random Effects Growth Curve model is defined as

$$\begin{aligned} \mathbf{Y} &= \mathbf{A}\mathbf{B}\mathbf{C} + \mathbf{Z}\mathbf{U} + \mathbf{E}, \\ \mathbf{E} &\sim \mathcal{N}_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}_e, \mathbf{I}_n), \\ \mathbf{U} &\sim \mathcal{N}_{r,n}(\mathbf{0}, \boldsymbol{\Sigma}_u, \mathbf{I}_n), \end{aligned} \tag{2.11}$$

where $\mathbf{Z} : p \times r$ is a known design matrix for random effects. It follows that

$$\mathbf{Y} \sim \mathcal{N}_{p,n}(\mathbf{ABC}, \mathbf{\Sigma}, \mathbf{I}_n),$$

where $\mathbf{\Sigma} = \mathbf{Z}\mathbf{\Sigma}_u\mathbf{Z}' + \mathbf{\Sigma}_e$. This class of models has been considered by many authors and estimation of parameters of interest has been discussed with different choices of $\mathbf{A}$ and $\mathbf{Z}$. See for example Nummi (1997); Ip et al. (2007); Lange and Laird (1989); Yokoyama and Fujikoshi (1993); Yokoyama (2001); Srivastava and Singull (2012) for more details. In what follows, we consider a particular case where $\mathbf{\Sigma}_e$ has simple structure, i.e., $\mathbf{\Sigma}_e = \sigma_e^2 \mathbf{I}_p$ and present the estimation and prediction framework in model (2.11).

For a special case when $\mathbf{Z} = \mathbf{A}$ ($r = q$), model (2.11) is called *Random Coefficient Regression* (RCR) model. We also have $\mathbf{A} = [\mathbf{Z} : \mathbf{Z}_r]$, where $\mathbf{Z}_r$ is a $p \times (q - r)$. In both cases, the ML estimator is identical to Ordinary Least Squares (OLS) estimator of parameter matrix $\mathbf{B}$ (Reinsel, 1982; Azzalini, 1987; Lange and Laird, 1989; Nummi, 1997)

$$\hat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1},$$

for full rank matrices $\mathbf{A}$ and $\mathbf{C}$ and $0 \leq r \leq q \leq p < n$. Note that for general design matrices $\mathbf{A}$ and $\mathbf{Z}$, the explicit MLEs of $\mathbf{B}$ usually do not exist (Ip et al., 2007). For the case when $\mathbf{A} = [\mathbf{Z} : \mathbf{Z}_r]$, the estimators of variance σ_e^2 and covariance matrix $\mathbf{\Sigma}_u$ can be expressed by

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{1}{n(p-r)} \text{tr}\{\mathbf{Q}_Z \mathbf{W}\}, \\ \hat{\mathbf{\Sigma}}_u &= \frac{1}{n} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} - \hat{\sigma}_e^2 (\mathbf{Z}'\mathbf{Z})^{-1},\end{aligned}$$

where

$$\mathbf{W} = (\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{C})(\mathbf{Y} - \mathbf{A}\hat{\mathbf{B}}\mathbf{C})'.$$

For the RCR model, the estimators of variance σ_e^2 and covariance matrix $\mathbf{\Sigma}_u$ can be expressed as

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{1}{n(p-q)} \text{tr}\{\mathbf{Y}'\mathbf{Q}_A \mathbf{Y}\}, \\ \hat{\mathbf{\Sigma}}_u &= \frac{1}{n} (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'\mathbf{S}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} - \hat{\sigma}_e^2 (\mathbf{A}'\mathbf{A})^{-1},\end{aligned}$$

where $\mathbf{S} = \mathbf{Y}\mathbf{Q}_C\mathbf{Y}'$. More details can be found for example in (Nummi, 1997; Pan et al., 2002).

For the second version of Random Effects Growth Curve model, we consider additional individual random effects $\mathbf{U} : p \times m$ to the the classical Growth Curve model given in

equation (2.10) and we formulate a Random Effects Growth Curve model as follows

$$\begin{aligned} \mathbf{Y} &= \mathbf{ABC} + \mathbf{UZ} + \mathbf{E}, \\ \mathbf{E} &\sim \mathcal{N}_{p,n}(\mathbf{0}, \mathbf{\Sigma}_e, \mathbf{I}_n), \\ \mathbf{U} &\sim \mathcal{N}_{p,m}(\mathbf{0}, \mathbf{\Sigma}_u, \mathbf{I}_m), \end{aligned} \tag{2.12}$$

where $\mathbf{Z} : m \times n$ is a known design matrix for random effects. It follows that

$$\text{vec} \mathbf{Y} \sim \mathcal{N}_{pn}(\text{vec}(\mathbf{ABC}), \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = \mathbf{Z}'\mathbf{Z} \otimes \mathbf{\Sigma}_u + \mathbf{I}_n \otimes \mathbf{\Sigma}_e$. For estimation and prediction of model parameters in model (2.12), we may refer to Ngaruye et al. (2016).

The main difference of the two versions of Random Effects Growth Curve models is that the first type of model (2.11) has the property of matrix normal distribution since the corresponding covariance matrix is separable while the second type (2.12) does not.

The prediction of random effects in model (2.11) and (2.12) is performed using Henderson's prediction theory of BLUP (Henderson, 1973; Robinson, 1991). This approach consists of maximizing the joint density $f(\mathbf{Y}, \mathbf{U})$ with respect to $\mathbf{U}$ assuming the variance σ_e^2 and covariance matrix $\mathbf{\Sigma}_u$ to be known. Hereafter we present the predictor of random effects in model (2.11), the predictor of random effects in the main model of this thesis which is similar to model (2.12) with some considerations is discussed in all included papers.

The predictor in model (2.11) is then given by

$$\hat{\mathbf{U}} = \mathbf{\Sigma}_u \mathbf{Z}' \mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{ABC}),$$

where $\hat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'\mathbf{Y}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}$. Usually, the dispersion components are unknown and replaced by their suitable estimators which yields

$$\hat{\mathbf{U}} = \hat{\mathbf{\Sigma}}_u \mathbf{Z}' \hat{\mathbf{\Sigma}}^{-1} (\mathbf{Y} - \mathbf{ABC}),$$

where

$$\hat{\mathbf{\Sigma}} = \mathbf{Z} \hat{\mathbf{\Sigma}}_u \mathbf{Z}' + \hat{\sigma}_e^2 \mathbf{I}_p.$$

2.4.3 The main model of the thesis and main results

The main model considered in the thesis is a Random Effects Growth Curve model with covariates formulated in the context of SAE. The objective is to borrow strength across

both small areas and over time by incorporating simultaneously the effects of areas and time. This model allows for finding the small area means at each time point, per group units and particularly the pattern of changes or mean growth profiles over time.

Ngaruye et al. (2016) consider repeated measurements on variable of interest y for p time points, $t_1, \dots, t_p$ from the finite population U of size N partitioned into m disjoint sub-populations or domains $U_1, \dots, U_m$ called *small areas* of sizes $N_i, i = 1, \dots, m$, such that $\sum_{i=1}^m N_i = N$. It is also assumed that in every area, there are k different groups of units of size N_{ig} for group g such that $\sum_i \sum_{g=1}^k N_{ig} = N$ and supposed that there are auxiliary data $\mathbf{x}_{ij}$ of r variables available for each population unit j in all m small areas. The working model which combines all small areas is expressed by

$$\begin{aligned} \mathbf{Y} &= \mathbf{ABHC} + \mathbf{1}_p \gamma' \mathbf{X} + \mathbf{UZ} + \mathbf{E}, \\ \mathbf{E} &\sim \mathcal{N}_{p,N}(\mathbf{0}, \Sigma_e, \mathbf{I}_N), \quad \mathbf{U} \sim \mathcal{N}_{p,m}(\mathbf{0}, \Sigma_u, \mathbf{I}_m), \\ \text{vec}(\mathbf{Y}) &\sim \mathcal{N}_{pN}(\text{vec}(\mathbf{ABHC} + \mathbf{1}_p \gamma' \mathbf{X}), \mathbf{Z}' \mathbf{Z} \otimes \Sigma_u + \mathbf{I}_N \otimes \Sigma_e), \end{aligned} \quad (2.13)$$

where $\mathbf{H} = [\mathbf{I}_r : \mathbf{I}_r : \dots : \mathbf{I}_r]$, $\Sigma_e = \sigma_e^2 \mathbf{I}$ and σ_e^2 is assumed to be known. The matrix $\mathbf{H}$ is included in the model for technical issues of estimation.

The estimation and prediction are performed within a likelihood based approach through orthogonal transformation of model (2.13). This transformation is based on orthogonal diagonalization of the idempotent matrix $(\mathbf{CC}')^{-1/2} \mathbf{CZ}' \mathbf{ZC}' (\mathbf{CC}')^{-1/2}$ by $\mathbf{\Gamma} = [\mathbf{\Gamma}_1 : \mathbf{\Gamma}_2]$, the orthogonal matrix of eigenvectors for m and $N-m$ elements. The orthogonal transformation leads to the partition of model (2.13) into three independent models

$$\begin{aligned} \mathbf{YR}_1 &\sim \mathcal{N}_{p,m}(\mathbf{M}_1, \Sigma_u + \Sigma_e, \mathbf{I}_m), \\ \mathbf{YR}_2 &\sim \mathcal{N}_{p,mk-m}(\mathbf{M}_2, \Sigma_e, \mathbf{I}_{mk-m}), \\ \mathbf{YC}'^o &\sim \mathcal{N}_{p,N-mk}(\mathbf{M}_3, \Sigma_e, \mathbf{I}_{N-mk}), \end{aligned}$$

where

$$\begin{aligned} \mathbf{M}_1 &= \mathbf{ABK}_1 + \mathbf{1}_p \gamma' \mathbf{XR}_1, \\ \mathbf{M}_2 &= \mathbf{ABK}_2 + \mathbf{1}_p \gamma' \mathbf{XR}_2, \\ \mathbf{M}_3 &= \mathbf{1}_p \gamma' \mathbf{XC}'^o, \\ \mathbf{K}_1 &= \mathbf{H}(\mathbf{CC}')^{1/2} \mathbf{\Gamma}_1, \quad \mathbf{K}_2 = \mathbf{H}(\mathbf{CC}')^{1/2} \mathbf{\Gamma}_2, \\ \mathbf{R}_1 &= \mathbf{C}'(\mathbf{CC}')^{-1/2} \mathbf{\Gamma}_1, \quad \mathbf{R}_2 = \mathbf{C}'(\mathbf{CC}')^{-1/2} \mathbf{\Gamma}_2. \end{aligned}$$

Then, the likelihood based estimators and the predictor of random effects for a general case and for a particular case when the matrices $\mathbf{X}$ and $\mathbf{K}_2^{o'} \mathbf{K}_1$ are of full rank and $p = q$ are presented in the following two theorems.

Theorem 1 (Ngaruye et al. (2016)). Consider the multivariate linear model for repeated measures data on the response variable of interest given in (2.13). The estimators of γ , B and Σ_u and the predictor of random effects U can be expressed by

$$\begin{aligned}\hat{\gamma} &= \frac{1}{p}(\mathbf{P}\mathbf{X}')^{-1}\mathbf{P}\mathbf{Y}'\mathbf{1}_p + (\mathbf{P}\mathbf{X}')^o\hat{\mathbf{t}}_2, \\ \hat{B} &= (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}\mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-} \\ &\quad - \frac{1}{p}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{1}_p\mathbf{1}_p'\mathbf{Y}\mathbf{P}'(\mathbf{X}\mathbf{P}')^{-}\mathbf{X}\mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-} \\ &\quad - (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{1}_p\hat{\mathbf{t}}_2'(\mathbf{P}\mathbf{X}')^{o'}\mathbf{X}\mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-} + \hat{\mathbf{T}}_1(\mathbf{K}_2\mathbf{K}_2')^{o'}, \\ \hat{\Sigma}_u &= \frac{1}{m}(\mathbf{V}_3 - \mathbf{A}\hat{\mathbf{T}}_1\tilde{\mathbf{C}}_1 - \mathbf{1}_p\hat{\mathbf{t}}_2'\tilde{\mathbf{C}}_2)(\mathbf{V}_3 - \mathbf{A}\hat{\mathbf{T}}_1\tilde{\mathbf{C}}_1 - \mathbf{1}_p\hat{\mathbf{t}}_2'\tilde{\mathbf{C}}_2)' - \Sigma_e, \\ \hat{U} &= \left(\Sigma_e\hat{\Sigma}_u^{-1} + \mathbf{I}_p\right)^{-1}(\mathbf{Y} - \mathbf{A}\hat{B}\mathbf{H}\mathbf{C} - \mathbf{1}_p\hat{\gamma}'\mathbf{X})\mathbf{Z}',\end{aligned}$$

where

$$\begin{aligned}\mathbf{P} &= \mathbf{X}\mathbf{C}''^o(\mathbf{C}''^o)' + \mathbf{X}\mathbf{R}_2\mathbf{R}_2' - \mathbf{X}\mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-}\mathbf{K}_2\mathbf{R}_2' \\ \hat{\mathbf{T}}_1 &= (\mathbf{A}'\mathbf{S}_2^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}_2^{-1}(\mathbf{V}_3 - \mathbf{1}_p\hat{\mathbf{t}}_2'\tilde{\mathbf{C}}_2)\tilde{\mathbf{C}}_1'(\tilde{\mathbf{C}}_1\tilde{\mathbf{C}}_1')^{-} + \mathbf{A}'\mathbf{T}_{11}\tilde{\mathbf{C}}_1^o, \\ \hat{\mathbf{t}}_2 &= (\mathbf{1}_p'\mathbf{S}_1^{-1}\mathbf{1}_p)^{-1}(\tilde{\mathbf{C}}_2\mathbf{Q}_{\tilde{\mathbf{C}}_1'}\tilde{\mathbf{C}}_2')^{-}\tilde{\mathbf{C}}_2\mathbf{Q}_{\tilde{\mathbf{C}}_1'}\mathbf{V}_3'\mathbf{S}_1^{-1}\mathbf{1}_p + (\tilde{\mathbf{C}}_2\mathbf{Q}_{\tilde{\mathbf{C}}_1'})^{o'}\mathbf{t}_{21}'\mathbf{1}_p, \\ \mathbf{S}_1 &= \mathbf{V}_3\mathbf{Q}_{(\tilde{\mathbf{C}}_1';\tilde{\mathbf{C}}_2')}\mathbf{V}_3', \\ \mathbf{S}_2 &= \mathbf{S}_1 + \mathbf{Q}_{\mathbf{1}_p,\mathbf{S}_1^{-1}}\mathbf{V}_3\mathbf{P}_{\mathbf{Q}_{\tilde{\mathbf{C}}_1'}\tilde{\mathbf{C}}_2'}\mathbf{V}_3'\mathbf{Q}_{\mathbf{1}_p,\mathbf{S}_1^{-1}}', \\ \tilde{\mathbf{C}}_1 &= (\mathbf{K}_2\mathbf{K}_2')^{o'}\mathbf{K}_1, \quad \tilde{\mathbf{C}}_2 = (\mathbf{P}\mathbf{X}')^{o'}\mathbf{X}(\mathbf{I} - \mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-})\mathbf{R}_1.\end{aligned}$$

Theorem 2 (Ngaruye et al. (2016)). Consider the model (2.13) and assume that the matrices $\mathbf{X}$ and $\mathbf{K}_2^{o'}\mathbf{K}_1$ are of full rank. For a particular case when $p = q$, the within design matrix $\mathbf{A}_{p \times p}$ is non singular, then the estimators for γ , Σ_u , the linear combination BHC and the predictor of U are given by

$$\begin{aligned}\hat{\gamma} &= \frac{1}{p}(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\mathbf{P}\mathbf{Y}'\mathbf{1}_p, \\ \hat{B}\mathbf{H}\mathbf{C} &= \mathbf{A}^{-1}\left(\mathbf{Y} - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\mathbf{Y}\mathbf{P}\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\right)\mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-}\mathbf{H}\mathbf{C} + \mathbf{A}^{-1}\mathbf{V}_3\mathbf{K}_1'\mathbf{P}_1\mathbf{H}\mathbf{C}, \\ \hat{\Sigma}_u &= \frac{1}{m}\mathbf{V}_3\mathbf{Q}_{\mathbf{K}_1'\mathbf{K}_2'}\mathbf{V}_3' - \Sigma_e \text{ assumed to be positive definite,} \\ \hat{U} &= (\Sigma_e + \hat{\Sigma}_u)^{-1}\hat{\Sigma}_u\mathbf{V}_3\mathbf{Q}_{\mathbf{K}_1'\mathbf{K}_2'}\mathbf{R}_1'\mathbf{Z}',\end{aligned}$$

where

$$\begin{aligned}\mathbf{P} &= \mathbf{C}''^o(\mathbf{C}''^o)' + \mathbf{R}_2\mathbf{Q}_{\mathbf{K}_2'}\mathbf{R}_2', \quad \mathbf{P}_1 = \mathbf{K}_2^o(\mathbf{K}_2^{o'}\mathbf{K}_1\mathbf{K}_1'\mathbf{K}_2^o)^{-1}\mathbf{K}_2^{o'}, \\ \mathbf{V}_3 &= \left(\mathbf{Y} - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\mathbf{Y}\mathbf{P}\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\right)\mathbf{G}_2\mathbf{R}_1, \\ \mathbf{G}_2 &= \mathbf{I} - \mathbf{R}_2\mathbf{K}_2'(\mathbf{K}_2\mathbf{K}_2')^{-}\mathbf{H}\mathbf{C}.\end{aligned}$$

The prediction of small area means is performed under the superpopulation approach to finite population in the sense that estimating the small area means is equivalent to predicting small area means of non sampled values, given the sample data and auxiliary data. The model in (2.13) is partitioned into two parts $\mathbf{Y}_i^{(s)}$ and $\mathbf{Y}_i^{(r)}$ of sampled and non sampled units and similar partitions apply for matrices $\mathbf{X}$ and $\mathbf{Z}$. Given the multivariate linear regression model defined in (2.13), the target small area means at each time point for each group and across all time points are given by

$$\hat{\boldsymbol{\mu}}_{ig} = \frac{1}{N_{ig}} \left(\mathbf{Y}_{ig}^{(s)} \mathbf{1}_{n_{ig}} + \left(\mathbf{A} \hat{\boldsymbol{\beta}}_g \mathbf{1}'_{N_{ig}-n_{ig}} + \mathbf{1}_p \hat{\boldsymbol{\gamma}}' \mathbf{X}_{ig}^{(r)} + \hat{\mathbf{u}}_i \mathbf{z}_{ig}^{(r)'} \right) \mathbf{1}_{N_{ig}-n_{ig}} \right), \quad (2.14)$$

$$\hat{\mu}_{ig} = \frac{1}{p} \mathbf{1}_p' \hat{\boldsymbol{\mu}}_{ig}, \quad g = 1, \dots, k, \quad t = 1, \dots, p. \quad (2.15)$$

Ngaruye et al. (2016) have proposed a parametric bootstrap method for estimating the MSE of the predicted small area means through the following steps

- 1) From the sample data, calculate estimators $\hat{\mathbf{B}}, \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Sigma}}_u$ for $\mathbf{B}, \boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}_u$, respectively, as well as predictor $\hat{\mathbf{U}}$;
- 2) Generate independent area-specific random effects $\mathbf{u}_i^*, i = 1, \dots, m$ from $\mathbf{u}_i^* \sim \mathcal{N}_p(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_u)$;
- 3) Generate bootstrap population $\{\mathbf{Y}_i^*\} : p \times N_i$ from the model

$$\mathbf{Y}_i^* = \mathbf{A} \hat{\mathbf{B}} \mathbf{C}_i + \mathbf{1}_p \hat{\boldsymbol{\gamma}}' \mathbf{X}_i + \mathbf{u}_i^* \mathbf{z}_i' + \mathbf{E}_i, \quad i = 1, \dots, m;$$

- 4) Compute the target small area means $\boldsymbol{\mu}_{ig}^*$ for the bootstrap population from step 3;
- 5) Fit the model to bootstrap sample $\{\mathbf{Y}_i^*\} : p \times n_i$, calculate $\hat{\mathbf{B}}^*, \hat{\boldsymbol{\gamma}}^*, \hat{\boldsymbol{\Sigma}}_u^*$ and derive the corresponding bootstrap area means $\hat{\boldsymbol{\mu}}_{ig}^*$ using these new estimators;
- 6) Repeat steps 2) – 5) R times and compute the estimated MSE by

$$\widehat{MSE}(\hat{\boldsymbol{\mu}}_{ig}) = \frac{1}{R} \sum_{r=1}^R \left(\hat{\boldsymbol{\mu}}_{ig}^*(r) - \boldsymbol{\mu}_{ig}^*(r) \right) \left(\hat{\boldsymbol{\mu}}_{ig}^*(r) - \boldsymbol{\mu}_{ig}^*(r) \right)'.$$

For a particular case when $p = q$, where the within individuals design matrix is non singular, the model has been applied to Seasonal Agricultural Survey (SAS), 2014 in Rwanda to produce district level estimates of crop yield for beans which comprise two varieties, bush beans and climbing beans in Rwanda during agricultural seasons 2014.

Ngaruye et al. (2016) considered the SAS 2014 for two seasons A and B carried out in Rwanda, a country located in central Africa and divided into 30 administrative districts, with the intention to provide accurate and reliable agricultural statistics in terms of land use, crop

By using crop yield estimates for the agricultural season 2013, proportion of usage of organic and inorganic fertilizers, proportion for usage of irrigation and anti erosion practices as relevant covariates, the beans yield estimates at district level for both varieties and both seasons were produced. The assessment of reliability of the obtained model-based estimates were conducted via model checking of underling assumptions and bias diagnostics. The following cluster map given in Figure 2.1 shows the distribution of average beans yield model-based estimates during agricultural seasons A and B, year 2014 per district.

Map of Rwanda: Bean Production Yield and Variety Distribution

Inset Map: Rwanda in Africa (Scale: 1:1 200 000)

Main Map: District boundaries and Lake Kivu are shown. The map displays bean production yield per district (Average Kg/ha) and the distribution of four bean varieties per season.

Legend:

- District boundary
- Lake
- Yield per district (Average Kg/ha):
 - 900 - 1000
 - 1100 - 1270
 - 1300 - 1500
- Bean varieties per season:
 - Bush beans A
 - Bush beans B
 - Climbing beans A
 - Climbing beans B

Districts and Approximate Yield (Kg/ha):

District	Approximate Yield (Kg/ha)
Nyagatare	900 - 1000
Musanze	900 - 1000
Butera	900 - 1000
Gicumbi	900 - 1000
Gatsibo	900 - 1000
Kayanza	900 - 1000
Kirehe	900 - 1000
Nigoma	900 - 1000
Bugesera	900 - 1000
Ruhango	900 - 1000
Nyanza	900 - 1000
Huye	900 - 1000
Gisagara	900 - 1000
Nyaruguru	900 - 1000
Rutsiro	1100 - 1270
Gakeme	1100 - 1270
Rulindo	1100 - 1270
Gashyamba	1100 - 1270
Kamonyi	1100 - 1270
Gashyamba	1100 - 1270
Kyungu	1100 - 1270
Rwamagana	1100 - 1270
Nyamagabe	1100 - 1270
Karongi	1100 - 1270
Nyamashene	1300 - 1500
Rusizi	1300 - 1500
Nyagatare	1300 - 1500
Musanze	1300 - 1500
Butera	1300 - 1500
Gicumbi	1300 - 1500
Gatsibo	1300 - 1500
Kayanza	1300 - 1500
Kirehe	1300 - 1500
Nigoma	1300 - 1500
Bugesera	1300 - 1500
Ruhango	1300 - 1500
Nyanza	1300 - 1500
Huye	1300 - 1500
Gisagara	1300 - 1500
Nyaruguru	1300 - 1500

Furthermore, the working model defined in (2.13) has been used to handle missing observations of repeated measures data in small area estimation. As pointed out by Liski (1985), when there are missing data, the consequence is usually that the original problem becomes more complex and standard approaches are no longer applicable. Consider the model (2.13) and suppose that the corresponding sample data has a two-step monotonic pattern. That is if we denote by n the sample size. Then for all p time points considered in the study, there are p_1 time periods where all the units remain in the sample and p_2 time periods where the sample size is reduced to n_1 with $n_2 = n - n_1$ drop outs.

The model (2.13) for sample data is presented by two model equations:

$$Y = A_1 B H C + 1_{p_1} \gamma' X + U_1 Z + E_1, \quad (2.16)$$

$$\tilde{Y} = A_2 B H \tilde{C} + 1_{p_2} \gamma' \tilde{X} + U_2 \tilde{Z} + E_2, \quad (2.17)$$

where

$$\begin{aligned} U_1 &\sim \mathcal{N}_{p_1, m}(\mathbf{0}, \Sigma_{11}^u, I_m), & E_1 &\sim \mathcal{N}_{p_1, n}(\mathbf{0}, I_{p_1}, \sigma_{E_1}^2 I_n), \\ U_2 &\sim \mathcal{N}_{p_2, m}(\mathbf{0}, \Sigma_{22}^u, I_m), & E_2 &\sim \mathcal{N}_{p_2, n_1}(\mathbf{0}, I_{p_2}, \sigma_{E_2}^2 I_{n_1}). \end{aligned}$$

Following Ngaruye et al. (2016), models (2.16) and (2.17) are transformed into six different models:

$$\begin{aligned} V_0 &= Y C'^o = 1_{p_1} \gamma' X C'^o + E_1 C'^o, \\ V_1 &= Y R_1 = A_1 B K_1 + 1_{p_1} \gamma' X R_1 + (U_1 Z + E_1) R_1, \\ V_2 &= Y R_2 = A_1 B K_2 + 1_{p_1} \gamma' X R_2 + E_1 R_2, \\ \tilde{V}_0 &= \tilde{Y} \tilde{C}'^o = 1_{p_2} \gamma' \tilde{X} \tilde{C}'^o + E_2 \tilde{C}'^o, \\ \tilde{V}_1 &= \tilde{Y} \tilde{R}_1 = A_2 B \tilde{K}_1 + 1_{p_2} \gamma' \tilde{X} \tilde{R}_1 + (U_2 \tilde{Z} + E_2) \tilde{R}_1, \\ \tilde{V}_2 &= \tilde{Y} \tilde{R}_2 = A_2 B \tilde{K}_2 + 1_{p_2} \gamma' \tilde{X} \tilde{R}_2 + E_2 \tilde{R}_2. \end{aligned}$$

Denote the likelihood for the observations Y with parameters Θ by $L(Y; \Theta)$ and using this notation the next theorem summarizes the estimation procedure.

Theorem 3. *For the model given by (2.16) and (2.17) let $\Psi_{22} = \Sigma_{22}^u + \sigma_{E_2}^2 I_{p_2}$, $\Psi_{11} = \Sigma_{11}^u + \sigma_{E_1}^2 I_{p_1}$, $\Theta = \Sigma_{21}^u \Psi_{11}^{-1}$ and $\Psi_{2\bullet 1} = \Psi_{22} - \Theta \Psi_{11} \Theta'$. Then B and γ are estimated by maximizing*

$$L(V_0; \gamma) \times L(\tilde{V}_0; \gamma) \times L(V_2; B, \gamma) \times L(\tilde{V}_2; B, \gamma) \quad (2.18)$$

and Σ^u is estimated by maximizing

$$L(\tilde{V}_1 | V_1; B, \gamma, \Psi_{2\bullet 1}, \Theta) \times L(V_1; B, \gamma, \Psi_{11}),$$

when inserting the estimated B and γ , and using $\Sigma_{11}^u = \Psi_{11} - \sigma_{E_1}^2 I_{p_1}$, $\Sigma_{22}^u = \Psi_{2\bullet 1} - \sigma_{E_2}^2 I_{p_2} + \Theta \Psi_{11} \Theta'$ and $\Sigma_{21}^u = \Theta \Psi_{11}$.

Finally, under the assumption of full rank matrices $\mathbf{X}$ and $\mathbf{K}_2^{o'}\mathbf{K}_1$ and $p = q$, the first and second moments for the proposed small area estimators are computed with the purpose to use these moments to estimate the mean-squared errors (MSE) for the predicted small area means as a measure of precision. At the first stage, we suppose that we have complete knowledge on the covariance matrix Σ_u or it has been estimated previously so that it is taken to be as known. At the second stage, parametric bootstrap is proposed for bias correction and for prediction error that reflects the uncertainty when the unknown covariance is replaced by its suitable estimator. For known Σ_u , then $\tilde{\mathbf{U}} = (\Sigma_e + \Sigma_u)^{-1}\Sigma_u\mathbf{V}_3\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{Z}'$. In the following two theorems, we present the moments of the proposed estimators presented in Theorem 2 with matrix $\mathbf{A}$ taken as $\mathbf{I}_p$ and the notation $(\mathbf{Y})(\cdot)'$ standing for $(\mathbf{Y})(\mathbf{Y})'$.

Theorem 4. *Given the estimators obtained from model (2.13) when $\mathbf{X}$ and $\mathbf{K}_2^{o'}\mathbf{K}_1$ are of full rank and $p = q$. The dispersion matrices $\mathbb{D}[\hat{\gamma}]$, $\mathbb{D}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}]$ and $\mathbb{D}[\tilde{\mathbf{U}}]$ are given by*

$$\begin{aligned}\mathbb{D}[\hat{\gamma}] &= \frac{\sigma_e^2}{p}(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}, \\ \mathbb{D}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}] &= \mathbf{C}'\mathbf{H}'\mathbf{P}_1\mathbf{H}\mathbf{C} \otimes (\Sigma_u + \Sigma_e) + \left(\mathbf{C}'\mathbf{H}'(\mathbf{I}_k - \mathbf{P}_1\mathbf{K}_1\mathbf{K}'_1)(\mathbf{K}_2\mathbf{K}'_2)^- \right. \\ &\quad \times (\mathbf{I}_k - \mathbf{K}_1\mathbf{K}'_1\mathbf{P}_1)\mathbf{H}\mathbf{C} \Big) \otimes \Sigma_e + \frac{\sigma_e^2}{p}\mathbf{C}'\mathbf{H}'\left((\mathbf{K}_2\mathbf{K}'_2)^- \mathbf{K}_2\mathbf{R}'_2 + \mathbf{P}_1\mathbf{K}_1\mathbf{R}'_1\mathbf{G}'_2 \right) \\ &\quad \times \mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\left(\mathbf{R}_2\mathbf{K}'_2(\mathbf{K}_2\mathbf{K}'_2)^- + \mathbf{G}_2\mathbf{R}_1\mathbf{K}'_1\mathbf{P}_1 \right)\mathbf{H}\mathbf{C} \otimes \mathbf{1}_p\mathbf{1}'_p, \\ \mathbb{D}[\tilde{\mathbf{U}}] &= \sigma_e^2\mathbf{Z}\mathbf{R}_1\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}(\mathbf{I}_m + \mathbf{K}'_1(\mathbf{K}_2\mathbf{K}'_2)^-\mathbf{K}_1)\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{Z}' \otimes \Sigma_u(\Sigma_e + \Sigma_u)^{-2}\Sigma_u \\ &\quad + \frac{\sigma_e^2}{p}\mathbf{Z}\mathbf{R}_1\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{G}'_2\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\mathbf{G}_2\mathbf{R}_1\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{Z}' \\ &\quad \otimes \Sigma_u(\Sigma_e + \Sigma_u)^{-1}\mathbf{1}_p\mathbf{1}'_p(\Sigma_e + \Sigma_u)^{-1}\Sigma_u.\end{aligned}$$

Theorem 5. *Given the estimators obtained from model (2.13) when $\mathbf{X}$ and $\mathbf{K}_2^{o'}\mathbf{K}_1$ are of full rank and $p = q$. Then, the covariances $\text{cov}[\tilde{\mathbf{U}}, \hat{\gamma}]$, $\text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \tilde{\mathbf{U}}]$ and $\text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \hat{\gamma}]$ are given by*

$$\begin{aligned}\text{cov}[\tilde{\mathbf{U}}, \hat{\gamma}] &= -\frac{\sigma_e^2}{p}\mathbf{Z}\mathbf{R}_1\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{G}'_2\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1} \otimes \Sigma_u(\Sigma_e + \Sigma_u)^{-1}\mathbf{1}_p, \\ \text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \tilde{\mathbf{U}}] &= \frac{\sigma_e^2}{p}\mathbf{C}'\mathbf{H}'\left((\mathbf{K}_2\mathbf{K}'_2)^- \mathbf{K}_2\mathbf{R}'_2 + \mathbf{P}_1\mathbf{K}_1\mathbf{R}'_1\mathbf{G}'_2 \right)\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1}\mathbf{X}\mathbf{G}_2\mathbf{R}_1\mathbf{Q}_{\mathbf{K}'_1\mathbf{K}_2}\mathbf{R}'_1\mathbf{Z}' \\ &\quad \otimes \mathbf{1}_p\mathbf{1}'_p(\Sigma_e + \Sigma_u)^{-1}\Sigma_u, \\ \text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \hat{\gamma}] &= -\frac{\sigma_e^2}{p}\mathbf{C}'\mathbf{H}'\left((\mathbf{K}_2\mathbf{K}'_2)^- \mathbf{K}_2\mathbf{R}'_2 + \mathbf{P}_1\mathbf{K}_1\mathbf{R}'_1\mathbf{G}'_2 \right)\mathbf{X}'(\mathbf{X}\mathbf{P}\mathbf{X}')^{-1} \otimes \mathbf{1}_p.\end{aligned}$$

Put

$$\mathbf{K} = \left(1 - \frac{n_{ig}}{N_{ig}}\right) \mathbf{A}, \quad \mathbf{L} = \frac{1}{N_{ig}} \mathbf{1}_p \mathbf{1}'_{N_{ig}-n_{ig}} \mathbf{X}_{ig}^{(r)'} , \quad \mathbf{M} = \frac{\sqrt{N_{ig}-n_{ig}}}{N_{ig}} \mathbf{I}_p, \quad (2.19)$$

$$i = 1, \dots, m, \quad g = 1, \dots, k.$$

Then, the linear prediction and empirical linear prediction quantities from small area means given by the second term of the right hand side in (2.14) can be written by

$$\begin{aligned} \tilde{\tau}_{ig} &= \mathbf{K} \hat{\beta}_g + \mathbf{L} \hat{\gamma} + \mathbf{M} \tilde{u}_i, \\ \hat{\tau}_{ig} &= \mathbf{K} \hat{\beta}_g + \mathbf{L} \hat{\gamma} + \mathbf{M} \hat{u}_i, \quad i = 1, \dots, m, \quad g = 1, \dots, k. \end{aligned}$$

Let ν_g and δ_i to be a k -vector and an m -vector of zeros with one on the g -th and i -th position, respectively. Choose $\alpha_g = \mathbf{C}' \mathbf{H}' (\mathbf{H} \mathbf{C} \mathbf{C}' \mathbf{H}')^{-1} \nu_g$, then $\hat{\beta}_g = \hat{\mathbf{B}} \mathbf{H} \mathbf{C} \alpha_g$, $\tilde{u}_i = \tilde{\mathbf{U}} \delta_i$ and $\hat{u}_i = \hat{\mathbf{U}} \delta_i$.

The MSE of $\hat{\tau}_{ig}$ is obtained by

$$\begin{aligned} \text{MSE}(\hat{\tau}_{ig}) &= \mathbb{E} \left[\left((\hat{\tau}_{ig} - \tilde{\tau}_{ig}) + (\tilde{\tau}_{ig} - \tau_{ig}) \right) \left(\right)' \right] \\ &= \mathbf{G}_{1i}(\Sigma_u) + \mathbf{G}_{2i}(\Sigma_u) + \mathbf{G}_{3i}(\Sigma_u), \end{aligned}$$

where

$$\begin{aligned} \mathbf{G}_{1i}(\Sigma_u) + \mathbf{G}_{2i}(\Sigma_u) &= \mathbb{E}[(\tilde{\tau}_{ig} - \tau_{ig})(\tilde{\tau}_{ig} - \tau_{ig})'] \\ \mathbf{G}_{3i}(\Sigma_u) &= 2\mathbb{E}[(\hat{\tau}_{ig} - \tilde{\tau}_{ig})(\tilde{\tau}_{ig} - \tau_{ig})'] + \mathbb{E}[(\hat{\tau}_{ig} - \tilde{\tau}_{ig})(\hat{\tau}_{ig} - \tilde{\tau}_{ig})']. \end{aligned}$$

Consider the bootstrap model

$$\mathbf{Y}_i^* \mid \mathbf{u}_i^* \sim \mathcal{N}_{p,n_i}(\mathbf{A} \hat{\mathbf{B}} \mathbf{C}_i + \mathbf{1}_p \hat{\gamma}' \mathbf{X}_i + \mathbf{u}_i^* \mathbf{z}_i', \Sigma_e, \mathbf{I}_{n_i}), \quad i = 1, \dots, m, \quad (2.20)$$

where $\mathbf{u}_i^* \sim \mathcal{N}_p(\mathbf{0}, \hat{\Sigma}_u)$. Then, the estimator of MSE can be expressed by

$$\begin{aligned} \widehat{\text{MSE}}(\hat{\tau}_{ig}) &= 2 \left[\mathbf{G}_{1i}(\hat{\Sigma}_u) + \mathbf{G}_{2i}(\hat{\Sigma}_u) \right] - \mathbb{E}_* \left[\mathbf{G}_{1i}(\hat{\Sigma}_u^*) + \mathbf{G}_{2i}(\hat{\Sigma}_u^*) \right] \\ &\quad + 2\mathbb{E}_* \left[\left(\hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) - \hat{\tau}_{ig}(\mathbf{Y}_i^*; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) \right) \right. \\ &\quad \times \left. \left(\hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) - \hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g, \hat{\gamma}, \hat{\Sigma}_u) \right)' \right] \\ &\quad + \mathbb{E}_* \left[\left(\hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) - \hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) \right) \left(\right)' \right], \quad (2.21) \end{aligned}$$

where

$$\begin{aligned} \mathbf{G}_{1i}(\hat{\Sigma}_u) + \mathbf{G}_{2i}(\hat{\Sigma}_u) &= \mathbf{K}(\alpha'_g \otimes \mathbf{A}^{-1}) \mathbb{D}[\hat{\mathbf{B}} \mathbf{H} \mathbf{C}] (\alpha_g \otimes \mathbf{A}'^{-1}) \mathbf{K}' + \mathbf{K}(\alpha'_g \otimes \mathbf{A}^{-1}) \text{cov}[\hat{\mathbf{B}} \mathbf{H} \mathbf{C}, \hat{\gamma}] \mathbf{L}' \\ &\quad + \mathbf{K}(\alpha'_g \otimes \mathbf{A}^{-1}) \text{cov}[\hat{\mathbf{B}} \mathbf{H} \mathbf{C}, \tilde{\mathbf{U}}] (\delta_i \otimes \mathbf{I}_p) \mathbf{M}' + \mathbf{L} \mathbb{D}[\hat{\gamma}] \mathbf{L}' \\ &\quad + \mathbf{L} \text{cov}[\tilde{\mathbf{U}}, \hat{\gamma}]' (\delta_i \otimes \mathbf{I}_p) \mathbf{M}' + \mathbf{M}(\delta'_i \otimes \mathbf{I}_p) \text{cov}[\hat{\mathbf{B}} \mathbf{H} \mathbf{C}, \tilde{\mathbf{U}}]' (\alpha_g \otimes \mathbf{A}'^{-1}) \mathbf{K}' \\ &\quad + \mathbf{L} \text{cov}[\hat{\mathbf{B}} \mathbf{H} \mathbf{C}, \hat{\gamma}]' (\alpha_g \otimes \mathbf{A}'^{-1}) \mathbf{K}' + \mathbf{M}(\delta'_i \otimes \mathbf{I}_p) \text{cov}[\tilde{\mathbf{U}}, \hat{\gamma}] \mathbf{L}' \\ &\quad + \mathbf{M} \Sigma_u \mathbf{M}' - \mathbf{M}(\delta'_i \otimes \mathbf{I}_p) \mathbb{D}[\tilde{\mathbf{U}}] (\delta_i \otimes \mathbf{I}_p) \mathbf{M}'. \quad (2.22) \end{aligned}$$

The dispersion matrices $\mathbb{D}[\hat{\gamma}]$, $\mathbb{D}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}]$, $\mathbb{D}[\tilde{\mathbf{U}}]$ and the covariance matrices $\text{cov}[\tilde{\mathbf{U}}, \hat{\gamma}]$, $\text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \tilde{\mathbf{U}}]$, $\text{cov}[\hat{\mathbf{B}}\mathbf{H}\mathbf{C}, \hat{\gamma}]$ involved in equation (2.22) are presented in Theorem 4 and Theorem 5 with Σ_u replaced by $\hat{\Sigma}_u$, respectively. The quantity $\mathbf{G}_{1i}(\hat{\Sigma}_u^*) + \mathbf{G}_{2i}(\hat{\Sigma}_u^*)$ is obtained similarly with Σ_u replaced by $\hat{\Sigma}_u^*$. Moreover, $\mathbb{E}_*$ is the expectation with respect to model (2.20), and the calculation of $\hat{\gamma}^*$, $\hat{\beta}_g^*$, $\hat{\Sigma}_u^*$ is performed similarly to that of $\hat{\gamma}$, $\hat{\beta}_g$, $\hat{\Sigma}_u$ except that $\hat{\gamma}^*$, $\hat{\beta}_g^*$, $\hat{\Sigma}_u^*$ are calculated based on $\mathbf{Y}_i^*$ instead of $\mathbf{Y}_i$ and

$$\begin{aligned}\hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g, \hat{\gamma}, \hat{\Sigma}_u) &= \mathbf{K}\hat{\beta}_g + \mathbf{L}\hat{\gamma} + \mathbf{M}\tilde{\mathbf{u}}_i, \\ \hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u) &= \mathbf{K}\hat{\beta}_g^* + \mathbf{L}\hat{\gamma}^* + \mathbf{M}\tilde{\mathbf{u}}_i^*, \\ \hat{\tau}_{ig}(\mathbf{Y}_i; \hat{\beta}_g^*, \hat{\gamma}^*, \hat{\Sigma}_u^*) &= \mathbf{K}\hat{\beta}_g^* + \mathbf{L}\hat{\gamma}^* + \mathbf{M}\hat{\mathbf{u}}_i^*, \quad i = 1, \dots, m, \quad g = 1, \dots, k.\end{aligned}$$

Summary of contributions, concluding remarks and further research

IN this chapter, we present a short summary of the contributions of the papers included in the thesis, discussion of the results, concluding remarks and suggestions for further research.

3.1 Summary of contributions

The main contribution of this thesis is that a new approach based on random effects growth curve modeling has been introduced in small area estimation for repeated measures data. The main contributions of each paper are summarized as follows:

- In Paper A, a random effects growth curve model is formulated in the context of small area estimation. A simple orthogonal transformation is used to partition the model into independent sub-models which are matrix normal distributed and a likelihood based approach is used to derive the estimators of the model parameters. A method consisting of maximizing the joint density of observable and non-observable random variable is used for prediction of random effects. A superpopulation approach to survey sampling for finite population yields predicted small area means. Also, a parametric bootstrap method is proposed for estimation of mean-squared errors of the predicted small area means.
- In Paper B, we apply the results obtained from Paper A to a real case study. The main focus is to estimate crop yield at district level for agricultural seasons 2014 in Rwanda. The crop yield of interest was beans comprising bush beans and climbing beans. Firstly, we identify the relevant covariates that influence the crop yield, adapt the model to the data set to accommodate the sampling weights and conduct the data analysis. Secondly, we derive the beans yield estimates and discuss the assessment

of the results. We end up by providing a cluster map that shows the distribution of average beans yield estimates per district across the country.

- In Paper C, the first and second moments of the proposed estimators in Paper A are presented. Also, an approximation and estimation of mean-squared errors for predicted small area means are investigated.
- In Paper D, the random effects growth curve model proposed in Paper A is extended to handle missing data that follow a monotonic pattern. First, the model equations of complete and incomplete sample data under the monotonic structure are formulated similarly as proceeded in Paper A. Then, the models are one-to-one transformed into six different models. The conditional distribution for incomplete data given the complete is derived and a conditional likelihood based approach is proposed for estimation of mean and covariance parameters. Both prediction of random effects and small area means are also presented.

3.2 Concluding remarks

The aims of this thesis have been met and discussed through relevant papers.

In Paper A, the knowledge from random effects growth curve model has been applied to small area estimation context. The working model is a random effects growth curve model with covariates. This model belongs to the class of extensions of unit-level model often known as nested linear regression model. When $\mathbf{U}$ and $\mathbf{E}$ are both random matrices which are matrix normal distributions, it is not the case that $\mathbf{UZ} + \mathbf{E}$, for some known matrix $\mathbf{Z}$ is matrix normally distributed. The reason is that the resulting covariance matrix is not always separable Kollo and von Rosen (2005). Therefore, in Paper A we have made a one-to-one transformation to split the working model into three independent models which are matrix normally distributed. Thus, the resulting estimators are not maximum likelihood estimators. It is worth noting that we may face the problem of negative definiteness of the covariance estimator $\hat{\Sigma}_u$ for big value of sampling variance. When the analytic expression of the mean-squared errors of the predicted quantities can not be obtained, a re-sampling method is a good alternative to asymptotic analytical approximation of MSE. With the advancement of modern computer based technology, the bootstrap method has been the most used re-sampling method. This technique consists of drawing a large number of independent re-samples or bootstrap samples with replacement from original sample and then performing inference about a sample from re-sampled data. For parametric bootstrap, when the population distribution is known, the method consists of repeatedly re-sampling from the

fitted distribution of the sample data and using the bootstrap sample calculate an estimator of the required statistic to estimate the uncertainty about the true value of this statistic (Vose, 2008). Parametric bootstrap produced more stable estimates of standard errors than the bootstrap in most of the conditions studied by Cui and Kolen (2008). Lahiri (2003) gave a comprehensive account of the bootstrap for small area estimation inference. Butar and Lahiri (2003) have developed a parametric bootstrap method to estimate MSE of predicted quantity of interest under the Fay-Herriot model. In Paper A, we have used a parametric bootstrap method to estimate the mean-squared errors for the predicted small area means.

As discussed in Paper B, when the sampling scheme used results in a selection of sample units with unequal probabilities within small area means, there should be a way to accommodate sampling weights in the prediction of small area means. One way of extending unweighted predictor under linear mixed model to account for design weights is by augmenting the design matrix by design weights. That is to include an additional regression coefficient for the effect of design weights on the variable of interest Burgard et al. (2014). In Paper B, an augmented model which includes design weights is fitted to the survey data by an additional unknown regression coefficient accounting for weight effect. We should point out that this empirical study was not relevant to extend the results at a smaller administrative unit and to investigate the trends of crop yield over time.

The results about estimation of the MSE obtained from Paper C can be applied to the empirical study carried out in paper B to assess the precision of the district level model-based beans yield estimates by making the comparison of their MSE with the MSE of the district level direct beans yield estimates.

In Paper D, the efficient method of expectation maximization which is widely used as inference based for missing observations under the assumption of multivariate normal distribution of the data (Liski, 1985; Liski and Nummi, 1990; Nummi, 1997) was not found as a suitable approach for our model, simply because at each maximization step the model has to be transformed in order to obtain likelihood equations that are linear with respect to model parameters.

3.3 Further research

Regarding the future work, we plan to complete Paper D by including an illustration example of a case study with relevant empirical sample data that follow a monotonic pattern. Furthermore, the working model considered in Paper A assumes the random area effects for different areas to be independently distributed and this assumption can be relaxed. We think of putting up a multivariate linear model that accounts for:

- time correlation effects,
- spatial correlation between areas,
- random area effects,
- group effects.

The idea is to take into account the correlation between neighboring areas. In that case, spatially correlated small area effects are described by a proximity matrix and a spatial autocorrelation parameter. There are different ways of choosing a spatial interaction matrix, one can consider a contiguity matrix as the proximity matrix, where its elements take nonzero values for those pairs of areas that are adjacent and zeros elsewhere. Once the model is formulated, the aim will be to propose an inference based method for estimating models parameters, predicting random effects and small area means. Finally, throughout this work, the target parameters of interest have been small area means. We plan to carry out prediction for other characteristics of interest such as small area totals and small area quantiles.

Bibliography

- Azzalini, A. (1987). Growth curves analysis for patterned covariance matrices, in: M.L. Puri, J.P. Vilaplana, W. Wertz (Eds.). *New Perspectives in Theoretical and Applied Statistics*, pages 61–74.
- Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.
- Benavent, R. and Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94:372–390.
- Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*. Springer-Verlag, New York.
- Brackstone, G. (1987). Small Area Data: Policy Issues and Technical Challenges, in R. Platek, J.N.K. Rao, C-E. Särndal, and M.P. Singh (Eds.). *Small Area Statistics*, pages 3–20.
- Burgard, J. P., Münnich, R., and Zimmermann, T. (2014). The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics*, 30(4):749–771.
- Butar, F. B. and Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112(1):63–76.
- Chambers, R. and Clark, R. G. (2012). *An introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford, New York.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, New York.
- Cui, Z. and Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement*, 32(4):334–347.
- Das, K., Jiang, J., and Rao, J. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32(2):818–840.
- Datta, G. S., Lahiri, P., Maiti, T., and Lu, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the US. *Journal of the American Statistical Association*, 94(448):1074–1082.
- Datta, G. S., Rao, J., and Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92(1):183–196.

- Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.
- Filipiak, K. and von Rosen, D. (2012). On MLEs in an extended multivariate linear growth curve model. *Metrika*, 75(8):1069–1092.
- Fuller, W. and Harter, R. (1987). The Multivariate Components of Variance Model for Small Area Estimation, in R. Platek, J.N.K. Rao, C-E. Särndal, and M.P. Singh (Eds.). *Small Area Statistics*, pages 103–123.
- Fuller, W. A. (2009). *Sampling Statistics*. John Wiley and Sons, New Jersey.
- Ghosh, M., Datta, G., and Fay, R. (1991). Hierarchical and empirical multivariate bayes analysis in small area estimation. In *Proceedings of the Bureau of the Census 1991 Annual Research Conference*, pages 63–79, Washington, DC. U.S. Bureau of the Census.
- Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9(1):55–76.
- Gonzalez, M. E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, pages 33–36.
- González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., and Santamaría, L. (2008a). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5):443–462.
- González-Manteiga, W., Morales, D., Molina, I., Lombardía, M. J., and Santamaría, L. (2008b). Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics and Data Analysis*, 52:5242–5252.
- Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):221–238.
- Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium):10–41.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.
- Ip, W. C., Wu, M. X., Wang, S. G., and Wong, H. (2007). Estimation for parameters of interest in random effects growth curve models. *Journal of Multivariate Analysis*, 98(2):317–327.

- Jiongo, V. D., Haziza, D., and Duchesne, P. (2013). Controlling the bias of robust small-area estimators. *Biometrika*, 100(4):843–858.
- Kackar, R. N. and Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics - Theory and Methods*, 10(13):1249–1261.
- Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–862.
- Khatri, C. G. (1973). Testing some covariance structures under a growth curve model. *Journal of Multivariate Analysis*, 3(1):102–116.
- Kollo, T. and von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht.
- Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 18(2):199–210.
- Lahiri, S. N., Maiti, T., Katzoff, M., and Parsons, V. (2007). Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, 94(2):469–485.
- Lange, N. and Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, 84(405):241–247.
- Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. *Handbook of Statistics*, 29:219–249.
- Liski, E. P. (1985). Estimation from incomplete data in growth curves models. *Communications in Statistics - Simulation and Computation*, 14(1):13–27.
- Liski, E. P. and Nummi, T. (1990). Prediction in growth curve models using the EM algorithm. *Computational Statistics and Data Analysis*, 10(2):99–108.
- Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A*, 154(3):421–441.
- Ngaruye, I., Nzabanita, J., von Rosen, D., and Singull, M. (2016). Small area estimation under a multivariate linear model for repeated measures data. *Communications in Statistics - Theory and Methods*. <http://dx.doi.org/10.1080/03610926.2016.1248784>.
- Nissinen, K. (2009). *Small Area Estimation with Linear Mixed Models from Unit-level panel and Rotating panel data*. PhD thesis, University of Jyväskylä.

- Nummi, T. (1997). Estimation in a random effects growth curve model. *Journal of Applied Statistics*, 24(2):157–168.
- Nzabanita, J., Singull, M., and von Rosen, D. (2012). Estimation of parameters in the extended growth curve model with a linearly structured covariance matrix. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 16(1):13–32.
- Ohlson, M. and Srivastava, M. S. (2010). Profile analysis for a growth curve model. *Journal of the Japan Statistical Society*, 40:1–21.
- Pan, J.-X., Fang, K., and Fang, K.-T. (2002). *Growth Curve Models and Statistical Diagnostics*. Springer Science & Business Media, New York.
- Pereira, L. N. and Coelho, P. S. (2010). Assessing different uncertainty measures of eblup: a resampling-based approach. *Journal of Statistical Computation and Simulation*, 80(7):713–727.
- Pfeffermann, D. (2002). Small area estimation-new developments and directions. *International Statistical Review*, 70(1):125–143.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.
- Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3/4):313–326.
- Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171.
- Rahman, A. (2008). A review of small area estimation problems and methodological developments. Discussion paper, NATSEM-University of Canberra.
- Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons, New York.
- Rao, J. N. K. and Isabel, M. (2015). *Small Area Estimation*. John Wiley and Sons, New York, 2 edition.
- Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77(377):190–195.

- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32.
- Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag, New York.
- Singh, B. and Sisodia, B. S. V. (2011). Small area estimation in longitudinal surveys. *Journal of Reliability and Statistical Studies*, 4(2):83–91.
- Srivastava, M. S. and Singull, M. (2012). Profile analysis with random-effects covariance structure. *Journal of the Japan Statistical Society*, 42(2):145–164.
- The EURAREA Consortium (2004). Enhancing small area estimation techniques to meet european needs. Technical report, Office for National Statistics, London.
- von Rosen, D. (1989). Maximum likelihood estimators in multivariate linear normal models. *Journal of Multivariate Analysis*, 31(2):187–200.
- Vose, D. (2008). *Risk Analysis: a Quantitative Guide*. John Wiley and Sons.
- Yokoyama, T. (2001). Estimation in a random effects model with parallel polynomial growth curves. *Hiroshima Mathematical Journal*, 31(3):425–433.
- Yokoyama, T. and Fujikoshi, Y. (1993). A parallel profile model with random-effects covariance structure. *Journal of Japan Statistical Society*, 23(1):83–89.

Papers

The articles associated with this thesis have been removed for copyright reasons. For more details about these see:

[http://urn.kb.se/resolve? urn:nbn:se:liu:diva-137206](http://urn.kb.se/resolve?urn:nbn:se:liu:diva-137206)