UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

Division of Social Statistics

Improved Direct Estimators for Small Areas

by

Hukum Chandra

Thesis for the degree of Doctor of Philosophy

August 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Social Sciences

Division of Social Statistics

Doctor of Philosophy

IMPROVED DIRECT ESTIMATORS FOR SMALL AREAS

by Hukum Chandra

Improved direct estimators for small area estimation (SAE) are investigated and extended in this thesis.

Unbiased direct estimators for small area quantities are usually considered too variable to be of any practical use. In this thesis we described a class of model based direct (MBD) estimators for small area quantities that appears to overcome this objection, in the sense that these estimators are comparable in efficiency to the indirect model-based small area estimators (e.g. empirical best linear unbiased predictors, or EBLUPs) that are now widely used. There are many practical advantages associated with such MBD estimators, arising from the fact that they are computed as weighted linear combinations of the actual sample data from the small areas of interest. In this case the weights 'borrow strength' via a model that explicitly allows for small area effects. One particular advantage that we explore in this thesis is that estimation of mean squared error (MSE) is then straightforward, using well-known methods that are in common use for population level estimates. Empirical results show that the MBD estimator represents a real alternative to the EBLUP, with the simple MSE estimator associated with the MBD estimator providing good coverage performance. Further, our results indicate that the MBD estimator may be more robust than the EBLUP when the small area model is incorrectly specified.

We extended the MBD approach to multipurpose SAE. Our results indicate these multipurpose weights are efficient across a range of variables, including variables that are ill-suited to EBLUP, e.g. variables that contain a significant proportion of zeros. We also show that these multipurpose weights remain efficient across a wide range of variables, even variables that have not been used in the definition of the multipurpose weights. We also extended the MBD approach to SAE for skewed data where the linear model provides poor fit and standard methods of small area estimation are inefficient. The proposed method based on the log-log transform model with random effects show significant gains in small area estimation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I, first of all, thank God for his love and care, for all the strength, courage, patience and determination He has given me.

I am deeply grateful to Professor Ray Chambers for his supervision, encouragement and patience during my PhD study. Apart from teaching me statistics and research methodology, Ray has also taught me the way in which to control my stress by always viewing the positive side things. I am indebted to Dr James Brown for his supervision. His advice has been absolutely essential for the progress of my research.

The friendly and outstanding academic and research environment of the Division of Social Statistics and of the Southampton Statistical Sciences Research Institute at the University of Southampton has been very beneficial to me. Acknowledgements are also due to Professor Danny Pfeffermann, Dr Nikos Tzavidis, Dr Sabu Padmadasan and Professor Pedro Silva. Their support was really valuable. A special thank to all Research Office staff and to Dr Darren Hampton and Dr Sandy MacKinnon for their support. I am indebted to my fellow colleagues for their friendship and support, especially Leonardo, Caroline and Solange.

I owe the deepest gratitude to my adored wife Shaily for being always at my side. I thank her for the patience, care, dedication and unconditional love. Without her I would have never been able to finish this thesis. I am thankful to my beloved daughter Amishi for her cooperation throughout my PhD research. I am also grateful to my family members and friends back home in India for their continuous support and inspiration during my study.

I want to thank the Commonwealth Scholarship Commission and British Council, United Kingdom for their financial support and Indian Council of Agricultural Research, Government of India for giving me the study leave from the office. The Division of Social Statistics sponsored my participation in most of the conferences and extra courses attended while undertaking my studies. Whilst undertaking this research, I was fortunate to visit the University of Wollongong, Australia and the Australian Bureau of Statistics in Canberra, Australia. I thank the University of Wollongong, and the Australian Bureau of Statistics for supporting my visit.

This thesis is dedicated to my parents Sh. (Late) Ram Shankar and Smt. Ramawati

CHAPTER 1

INTRODUCTION

1.1 Introduction

Small area estimation plays a prominent role in survey sampling due to growing demands for reliable small area statistics from both public and private sectors. The use of small area statistics have existed for a long time. The existence of the Domesday Book in eleventh century England and in the seventeenth century Canadian small area data based on the 1666 census is described in Brackstone (1987). In those early days the small area statistics were based either on a census or on administrative records. In either case the process relies on the complete enumeration of the domain of interest, no sampling is involved. However, for the past 40 years, sample surveys have been recognised as a mean for providing efficient and cost-effective national and sub-national estimates at frequent intervals and consequently, for most purposes, have replaced the complete enumeration.

Sample surveys, whether they are conducted by government organisations or by private entities, aim to produce reasonably accurate direct estimators, not only for the characteristics of whole population but also for a variety of subpopulations or domains. These direct estimators are based on domain specific sample data. However, many policymakers and researchers also want to obtain statistics for small domains. A domain is regarded as 'small' if the domain-specific sample is not large enough to

support a direct estimator of adequate precision. In other words, the estimator is likely to have a large standard error due to the small size of the sample in the domain (Ghosh and Rao, 1994). These small domains are also called small areas, so called because the sample size in the area or domain from the survey is small. Thus, we need special methods to estimate the characteristics of these small areas, referred to as the small area estimation techniques.

1.2 Small Area Problem and Associated Estimators

Each small area typically denotes a subset of the population for which very little information is available from the sample survey. These subsets refer to a small geographic area (e.g., a county, a municipality, a census division etc.) or a demographic group (e.g., a specific age-sex-race group of people within a large geographical area) or a cross classification of both. A small area can be any part of the population defined by any method of stratification. The statistics related to these small areas are often termed as small area statistics. The term small area and small domain are interchangeably used in the literature.

In recent years, many countries in the world are transferring the responsibilities for many social and economic policies from national governments to the local governments. Policy planners want to make sure that resources are targeted effectively and efficiently at the areas most in need and for the evaluation of the success of this targeting at a local level, they need reliable small area statistics. The private sector also needs small area statistics for policy making since many businesses and industries rely on local socio-economic conditions. Feasibility studies, for

example, require the use of small area statistics. Small area estimates can be made available from various censuses of population, businesses, housing and agriculture. However, the demand for small area estimate also exists for the intercensal period when data usually come from sample surveys.

Due to the increasing demand, survey organizations are faced with producing the small area estimates from existing sample surveys. Unfortunately, sample sizes in small areas tend to be too small, sometimes non-existent, to provide domains specific reliable direct estimates for these small areas. In other words, for small domains (small in terms of sample size), the domain specific usual design-based direct estimates (see section 2.2) are too unstable to be used for planning and policy-making purposes as they are likely to produce unacceptably large standard errors due to the small sample size. Accurate direct estimates for small areas would require a substantial increase in the overall sample size which in turn could overwhelm an already constrained budget and which could further lengthen the data processing time. Consequently, there has been growing interest in developing a range of estimation techniques to answer this need for small area statistics without further burdening the resources of already constrained survey organizations.

Small area estimation (SAE) methods look at producing estimates with adequate precision for such small areas or domains, through an estimation procedure that 'borrows strength' from related areas or time periods (or both) and thus increase the overall (effective) sample size and precision. These estimation procedures are based on either implicit or explicit models that provide a link to related areas or time periods

(or both) through the use of supplementary data (auxiliary information) such as recent census counts and current administrative records, see Pfeffermann (2002).

The traditional estimation techniques based on implicit linking models are synthetic and composite estimation methods. In these methods, an unbiased estimator for a large area is used to derive estimators for smaller areas under the assumption that these small areas exhibit the same structure (with regard to the phenomenon being studied) as the initial large area. If this condition is not met, the result could be biased estimators.

We notice that the usual design-based direct estimators based on the area-specific sample data are unbiased but in general not very precise. The traditional indirect (synthetic) estimators obtained through the use of auxiliary information have smaller variance but are generally biased. Statistical theory of SAE proposes a way of combining both estimators in a linear fashion so that the resulting estimator represents a compromise between the absence of bias and minimal variance. The resulting composite estimator is the linear combination of the direct and indirect estimators that minimises the mean squared error (Ghosh and Rao, 1994).

The traditional indirect estimators such as synthetic and composite estimates have the advantage of being simple to implement. In addition, these estimation techniques provide a more efficient estimate than the corresponding design-based direct estimator for each small area through the use of implicit models which 'borrow strength' across the small areas. These models assume that all the areas of interest behave similarly with respect to the variable of interest and do not take into account the area specific

variability. However, we can find situations where validity of assumed model fails leading to a biased estimator. Consequently, explicit linking model which incorporate random area-specific effects that account for between area variation beyond that explained by the auxiliary variables included in the model provides a better approach to SAE. These random area effects in the mixed model capture the dissimilarities between the areas. In general, estimation methods based on an explicit models are more efficient than traditional methods based on an implicit model. The explicit models used in SAE are a special case of the linear mixed model and are very flexible in formulating and handling complex problems in SAE. However, availability of good auxiliary information and the determination of a suitable linking model is crucial. In this thesis, our emphasis will be on mixed model based SAE methods. See Saei and Chambers (2003) and Jiang and Lahiri (2006), among others, for an extensive review of SAE based on mixed models. The related references for comprehensive review on SAE methods are Ghosh and Rao (1994), Pfefferman (2002) and Rao (1999, 2003). In chapter 2 we shall return with a brief outline of some of the important SAE techniques existing in the literature. In this chapter we shall also elaborate some analytical expressions to illustrate different SAE methods.

1.3 Motivation and Aim of the Thesis

Several methods for SAE have been proposed in the literature. However, research is still continuing on the important problem of identifying SAE techniques that are efficient and also simple to implement, with estimation of mean squared error (MSE) a particular problem. The model-based predictive approach or empirical best linear unbiased (EBLUP) approach under mixed effect models is very common and proven

to be efficient for the SAE. Prasad and Rao (1990) using results obtained from Kackar and Harville (1984) developed approximations to the MSE of the EBLUP which account for variability due to estimation of the variance components. They also obtained nearly unbiased MSE estimators under normality. However, in this EBLUP approach, survey weights have got little or no relevance. Consequently, many practical advantages of weighted linear estimation are lost. Perhaps the most important of these is the simplicity of the estimation process. The calibrated weighting approach to SAE introduced in Chambers (2005) defines the model-based direct estimator for small area quantities, with a simple estimator of the mean squared error of this estimator. The simplicity and ease of implementation of this approach motivated us to undertake this detailed study.

The main aim of this thesis is to study the model based direct (MBD) estimation method of Chambers (2005) and compare it with the EBLUP method (Prasad and Rao, 1990), and to extend the MBD approach to multipurpose small area estimation and to small area estimation for skewed data.

1.4 Outline of the Thesis

This thesis is organised in seven chapters. The present chapter has provided an overview of the importance and need for small area statistics. It has also indicated issues and challenges in SAE. In addition, our motivations and the aims of our research topic have been summarized in previous section. The remaining part of this thesis is organised as below.

Chapter 2 of the thesis presents the review of some of the important SAE techniques, emphasis has been given to the mixed model based SAE methods. Further, a brief discussion on some recent developments in SAE methods is outlined. In addition, the gaps existing in the literature that this thesis study intends to address are discussed. This chapter prepares the foundation for the other chapters.

Chapter 3 introduces the calibrated weighting approach in SAE. The model-based direct (MBD) estimators for small areas are defined. This approach uses sample weights derived from a population level version of the mixed effects model to define weighted linear small area estimators as well as a simple expression for their mean squared error. An empirical result using Australian Agricultural and Grazing Industry Survey (AAGIS) data is reported, which evaluates the performance of the empirical best linear unbiased predictor (EBLUP) and the MBD methods of SAE. Further, robustness of these SAE methods under model misspecifications is examined. Furthermore, some discussion on practical issues to provide an argument that supports our empirical results is included. The results of this chapter also appear in Chandra and Chambers (2005, 2006c, 2006d) and Chambers and Chandra (2006).

Chapter 4 presents the SAE techniques in context of multivariate surveys. The multipurpose sample weights for SAE are introduced. The MBD estimators for small areas using multipurpose weights are described. Theoretical aspects on how such multipurpose sample weights can be constructed when small area estimates of more than one survey variable are required is discussed. An empirical result using AAGIS data is reported to examine the performance of proposed multipurpose SAE method. In addition, an empirical illustration is presented to see how much efficiency (if any)

is lost if the linear assumption based MBD estimation is applied to the categorical variables. The suitable estimator in this case is the indirect estimator under a generalized linear mixed model. Application of the MBD method to categorical data is examined and the performance is evaluated against the indirect estimator via simulation studies. The main results of this chapter are also reported in Chandra and Chambers (2006b).

In chapter 5 and 6 we have addressed the issues related to SAE for business surveys where the data are skewed, and linear models provide a bad fit. Chapter 5 focuses on theoretical development for SAE methods with skewed data. A transform variable based SAE method is develop for skewed data that is linear on the log-log transform scale. The MBD estimators for small areas are derived under a log-log linear mixed model. In deriving these methods both normal and gamma distribution for the random errors are assumed.

Chapter 6 is devoted to simulation studies that evaluate the performance of the different methods of SAE for skewed data proposed in chapter 5. Two types of simulation studies are considered. The first type of study uses model-based simulation to generate data. These data are then used to contrast the performance of proposed MBD estimators for skewed data derived under a log-log linear mixed model with the MBD and EBLUP under a linear mixed model. The robustness of these SAE methods is also examined under the model misspecification. The second type of simulation study was carried out using real data (AAGIS data) and design-based simulations to test these methods in the context of a real population and realistic sampling methods.

The results from chapter 5 and chapter 6 also appear in Chandra (2006) and Chandra and Chambers (2006a).

Finally, chapter 7 provides the summary of main findings and conclusions of this research. In addition, some possible further research topics are suggested.

CHAPTER 2

OVERVIEW OF SMALL AREA ESTIMATION TECHNIQUES

2.1 Introduction

In chapter 1 we briefly described the need of small area data and the problem of small area estimation (SAE). In this chapter we review some of the important and commonly employed methods of SAE existing in the literature. This chapter prepares a foundation for the proceeding chapters. The chapter is organised as follows. The direct, synthetic and composite methods of small area estimation are illustrated in sections 2.2-2.4 respectively. Section 2.5 is devoted to the application of mixed effect models in small area estimation with attention to unit level random effects model. In section 2.6 and 2.7 we introduce and discuss some recent developments in small area estimation such as the pseudo-EBLUP approach and model-based direct estimation. Section 2.8 elaborates some further extensions of the mixed effect model to small area estimation. Finally, section 2.9 summarizes the key points from this chapter.

2.2 Direct Estimators

As noted in the previous chapter, in many cases existing large national sample surveys are also used to produce estimates for domains (these can be planned or unplanned) of the population. When sample sizes are small these domains are called small areas. That is an area is regarded as *small* if the sample drawn from the area is not large

enough to yield direct estimates of adequate prevision. The estimation method defined for large domain or population level quantities becomes impossible to apply, mainly because sample sizes are typically small or even zero in some small areas of interest, so the direct estimates (i.e. domain-specific estimates) tend to be quite unstable. The direct estimates use the data on the survey variable from the domain of study and time of interest. For example, suppose a linear estimator based on sample weights $\{w_j; j \in s\}$ is used to make inference about population level quantities. Here, s denotes the sample of size n drawn with sampling design $p(s)$ from a population $U = \{1, ..., N\}$ of size N. Further, if $\pi_j = \sum_{j \in s} p(s)$ are the first order inclusion probabilities then $w_j = \pi_j^{-1}$ defines the design weight of element j. Under simple random sampling, $\pi_j = nN^{-1}$ and $w_j = Nn^{-1}$. Let a subscript of i denote restriction to small area $i (i = 1, ..., m)$. We assume that the population consists of m non-overlapping domains or small areas U_i each with population of size N_i such that $U = \bigcup_{i=1}^{m} U_i$ and $N = \sum_{i=1}^{m} N_i$. Let s_i be the part of the sample of size n_i that falls in small area i and $n = \sum_{i=1}^{m} n_i$. We denote by y_{ij} the value of j^{th} population unit in small area i for the characteristic of interest Y. The population mean of Y in the area i, $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_j$ could be then estimated using the same weights leading to estimator

$$\hat{\bar{Y}}_i^{Hajek} = \left(\sum_{j \in s_i} w_j \right)^{-1} \left(\sum_{j \in s_i} w_j y_j \right) \tag{2.1}$$

or, if the population size N_i of the small area i is known,

$$\hat{\bar{Y}}_i^{HT} = N_i^{-1} \left(\sum_{j \in s_i} w_j y_j \right) \tag{2.2}$$

The estimators (2.1) and (2.2) are sometimes referred to as direct estimators of small area i mean $\bar{Y}_i$. More precisely, the estimator (2.1) is referred as the Hájek type of the

direct estimator, and (2.2) as the Horvitz-Thompson (HT) type of the direct estimator. These names refer to alternative approaches to estimating finite population means in the classical sampling literature, see Cochran (1977) and Särndal, Swensson and Wretman (1992). Irrespective of which form of direct estimator is used, it is easy to see that its variance can be large when the area sample size n_i is small. For example, under simple random sampling, with no auxiliary information, a design-based direct estimator of the mean of Y for small area i ($\bar{Y}_i = N_i^{-1}\sum_{j\in U_i} y_j$) is

$$\hat{\bar{Y}}_i = \begin{cases} \bar{y}_i & \text{if } n_i \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where $\bar{y}_i = \sum_{s_i} w_j y_j \Big/ \sum_{s_i} w_j = \sum_{s_i} Nn^{-1} y_j \Big/ \sum_{s_i} Nn^{-1} = n_i^{-1}\sum_{j\in s_i} y_j$ is sample mean of Y in area i. The estimator (2.3) is conditionally unbiased for a fixed $n_i \geq 1$ since

$$E_p(\hat{\bar{Y}}_i) = E_p(\bar{y}_i) = E_{n_i}\left[E_p(\bar{y}_i \mid n_i)\right] = E_{n_i}(\bar{Y}_i) = \bar{Y}_i .$$

The conditional variance of (2.3) is

$$Var_p(\hat{\bar{Y}}_i \mid n_i) = (1-f_i)\,S_i^2/n_i \tag{2.4}$$

with $f_i = n_i/N_i$ and $S_i^2 = (N_i-1)^{-1}\sum_{j=1}^{N_i}(y_j - \bar{y}_i)^2, N_i \geq 2$. Here E_p and Var_p denotes the expectation and variance respectively under the design-based[1] approach. An unbiased estimator for S_i^2 is $s_i^2 = (n_i-1)^{-1}\sum_{j\in s_i}(y_j - \bar{y}_i)^2$. Thus, an unbiased estimator for variance (2.4) is given by $v(\hat{\bar{Y}}_i \mid n_i) = n_i^{-1}(1-f_i)s_i^2$ when N_i is known.

[1] In the design-based approach, an estimator $\hat{T}$ of T is said to be design-unbiased (or p-unbiased) if the design expectation of $\hat{T}$ equals to T i.e., $E_p(\hat{T}) = \sum p(s)\hat{T}_s = T$, where the summation is over all possible samples s under the specified design and $\hat{T}_s$ is the value of $\hat{T}$ for the sample s. In this approach, the population is fixed and the only randomness or stochastic process involved is the selection of random samples. No distribution and no model is involved, and expectation is over all possible samples from the population.

For unknown N_i, the sampling fraction $f_i = n_i/N_i$ is replaced by $f = n/N$ and then the estimator for variance (2.4) is $v(\hat{\bar{Y}}_i \mid n_i) = (1-f)\, s_i^2/n_i$.

From (2.4) it is obvious that for small sample size n_i, the variance will be larger unless the variability of the Y values is sufficiently small. Suppose that in addition to survey variable Y, values of p-auxiliary variables are also known. Let us denote by x_{ij} a $p\times 1$ vector of auxiliary variable X for the unit j in area i. Then with known auxiliary information, a more efficient design-based direct estimator for the i^{th} small area mean $\bar{Y}_i$ is the regression estimator defined as

$$\hat{\bar{Y}}_i^{reg} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'\beta_i \tag{2.5}$$

where β_i is the vector of regression coefficients in area i, $\bar{x}_i = n_i^{-1}\sum_{j\in s_i} x_j$ and $\bar{X}_i = N_i^{-1}\sum_{j=1}^{N_i} x_j$ are the sample mean and population mean of auxiliary variable X in the area i respectively. The variance of (2.5) is

$$Var_p(\hat{\bar{Y}}_i^{reg} \mid n_i) \approx n_i^{-1}(1-f_i)S_i^2(1-\rho_i^2) = (1-\rho_i^2)Var_p(\hat{\bar{Y}}_i \mid n_i) \tag{2.6}$$

where ρ_i is the multiple correlation between survey variable Y and auxiliary variables X in area i. An estimate of variance (2.6) is then $v(\hat{\bar{Y}}_i^{reg} \mid n_i) = (1-\hat{\rho}_i^2)(1-f_i)\, s_i^2/n_i$.

From (2.6) we notice that by use of auxiliary variables, the variance is reduced by the factor $(1-\rho_i^2)$. This indicates that use of good auxiliary information, in the sense of high correlation with survey variable Y, increases the accuracy in SAE. However, the problem with the regression estimator (2.5) is that in practice the regression coefficients β_i are seldom known. Replacing β_i by its ordinary least square (OLS)

estimates $\hat{\beta}_i$ is not effective because of small sample sizes in each area *i*. See Cochran (1977) and Särndal, Swensson and Wretman (1992). A large enough sample size to support direct estimation for all areas of interest rarely exists. Budget and other constraints usually prevent drawing sufficiently large samples from each small area. Often these small areas are defined after the survey has been carried out. The problem is therefore how to produce reliable estimates of characteristics of interest for small areas and how to assess the estimation error with these small sample sizes. This sensitivity to sample size has led many researchers to refer to the theory that has been developed to overcome this problem as the theory of small area estimation (SAE). SAE is based on model-based methods. The idea is to use statistical models to link the variable of interest with auxiliary information, e.g. census and administrative data, for the small areas.

Note that in the case of a design-based estimator, the estimate produced is unique to each individual small area under consideration. The estimate is unbiased for that area, in the sense that, under repeated sampling the mean of successive estimates will tend towards the true value. In contrast, a model-based[2] estimator utilises auxiliary information to produce an estimate of the target variable that is applicable to all small areas that share similar characteristics. Thus, if two small areas have exactly the same auxiliary information, exactly the same estimate will be produced for each by the model-based procedure.

---

[2] In the model-based approach, the population is random and expectation is under the model i.e., over all possible populations drawn from an assumed model. Here, only one sample is drawn from each population but every time we generate one new population from the model. An estimator $\hat{T}$ of T is said to be model-unbiased (or ξ-unbiased) if $E_\xi(\hat{T}-T)=0$. In this approach T is also a random variable, not fixed like the design based approach, so expectation is taken for $(\hat{T}-T)$ under the model, i.e. we cannot write $E_\xi(\hat{T})=T$. However, in the design-based approach we can also write $E_p(\hat{T}-T)=0$, since T is fixed quantity in this case. Further, model based SAE methods depend on assumptions e.g., normality and these should be tested.

When the sample size for each small area is sufficiently large to give reasonably accurate estimates, the design-based direct estimator is the most desirable. However, as the sources of data are usually sample surveys designed to produce larger or higher level statistics, sample sizes for the small areas are usually small. Consequently, the associated variances of these estimators are likely to be unacceptably large. Therefore, for estimating the small areas, it is necessary to employ the estimation methods that 'borrow strength' from related areas. These estimators are often referred as the indirect estimators since they use values of survey variables (and auxiliary variables) from other small areas or times, and possibly from both. They borrow information (data) from other small areas or times (or both) by use of statistical models either based on implicit or explicit models that link related small areas through auxiliary information. This auxiliary information can be values of the variable of interest in other similar areas, values of this variable in the same area in the past, or values of other variables that are related to the variable of interest. However, the effectiveness of the approach depends on the strength of the relationship between the survey variables and the auxiliary variables, and the closeness in the behaviour of the data between different areas or over time. A good model is important but the availability of auxiliary information related to the survey variable is also crucial for small area estimation. Further, the smaller the small area sample size, the more important is the auxiliary variables. Furthermore, model diagnostics are very important for the model-based methodologies since misspecification of the model may induce bias, see Pfeffermann (2002).

2.3 Synthetic Estimators

In producing the synthetic estimates for small areas, availability of direct estimates for a set of larger domains of the population is assumed. Appropriate weights or proportions are then applied to these large population domain estimates to obtain the desired small area estimates. This class of estimators implicitly assumes that small areas which are being considered are similar, in some sense, to some larger areas which contain them and for which the reliable direct estimate is available. The synthetic estimation procedure was first used by the United States National Centre for Health Statistics (NCHS) for estimation of long and short-term physical disabilities based on the National Health Interview Survey (1968).

Over time, several definitions and descriptions of synthetic estimation have been given in the literature.

Gonzales (1973) described synthetic estimator as one in which an unbiased estimator of a large area is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the larger areas. Holt, Smith and Tomberlin (1979) defined it as the method of borrowing information from related sub-areas in order to increase the effective sample size for estimation and hence the accuracy of the resulting estimates. Pfeffermann (2002) stated that the term 'synthetic' refers to the fact that an estimator computed from a large domain is used for each of the separate areas comprising that domain, assuming that the areas are 'homogeneous' with respect to the quantity that is estimated. Thus, synthetic estimators already borrow information from other 'similar areas'.

Gonzalez and Waksberg (1973) and Levy and French (1977) developed the statistical properties of the synthetic estimator such as its variance, bias and mean squared error and methods of estimation for these parameters. Purcell and Linacre (1976) in their empirical studies developed synthetic estimates of income and work force status for Australian Census Statistical Divisions at the Australian Bureau of Census. Synthetic estimation has extensively been used and found wide acceptance because of its simplicity and intuitive appeal. However, at the same time it was recognised that it is a crude method for SAE and needs further improvement.

In synthetic estimation (a scale down approach), we assume availability of reliable direct estimates $\hat{T}_{y.g} = \sum_{i=1}^{m} \hat{T}_{y_{ig}}$ for the totals of larger group or class $g(g = 1,....,G)$ that encompass the small areas $i(i = 1,....,m)$ for a given survey, where $\hat{T}_{y_{ig}}$ is the estimate of population total ($T_{y_{ig}} = \sum_{j=1}^{N_{ig}} y_{jig}$) of Y in the $(i,g)^{th}$ cell with population of size N_{ig}. Here y_{jig} is the value of unit $j(j = 1,....,N_{ig})$ for variable of interest Y in the cell (i,g). From the available estimates for population $\hat{T}_{y.g}$, estimates of population means for group g are obtained as $\hat{\bar{Y}}_{.g} = \left(\sum_{i=1}^{m} \hat{T}_{y_{ig}}\right) / \left(\sum_{i=1}^{m} N_{ig}\right) = \hat{T}_{y.g} / N_{.g}$. A suitable auxiliary information available from a census or some other source is used to compute a series of weights or proportions w_{ig} such that $\sum_{g} w_{ig} = 1$. The weights w_{ig} are then applied to the group means to derive the synthetic estimator for the i^{th} small area mean $\bar{Y}_i$ as $\hat{\bar{Y}}_i^{syn} = \sum_{g=1}^{G} w_{ig} \hat{\bar{Y}}_{.g}$. This estimator is referred to as the design-based synthetic estimator. See Gonzales and Hoza (1978). Skinner (1993) referred to this approach as simple synthetic estimation.

The synthetic estimator proposed by Gonzales and Hoza (1978) and elaborated by Holt, Smith and Tomberlin (1979) uses the weights w_{ig} from the census or some other sources of accurate information. They suggested the weights based on the population size and assumed that the population size N_{ig} and weights $w_{ig} = N_{ig} / N_{i.}$, $N_{i.} = \sum_i N_{ig}$ with $\sum_g w_{ig} = 1$, are known from a previous census or some other source. Then the synthetic estimator for the mean of Y in small area i is $\hat{\bar{Y}}_i^{syn} = \sum_{g=1}^{G} \left(N_{ig} / N_{i.} \right) \hat{\bar{Y}}_{.g}$. Purcell and Kish (1979) and Ghosh and Rao (1994) propose a different series of weights $w_{ig} = \hat{T}_{x_{ig}} / \hat{T}_{x_{.g}}$ such that $\sum_i w_{ig} = 1$ but $\sum_g w_{ig} \neq 1$. Here $\hat{T}_{x_{ig}} = \sum_{j=1}^{N_{ig}} x_{jig}$ and $\hat{T}_{x_{.g}} = \sum_{i=1}^{m} \hat{T}_{x_{ig}}$ are the estimates of population total of X in cell (i, g) and the totals of X in larger group g respectively. The synthetic estimator of the mean of Y for small area i is $\hat{\bar{Y}}_i^{syn} = \sum_{g=1}^{G} \left(\hat{T}_{x_{ig}} / \hat{T}_{x_{.g}} \right) \hat{\bar{Y}}_{.g}$.

Rao and Choudry (1995) suggested the use of a ratio synthetic estimator, a modification of the earlier method used by Gonzales and Hoza (1978). The ratio synthetic estimator for the population total of Y in small area i is $\hat{T}_{y_i}^{synR} = \hat{R}_i T_{x_i}$. They assumed that area i population ratios $R_i = T_{y_i} / T_{x_i}$, $T_{y_i} = \sum_{j=1}^{N_i} y_j$ and $T_{x_i} = \sum_{j=1}^{N_i} x_j$ respectively being the population total of the characteristic of interest Y and covariate X for the i^{th} small area, are homogeneous. Thus, $R_i = R_U = T_y / T_x$, where R_U, T_y and T_x are the values for the whole population. Here R_U is estimated by $\hat{R}_U = \bar{y} / \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the overall sample means. We use a subscript of U to denote the population level quantities.

The design-variance (or p-variance) of a synthetic estimator $\hat{T}_{y_i}^{syn}$ of the population total of Y in small area i (of order O($1/n$)) will be small relative to the p-variance of a direct estimator $\hat{T}_{y_i}^{d}$ (of order O($1/n_i$)) because it depends on the precision of direct estimators at a large area level. This variance can be estimated using standard design-based methods but it is more difficult to estimate the MSE of $\hat{T}_{y_i}^{syn}$ because it is hard to estimate the bias. See Ghosh and Rao (1994). The mean squared error (MSE) of design-based synthetic estimators $\hat{T}_{y_i}^{syn}$ for the population total of Y in small area i is

$$MSE_p(\hat{T}_{y_i}^{syn}) = E_p(\hat{T}_{y_i}^{syn} - \hat{T}_{y_i}^{d})^2 - Var_p(\hat{T}_{y_i}^{syn} - \hat{T}_{y_i}^{d}) + Var_p(\hat{T}_{y_i}^{syn}) \tag{2.7}$$

where $\hat{T}_{y_i}^{d}$ is a design unbiased direct estimator for the i^{th} small area population total of Y and subscript of p denotes the operation under the design, see Rao, 2003, page 52. Under the assumption of $Cov_p(\hat{T}_{y_i}^{d}, \hat{T}_{y_i}^{syn}) = 0$, an approximately design-unbiased (or p-unbiased) estimator of (2.7) is

$$\begin{aligned} mse(\hat{T}_{y_i}^{syn}) &= (\hat{T}_{y_i}^{syn} - \hat{T}_{y_i}^{d})^2 - v(\hat{T}_{y_i}^{syn} - \hat{T}_{y_i}^{d}) + v(\hat{T}_{y_i}^{syn}) \\ &\approx (\hat{T}_{y_i}^{syn} - \hat{T}_{y_i}^{d})^2 - v(\hat{T}_{y_i}^{d}) \end{aligned} \tag{2.8}$$

where $v(\hat{T}_{y_i}^{d})$ is a design-unbiased estimator of $Var_p(\hat{T}_{y_i}^{d})$ [†]. The variance $Var_p(\hat{T}_{y_i}^{d})$ can be readily estimated by $v(\hat{T}_{y_i}^{d})$, but it is difficult to estimate bias of $\hat{T}_{y_i}^{syn}$. This MSE estimator is approximately p-unbiased, but is very unstable and can take negative values (since $v(\hat{T}_{y_i}^{syn}) << v(\hat{T}_{y_i}^{d})$). Consequently, it is customary to average these estimators over different small areas belonging to large area to obtain a global estimator of $MSE_p(\hat{T}_{y_i}^{syn})$ (Gonzalez, 1973). This average MSE estimator is expected to be stable, but it is not an area-specific measure of accuracy (Rao, 2003, chapter 4).

[†] The variance estimator $v(\hat{T})$ is design-unbiased (or p-unbiased) for $Var(\hat{T})$ if $E_p[v(\hat{T})] \equiv V_p(\hat{T})$.

We now turn to model-based synthetic estimation. Let us consider the regression model of the form

$$y_{ij} = x_{ij}'\beta + e_{ij} \tag{2.9}$$

where y_{ij} is value of variable of interest for the $j^{th}(j=1,...,n_i)$ unit in the small area $i(i=1, ..,m)$ and x_{ij} is the $p\times 1$ vector of auxiliary variables, β is a $p\times 1$ vector of regression coefficients. The error term e_{ij} is often assumed to be normally distributed with mean zero and variance σ^2. With this notation, and under model (2.9), two indirect estimators for small areas are defined.

The regression synthetic estimator for the mean of Y in small area i is defined as

$$\hat{\bar{Y}}_i^{SynREG} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'\hat{\beta} \tag{2.10}$$

where $\bar{x}_i = n_i^{-1}\sum_{j\in s_i} x_j$ and $\bar{X}_i = N_i^{-1}\sum_{j\in U_i} x_j$ are the sample and population means for the auxiliary variables X in area i. Here $\hat{\beta}$ is the full sample estimate, i.e. calculated using data from entire areas. The regression synthetic estimator (2.10) uses the same value of $\hat{\beta}$ in all small areas and thus the different from direct regression estimator (2.5). However, the regression synthetic estimator (2.10) can be calculated only when a small area has sample data.

For the areas with no sample data, the model-based synthetic estimator of the mean of Y for small area i, $\bar{Y}_i$ is defined as

$$\hat{\bar{Y}}_i^{MSyn} = \bar{X}_i'\hat{\beta} \tag{2.11}$$

The estimator (2.11) will be very efficient when small area i does not exhibit strong individual effect with respect to the regression coefficient. For a single auxiliary

variable (under model (2.9) with no intercept), the estimator (2.11) is the same as the ratio-synthetic estimator $\hat{\bar{Y}}_i^{SynR} = \bar{X}_i(\hat{\bar{Y}}_U / \hat{\bar{X}}_U) = \bar{X}_i(\bar{y}/\bar{x})$, where $\hat{\bar{Y}}_U$ and $\hat{\bar{X}}_U$ are the estimators for the population total of Y and X respectively.

The model (2.9) uses unit level auxiliary information at small area level, but one can use the area-level regression models when only at small area level auxiliary information is available. See Skinner (1993). Erickson (1974) applied the area level regression methods for the estimation of local area population change. This approach has been referred as 'the sample regression method' in Purcell and Kish (1979). Holt, Smith and Tomberlin (1979) incorporating the implicit assumption of the synthetic estimators $\hat{\bar{Y}}_i^{syn}$ derived the modified synthetic estimator under a simple one-way fixed effect analysis of variance model, referred as the prediction-synthetic estimators. For $N_{ig} >> n_{ig}$, this estimator leads to a design-based synthetic estimator. Laake (1979) showed that the variance of the prediction synthetic estimator is smaller than that of the design-based synthetic estimator.

Synthetic estimation, apart from the ease of calculation, addresses the issue of the small sample size by borrowing the strength from larger areas, and has prominent advantage due to its variance reduction. However, it can sometimes lead to severe bias if the assumption of homogeneity within the larger domain is violated or the structure of the population changed since the previous census. Also, unless the grouping variables are highly correlated with the variable of interest, the synthetic estimators fail to account for local factors.

2.4 Composite Estimators

Gonzalez and Waksberg (1973) and Schaible, Brock and Schnack (1977) compared the synthetic and design-based direct estimator for small areas and concluded that when area sample sizes are relatively small the synthetic estimator outperformed the simple direct, whereas, when the sample sizes are large the direct estimator outperformed the synthetic. Thus, as the sample size in a small area increases, a direct estimator becomes more desirable than a synthetic estimator. This is true whether or not the sample was designed to produce estimates for small areas. These results motivated the use of a weighted sum of direct estimator (with small or no bias but larger variance) and synthetic estimator (with small variance but possibly large bias) as a desirable alternative than choosing one over the other. This weighted estimator is termed as the composite estimator.

The composite estimators are of interest because they permit trade-off among the advantages and disadvantages of direct and synthetic estimators through their weighted combination. In fact, many estimators both design-based and model-based referred to by different terminology can also be regarded as composite estimators. See for example Battese, Harter and Fuller (1988).

In general, the composite estimator for the population total of Y in small area i is defined as

$$\hat{T}_{y_i}^{c} = \phi_i \hat{T}_{y_i}^{d} + (1-\phi_i)\hat{T}_{y_i}^{syn} \tag{2.12}$$

where $\hat{T}_{y_i}^{d}$ is the direct estimator and $\hat{T}_{y_i}^{syn}$ is the synthetic estimator for the population total of Y for small area i, and ϕ_i $(0 \leq \phi_i \leq 1)$ is a suitably chosen weight. The estimator (2.12), a weighted sum of two component estimators can have a mean squared error (MSE) smaller than that of either component estimator when an appropriate weighting scheme is used. However, deriving the optimal weighing has generally been a challenging problem in SAE since these estimators are surprisingly sensitive to poor estimates of the optimum weight. Ideally, the weights should be selected as to minimise the MSE but this is problematic since the MSE of the synthetic estimator is generally unknown because of its bias (Pfeffermann, 2002).

Several methods of weight selection have been proposed in the literature. Schaible (1978) assigned the weights of each component proportional to the inverse of its MSE and then the two component weights normalised so that they sum to unity. Purcell and Kish (1979) suggested the use of a common weight which minimizes the average MSE. However, use of a common weight is not recommended when the individual variances vary considerably. Drew, Singh and Choudry (1982) proposed a sample size dependent (SSD) estimator that has the form of the composite estimator with ratio type direct and ratio type synthetic estimator, with simple weights, dependent on domain counts. An alternative estimator termed as the "Dampened regression estimator" was suggested by Särndal and Hidiroglou (1989). Lui and Cumberland (1989, 1991) proposed a model-based approach to derive the optimal weight. See Ghosh and Rao (1994), Marker (1999), Rao (2003) and Schaible, (1978) for several possible weight choices proposed in the literature of SAE.

2.5 Mixed Models in Small Area Estimation

The traditional indirect estimators such as synthetic and composite commonly lead to more efficient estimators than the corresponding design-based direct estimator for small areas through the use of the implicit models which 'borrow strength' across the areas. These models assume that all the areas of interest behave similarly with respect to the variable of interest and do not take into account the area specific variability. However, in the situation where the validity of the assumed model fails, it leads to a biased estimator. That is area specific variability typically remains even after accounting for the auxiliary information. This limitation is handled by an alternative estimation technique based on an explicit linking model, which provides a better approach to SAE by incorporating random area-specific effects that account for the between area variation beyond that is explained by auxiliary variables included in the model. An area effect indicates how different one area is from another after allowing for differences in their auxiliary variable distributions. Estimating the effect for a particular area requires using data from all areas and not just the data from the particular area and thus increases the effective sample size for that area (this is known as borrowing strength across the areas). Consequently, the estimators based on such models are more efficient than traditional indirect estimators. The mixed effect model based SAE has received a considerable importance in the last two decades due to a number of advantages. These methods make specific allowance for local variation through complex error structures, models can be validated from the sample data and methods can handle complex cases such as cross-sectional, time series and multivariate data. Note that use of these model dependent methods overcomes the

problems encountered with design-based methods but at the expense of making further assumptions that need to be tested carefully.

Several methods for SAE based on the nested error regression model (Battese, Harter and Fuller, 1988), the random regression coefficients model (Dempster, Rubin and Tsutakawa, 1981) and the simple random effects model (Fay and Herriot, 1979) as special cases of the mixed model have been proposed in the literature. The estimators based on such models, include empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) estimators. Based on the level of auxiliary information available and utilised, two types of random effects model for SAE are described in the literature. The area level random effect model which uses area-specific auxiliary information (Fay and Herriot, 1979) and unit level random effect model which uses the unit level auxiliary information (Battese, Harter and Fuller, 1988). These are special cases of the linear mixed model, usually referred as area level and unit level small area models. See Pfefferman (2002), Rao (1999, 2003) and Saei and Chambers (2003).

2.5.1 Unit Level Random Effect Models

Battese, Harter and Fuller (1988) proposed and applied a nested error unit level regression model in the context of predicting mean acreage under corn and soybean crops in 12 counties (small areas) of the state of Iowa in the United States using LANDSAT satellite data in conjunction with survey data. Their model is of the form

$$y_{ij} = x_{ij}'\beta + u_i + e_{ij} \tag{2.13}$$

where as in (2.9) y_{ij} denotes the value of variable of interest for $j^{th}(j=1,....,n_i)$ sampled unit in area $i(i=1,....,m)$, x_{ij} is a $p\times 1$ vector of unit level auxiliary variables, β is a $p\times 1$ vector of the unknown fixed effects, n_i is the number of sample units in area i, u_i is the area specific random effect associated with area i with mean zero and variance σ_u^2, and e_{ij} is individual level random error with mean zero and variance σ_e^2. The two error terms are mutually independent. The random error u_i represents the joint effect of small areas that are not accounted for by the auxiliary variables, also known as the model error for area i. The normality of u_i and e_{ij} is often assumed. The model (2.13) assumes that samples are drawn independently across small areas according to a specified sampling design so sample design within small areas is ignorable. The model (2.13) also holds for non-sampled units and for the whole population, in the other words model (2.13) applies with n_i replaced by N_i.

In matrix notation, model (2.13) is expressed as

$$Y_i = X_i\beta + u_i 1_{n_i} + e_i \tag{2.14}$$

where $Y_i = (y_{i1},...., y_{in_i})'$, $X_i = (x_{i1},...., x_{in_i})'$ is a $n_i \times p$ matrix and $e_i = (e_{i1},..., e_{in_i})'$. The covariance matrix of Y_i is $Var(Y_i) = V_i = \sigma_e^2 I_{n_i} + \sigma_u^2 1_{n_i} 1_{n_i}'$, which depends on a vector of fixed parameters $\theta = (\sigma_u^2, \sigma_e^2)$, usually called the variance components of the model. Here 1_{n_i} is the unit vector of length n_i and I_{n_i} is a identity matrix of order n_i. The model (2.13) is also referred as a random intercept model since with $x_{ij1} = 1$ and $\beta_1 = \alpha$, we can write $\alpha_i = \alpha + u_i$ as the random intercept.

Assuming model (2.13) holds, population mean of the survey variable Y in area i is $\bar{Y}_i = \bar{X}_i'\beta + u_i + \bar{e}_i$, where $\bar{X}_i = N_i^{-1}\sum_{j=1}^{N_i} x_j$, is assumed to be known. For sufficiently large N_i, $\bar{e}_i = N_i^{-1}\sum_{j=1}^{N_i} e_j \approx 0$ and then mean of the survey variable Y in small area i is approximated by $\mu_i = \bar{X}_i'\beta + u_i = E(\bar{Y}_i \mid \bar{X}_i, u_i)$. This involves the prediction of the sum of a known linear function of unknown fixed parameters and unbiased random effects u_i. This is a special problem in predicting a linear combination of fixed effects and a realised value of random effects. There are a variety of approaches that deal with the estimation problem in mixed models, see Harville (1977), Henderson (1975), Kackar and Harville (1984) and Peixoto and Harville (1986).

2.5.1.1 Empirical Best Linear Unbiased Predictor

For known $\theta = (\sigma_u^2, \sigma_e^2)$, under model (2.13), following the proposal of Henderson (1975) the best linear unbiased predictor (BLUP) for the mean of Y for small area i, $\bar{Y}_i$ (Rao, 2003, chapter 7 page 141 and Royall, 1976) is

$$\begin{aligned}\tilde{\bar{Y}}_i^{BLUP} &= N_i^{-1}\left[\sum_{j\in s_i} y_j + \sum_{j\in r_i}(x_j'\tilde{\beta} + \tilde{u}_i)\right] \\ &= f_i\bar{y}_i + (1-f_i)\left\{\bar{X}_{ir}'\tilde{\beta} + \tilde{u}_i\right\} \\ &= f_i\bar{y}_i + (1-f_i)\left\{\bar{X}_{ir}'\tilde{\beta} + \gamma_i(\bar{y}_i - \bar{x}_i'\tilde{\beta})\right\}\end{aligned} \tag{2.15}$$

where s_i and r_i denote the sample and non-sample part of the population respectively in small area i, $f_i = n_i/N_i$ are sampling fractions, $\bar{y}_i$ and $\bar{x}_i$ are the sample means of y and x for small area i, $\bar{X}_{ir} = (N_i\bar{X}_i - n_i\bar{x}_i)/(N_i - n_i)$, is the mean of x for $(N_i - n_i)$ non-sampled units for small area i, $\tilde{\beta} = \left(\sum_i X_i'V_i^{-1}X_i\right)^{-1}\left(\sum_i X_i'V_i^{-1}Y_i\right)$ is the best linear unbiased estimate (BLUE) of β and $\gamma_i = \sigma_u^2\left(\sigma_u^2 + n_i^{-1}\sigma_e^2\right)^{-1}$.We can also obtain the results (2.15) from the general result given by theorem 3.1 in chapter 3.

For sufficiently large N_i, $f_i = (n_i/N_i) \to 0$ and then the approximate BLUP of the mean of Y for small area i, $\bar{Y}_i$ is given by

$$\tilde{\mu}_i = \bar{X}_i'\tilde{\beta} + \gamma_i(\bar{y}_i - \bar{x}_i'\tilde{\beta}) = \gamma_i\left\{\bar{y}_i + (\bar{X}_i - \bar{x}_i)'\tilde{\beta}\right\} + (1-\gamma_i)\bar{X}_i'\tilde{\beta}. \tag{2.16}$$

The weight γ_i $(0 \le \gamma_i \le 1)$ called 'shrinkage factor', provides a trade off between the approximately design-unbiased regression estimator (2.10) and the synthetic estimator (2.11) and measures the model variance σ_u^2 relative to total variance $\left(\sigma_u^2 + n_i^{-1}\sigma_e^2\right)$. For a small value of σ_u^2, weight γ_i will be small and consequently the synthetic part in (2.16) get more weight and vice versa. For $n_i = 0$, i.e. areas with no samples, $\gamma_i \to 0$ and $\tilde{\mu}_i = \bar{X}_i'\tilde{\beta}$. For large n_i, i.e. as n_i increases, $\gamma_i \to 1$ and then it tend to regression estimator.

Further, $\tilde{\beta}$ and $\tilde{\mu}_i$ depends on variance components θ that define the covariance matrix $V_i = \sigma_e^2 I_{n_i} + \sigma_u^2 1_{n_i} 1_{n_i}'$. In practice the variance components are unknown and estimated from sample data using standard method of estimation such as ANOVA, maximum likelihood (ML) or restricted maximum likelihood (REML) methods of estimation (Harville, 1977). We use "hat" to denote an estimate and then, a two stage estimators known as the empirical best linear unbiased predictor (EBLUP) of the mean of Y for small area i is

$$\hat{\bar{Y}}_i^{EBLUP} = f_i\bar{y}_i + (1-f_i)\left\{\bar{X}_{ir}'\hat{\beta} + \hat{\gamma}_i(\bar{y}_i - \bar{x}_i'\hat{\beta})\right\} \tag{2.17}$$

where $\hat{\gamma}_i$ and $\hat{\beta}$ are the estimates of γ_i and $\tilde{\beta}$ respectively obtained by replacing θ by $\hat{\theta}$.

For sufficiently large N_i, the approximate EBLUP of the mean of Y for small area i is

$$\hat{\mu}_i = \bar{X}_i'\hat{\beta} + \hat{\gamma}_i(\bar{y}_i - \bar{x}_i'\hat{\beta}) . \tag{2.18}$$

Note that the EBLUP given in (2.18) is an approximate EBLUP for sufficiently large population sizes. In finite population sampling the EBLUP for the mean $\bar{Y}_i$ of area i is given by (2.17).

Another popular application of a mixed effect model to the small area problem is provided by Fay and Herriot (1979) in the context of estimating per capita income for small places (population less than 1000) from the 1970 census of population and housing in the United States. The proposed model is known as the Fay-Herriot model in the literature. In this model auxiliary information are assumed to be available at the area level. Prasad and Rao (1990) working on this model, using BLUP concepts, showed that Fay and Herriot's estimator is a combination of direct survey estimator and regression estimator at area level. An advantage of the area model is that the survey weights are accounted for through the direct estimators while this is not the case for unit level model. Further, in the EBLUP approach for SAE, normality of random errors is not needed for the point estimation, but it is assumed for getting accurate MSE estimate. However, the MSE estimator for the Fay-Herriot model remains valid under non-normality of random effects (Prasad and Rao, 1990 and Lahiri and Rao, 1995).

2.5.1.2 Mean Squared Error of EBLUP

The mean squared error (MSE) of the EBLUP is evaluated to observe the variability in the estimator, but no closed form of MSE exists except in some special cases. Thus,

MSE estimation has got lot of attention in the SAE literature in recent years. Here we describe some approximations for the MSE of the EBLUP proposed in the literature. For analytical simplicity, we start with the MSE of an approximate EBLUP (2.18) and then we write down MSE for the EBLUP (2.17).

For known $\theta = (\sigma_u^2, \sigma_e^2)$, following Henderson (1975), the MSE of the approximate BLUP (2.16) is

$$MSE(\tilde{\mu}_i) = g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) \tag{2.19}$$

where

$g_{1i}(\sigma_u^2, \sigma_e^2) = (1-\gamma_i)\sigma_u^2 = \gamma_i(\sigma_e^2/n_i)$, and

$g_{2i}(\sigma_u^2, \sigma_e^2) = (\bar{X}_i' - \gamma_i \bar{x}_i')\left(\sum_i X_i' V_i^{-1} X_i\right)^{-1} (\bar{X}_i' - \gamma_i \bar{x}_i')'$.

Here $g_{1i}(\sigma_u^2, \sigma_e^2)$ is the leading term in (2.19) whereas in MSE of the simple regression estimator leading term is σ_e^2/n_i. This shows that the BLUP is superior to the simple regression estimator in terms of MSE if the shrinkage factor γ_i is small. This first term $g_{1i}(\sigma_u^2, \sigma_e^2)$ in (2.19) shows the variability of the BLUP (2.16) when all the parameters are known and is of order $o(1)$. The second term $g_{2i}(\sigma_u^2, \sigma_e^2)$ due to estimating the fixed effects β is of order $o(m^{-1})$ for large m. See Henderson (1975).

The MSE of the BLUP (2.16), $\tilde{\bar{Y}}_i^{BLUP}$ is evaluated as

$$MSE(\tilde{\bar{Y}}_i^{BLUP}) = (1-f_i)^2 \left\{MSE(\tilde{\mu}_i^*)\right\} + N_i^{-1}(1-f_i)\sigma_e^2 \tag{2.20}$$

where $MSE(\tilde{\mu}_i^*)$ is same as $MSE(\tilde{\mu}_i)$ except that $\bar{X}_i$ replaced by $\bar{X}_{ir}$ in $g_{2i}(\sigma_u^2, \sigma_e^2)$ and denoted by $g_{2i}^*(\sigma_u^2, \sigma_e^2) = (\bar{X}_{ir}' - \gamma_i \bar{x}_i')\left(\sum_i X_i' V_i^{-1} X_i\right)^{-1} (\bar{X}_{ir}' - \gamma_i \bar{x}_i')'$.

The näive approximation to the estimate of MSE of EBLUP $\hat{\mu}_i$ is obtained by replacing $\theta = (\sigma_u^2, \sigma_e^2)$ by $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ in (2.19) as

$$mse_{näive}(\hat{\mu}_i) = g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \tag{2.21}$$

However, this approximation to MSE seriously underestimates the true MSE because the BLUP assumes known variances and hence the MSE estimator obtained by replacing the unknown variances by their sample estimates $\hat{\theta}$ fails to account for the error resulting from variance estimation.

Kackar and Harville (1984) proposed the correction for this underestimation in the MSE estimator. The MSE of the approximate EBLUP (2.18), $\hat{\mu}_i$ is

$$\begin{aligned} MSE(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2 &= E(\tilde{\mu}_i - \mu_i)^2 + E(\hat{\mu}_i - \tilde{\mu}_i)^2 \\ &\quad + 2E\{(\tilde{\mu}_i - \mu_i)(\hat{\mu}_i - \tilde{\mu}_i)\} \\ &= MSE(BLUP) + E(\hat{\mu}_i - \tilde{\mu}_i)^2 + 2E\{(\tilde{\mu}_i - \mu_i)(\hat{\mu}_i - \tilde{\mu}_i)\} \end{aligned} \tag{2.22}$$

The cross-product term in (2.22) vanishes under the assumption of translation invariance of $\hat{\theta}$ and normality of two errors terms. An approximation to the second term on the right-hand side of (2.22) using the "Delta method" is

$$E(\hat{\mu}_i - \tilde{\mu}_i)^2 \cong E\left\{d_i(\theta)'(\hat{\theta} - \theta)^2\right\} \text{ with } d_i(\theta) = \frac{\partial \tilde{\mu}_i}{\partial \theta}$$

Under the approximate independence of $\hat{\theta}$ and $d_i(\theta)$, they proposed a further approximation of this term as

$$E(\hat{\mu}_i - \tilde{\mu}_i)^2 \cong tr\left[A(\theta)E\left\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)'\right\}\right] \tag{2.23}$$

where $A(\theta) = Var[d_i(\theta)]$ and $E\{(\hat{\theta}-\theta)(\hat{\theta}-\theta)'\} = Var(\hat{\theta})$ is the asymptotic covariance matrix of $\hat{\theta}$. Note that $\hat{\theta}$ is estimated from all data in all small areas whereas computation of $d_i(\theta) = \frac{\partial \tilde{\mu}_i}{\partial \theta}$ is based on the data in only a single small area *i*.

Prasad and Rao (1990) justified the approximate independence of $\hat{\theta}$ and $d_i(\theta)$, and concluded that for the method of fitting constants (MFC) estimate $\hat{\theta}$ of θ (also called Henderson's Method-III) , the neglected term in the Kackar-Harville approximation is order of $o(m^{-1})$, which is of smaller order than the order of the term retained. They proposed a further approximation to the second term of (2.22) as

$$E(\hat{\mu}_i - \tilde{\mu}_i)^2 \cong tr\left[\left(\frac{\partial b_i'}{\partial \theta}\right) V_i \left(\frac{\partial b_i'}{\partial \theta}\right)' Var(\hat{\theta})\right] = g_{3i}(\sigma_u^2, \sigma_e^2) \tag{2.24}$$

where $b_i' = \gamma_i$. They used the well known method of fitting of constant (MFC) to estimate σ_u^2 and σ_e^2. Neglected terms in (2.24) are of the lower order.

Bringing together these approaches, the Prasad-Rao mean squared error approximation for the approximate EBLUP (2.18) is

$$MSE(\hat{\mu}_i) \approx g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) + g_{3i}(\sigma_u^2, \sigma_e^2) \tag{2.25}$$

with bias of order $o(m^{-1})$, where *m* is the number of small areas. Similarly, the mean squared error of the EBLUP $\hat{\bar{Y}}_i^{EBLUP}$ is

$$MSE(\hat{\bar{Y}}_i^{EBLUP}) = (1-f_i)^2 MSE(\hat{\mu}_i^*) + N_i^{-1}(1-f_i)\sigma_e^2 \tag{2.26}$$

where $MSE(\hat{\mu}_i^*)$ is obtained from $MSE(\hat{\mu}_i)$ replacing $g_{2i}(\sigma_u^2, \sigma_e^2)$ by $g_{2i}^*(\sigma_u^2, \sigma_e^2)$.

Prasad and Rao (1990) proposed an approximately model-unbiased estimator for the mean squared error (2.25) as

$$mse(\hat{\mu}_i) \approx g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \tag{2.27}$$

where $g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$, $g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ are obtained from $g_{1i}(\sigma_u^2, \sigma_e^2)$, $g_{2i}(\sigma_u^2, \sigma_e^2)$ and $g_{3i}(\sigma_u^2, \sigma_e^2)$ respectively, replacing $\theta = (\sigma_u^2, \sigma_e^2)$ by $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)$. The order of the bias being $o(1/m)$ since $g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ have biases of order $o(1/m)$. This is an approximately model unbiased estimator in the sense that its bias is of order $o(m^{-1})$ and therefore considered as a second order approximation. This estimator is valid for both the MFC and REML method of estimating variances under certain regularity conditions and under the normality of random errors u_i and e_{ij}, but not for the maximum likelihood (ML) estimator (Datta and Lahiri, 2000).

Datta and Lahiri (2000) derive the estimator for mean squared error of the EBLUP when ML estimates $\hat{\theta}_{ML} = (\hat{\sigma}_{u,ML}^2, \hat{\sigma}_{e,ML}^2)$ of variance components $\theta = (\sigma_u^2, \sigma_e^2)$ are used. Their expression for mean squared error estimate includes one extra term for bias correction that arises due to the use of ML estimates given as

$$mse(\hat{\mu}_i) \approx g_{1i}(\hat{\theta}_{ML}) + g_{2i}(\hat{\theta}_{ML}) + 2g_{3i}(\hat{\theta}_{ML}) - B_i'(\hat{\theta}_{ML})\nabla g_{1i}(\hat{\theta}_{ML}) \tag{2.28}$$

where $\nabla g_{1i}(\hat{\theta}_{ML})$ is the first order derivative of $g_{1i}(\theta_{ML})$ with respect to θ at $\theta = \hat{\theta}_{ML}$, and $B_i'(\hat{\theta}_{ML}) = \frac{1}{2m}\left\{I^{-1}(\hat{\theta}_{ML}) col_{1\le j\le m} tr\left[\left(\sum_i X_i' V_i^{-1} X_i\right)^{-1}\left(\sum_i X_i' V_i^{(j)} X_i\right)\right]\right\}$ is the bias in estimating the variances of $\hat{\theta}_{ML}$, with $V_i^{(j)} = \partial V_i^{-1}/\partial\theta_j = -V_i^{-1}\left(\partial V_i/\partial\theta_j\right)V_i^{-1}$ and $I^{-1}(\hat{\theta}_{ML})$ is the inverse of information matrix $I(\hat{\theta}_{ML})$.

The estimator for the MSE of the EBLUP $\hat{\bar{Y}}_i^{EBLUP}$ given in (2.26) is expressed as

$$mse(\hat{\bar{Y}}_i^{EBLUP}) = (1-f_i)^2 mse(\hat{\mu}_i^*) + N_i^{-1}(1-f_i)\hat{\sigma}_e^2 \tag{2.29}$$

where $mse(\hat{\mu}_i^*)$ is obtained from $mse(\hat{\mu}_i)$ by replacing $g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ by $g_{2i}^*(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$.

2.6 Pseudo-EBLUP

The model (2.13) assumes that samples are drawn independently across areas according to a specified sampling design such that the sample design within small areas is ignorable or alternatively selection bias is absent. The estimation based on such models do not make use of unit level survey weights and the corresponding estimators are not design consistent unless the sampling design is self weighting within small areas (Prasad and Rao, 1999). In contrast, the design-based direct estimators are design consistent but fail to borrow strength from the related areas. In recent years, some methods proposed in the literature make use of survey weights in model-based small area estimation.

Kott (1989) proposed a design consistent estimator, also model unbiased under the simple random effect model with the same assumption of random errors as in (2.13). He showed that this estimator is robust with respect to model failure under certain conditions and derived an estimator of mean squared error without including the random effect component. Empirical results show the mean squared error estimates are quite unstable and even take negative values. Consequently, this approach cannot be used to the compare proposed design-consistent small area estimator and the conventional design-based direct estimator.

Prasad and Rao (1999) proposed a model assisted estimator for small area estimation called the *pseudo* empirical best linear unbiased predictor (pseudo-EBLUP), which depends on the survey weights and remains design consistent as the sample sizes in the small areas increased. Using the Prasad and Rao (1990) approach they also derived the mean squared error of this estimator. Their results indicate this estimator of mean squared error performs well even under moderate deviations of the linking model and often more stable than Kott (1989). The pseudo-EBLUP approach is described below.

As noted earlier, the EBLUP (2.18) does not depend on the unit level survey weights, w_{ij} attached to $y_{ij}(j=1,...,n_i;\ i=1,....,m)$, so that design consistency as the sample size n_i increases is forsaken except when the design is self weighting within areas (i.e. $w_{ij}=w_i$). A design-based direct estimator for the mean of Y in area i, $\bar{y}_{iw}=\sum_{j\in s_i}\tilde{w}_{ij}y_{ij}$, with $\tilde{w}_{ij}=w_{ij}/\sum_{j\in s_i}w_{ij}$, uses sampling weights and is therefore design consistent but fails to borrow strength. Under the model (2.13), let us define

$$\bar{y}_{iw}=\sum_{j\in s_i}\tilde{w}_j y_j=\sum_{j\in s_i}\tilde{w}_j(x_j'\beta+u_i+e_j)=\bar{x}_{iw}'\beta+u_i+\bar{e}_{iw} \tag{2.30}$$

with $E(\bar{e}_{iw})=0$ and $Var(\bar{e}_{iw})=\sigma_e^2\left(\sum_{j\in s_i}\tilde{w}_j^2\right)=\sigma_e^2\delta_i^2$, where $\hat{\bar{x}}_{iw}=\sum_{j\in s_i}\tilde{w}_j x_j$. This is an aggregated (survey-weighted) area-level model, equivalent to the well known area level Fay-Herriot model (Fay and Herriot, 1979) so the usual results of this model are applicable. Assuming an aggregated area level model (2.30) holds, for given σ_u^2 and σ_e^2, the BLUP for the mean of Y for small area i (Prasad and Rao, 1999) is

$$\tilde{\mu}_{iw}=\bar{X}_i'\tilde{\beta}_w+\gamma_{iw}(\bar{y}_{iw}-\bar{x}_{iw}'\tilde{\beta}_w) \tag{2.31}$$

where $\gamma_{iw} = \sigma_u^2 \left(\sigma_u^2 + \sigma_e^2 \delta_i^2\right)^{-1}$ and $\tilde{\beta}_w = \left(\sum_i \gamma_{iw} \bar{x}_{iw} \bar{x}_{iw}'\right)^{-1} \left(\sum_i \gamma_{iw} \bar{x}_{iw} \bar{y}_{iw}\right)$ with $E(\tilde{\beta}_w) = \beta$ and $Var(\tilde{\beta}_w) = \sigma_u^2 \left(\sum_i \gamma_{iw} \bar{x}_{iw} \bar{x}_{iw}'\right)^{-1} \equiv \Phi_w$. Note that the BLUP $\tilde{\mu}_{iw}$ in (2.31) is different from the BLUP $\tilde{\mu}_i$ given in (2.16). The variance components $\theta = (\sigma_u^2, \sigma_e^2)$ are unknown in practice and they are estimated under the full model (2.13). Using these estimates, a weighted estimator of β is $\hat{\beta}_w = \beta_w(\hat{\theta})$ and the EBLUP for the mean of Y for small area i is

$$\hat{\mu}_{iw} = \bar{X}_i' \hat{\beta}_w + \hat{\gamma}_{iw} (\bar{y}_{iw} - \bar{x}_{iw}' \hat{\beta}_w). \tag{2.32}$$

Note that the pseudo-EBLUP given in (2.32) depends on the survey weights and satisfies the design consistency property. The estimator (2.32) is model-assisted and approximately model and design unbiased even if the sample design is nonignorable.

From Prasad and Rao (1999), the approximate MSE of the EBLUP (2.32) is

$$MSE(\hat{\mu}_{iw}) \approx g_{1iw}(\theta) + g_{2iw}(\theta) + g_{3iw}(\theta) \tag{2.33}$$

where

$$g_{1iw}(\theta) = (1 - \gamma_{iw}) \sigma_u^2,$$

$$g_{2iw}(\theta) = \sigma_u^2 (\bar{X}_i' - \gamma_{iw} \bar{x}_{iw}') \left(\sum_i \gamma_{iw} \bar{x}_{iw} \bar{x}_{iw}'\right)^{-1} (\bar{X}_i' - \gamma_{iw} \bar{x}_{iw}')', \text{ and}$$

$$g_{3iw}(\theta) \approx \gamma_{iw} (1 - \gamma_{iw})^2 \sigma_e^{-4} \sigma_u^{-2} \left\{(\sigma_u^2)^2 Var(\hat{\sigma}_e^2) + (\sigma_e^2)^2 Var(\hat{\sigma}_u^2) - 2\sigma_u^2 \sigma_e^2 Cov(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\right\}.$$

An approximately model-unbiased estimator of the MSE (2.33) is

$$mse(\hat{\mu}_{iw}) \approx g_{1iw}(\hat{\theta}) + g_{2iw}(\hat{\theta}) + 2 g_{3iw}(\hat{\theta}) \tag{2.34}$$

where $g_{1iw}(\hat{\theta})$, $g_{2iw}(\hat{\theta})$ and $g_{3iw}(\hat{\theta})$ are obtained from $g_{1iw}(\theta)$, $g_{2iw}(\theta)$ and $g_{3iw}(\theta)$ respectively by replacing θ with $\hat{\theta}$.

You and Rao (2002) indicated that $\hat{\beta}_w$ based on the aggregated area level model suffers significant efficiency loss compared to the estimates based on the unit level model. This could in turn lead to some efficiency loss in the estimation of small area means. They proposed an iterative weighted estimating equation approach to estimate β using the sampling weights w_{ij} but they used the same estimate of the variance components as used in Prasad and Rao (1999). Consequently, You and Rao (2002) suggested the modified pseudo-EBLUP. Unlike EBLUP (Prasad and Rao, 1990), the modified pseudo-EBLUP is design consistent like pseudo-EBLUP (Prasad and Rao, 1999) as n_i becomes larger. Further, this estimator satisfies the benchmarking property without any adjustment when aggregated over small areas *i*. Furthermore, they showed that their estimator benchmark to the direct survey estimator of *Y*, in contrast to the EBLUP and pseudo-EBLUP.

As noted earlier, You and Rao (2002) proposed an iterative weighting estimating equations to estimate the fixed effects. You, Rao and Kovacevic (2003) proposed an extension of You and Rao (2002). They extended this approach to estimate both the fixed effects and the variance components in a random intercept model using sampling weights. Their approach updates the estimates of the fixed effects and variance components alternatively until convergence is achieved. Their approach produces simultaneous sampling weighted estimates of fixed effects and variance components.

Militino et al (2007) applied SAE to agricultural data. In their application of SAE, they used design weights and weights to account for heteroscedasticty (they named as model weights) in the pseudo-EBLUP method. They considered weighted estimation

of the variance components and the fixed effects. These authors argued that by combining both type of weights, models can be very useful for practitioners because the within error variance heterogeneity is accounted for and design consistency is achieved, providing protection against model failures as the small area sample sizes increase. We note that although the pseudo-EBLUP uses the survey weights in SAE, implementation of the approach is not straightforward, especially for MSE estimation.

2.7 Model-Based Direct Estimators

Chambers (2005) introduced the calibrated weighting based approach in SAE and defined the model-based direct (MBD) estimator for small area means. This approach uses the calibrated sample weights derived under a population level version of the linear mixed model to define weighted linear small area estimators as well as a simple expression for the MSE. In contrast to design-based direct estimators, these estimators borrow strength from other areas via the linear mixed model used in defining the weights. There are many practical advantages associated with this approach, arising from the fact that the estimators are computed as weighted linear combinations of the actual sample data from the small areas of interest. Perhaps the most important of these are the simplicity of both the estimation process and the estimation of the MSE. Further, the MBD estimator is easy to interpret and to build into a survey processing system. This motivates the use of the MBD approach in SAE. Consequently, in this thesis we study the MBD approach to SAE outlined in Chambers (2005) and proposed several extensions of his work. The next chapter of the thesis is devoted to the MBD method of small area estimation.

2.8 Extension of Mixed Models in Small Area Estimation

Stukel and Rao (1999) considered a two-way nested error regression model to derive the EBLUP and associated approximately unbiased second order MSE estimator, appropriate for two-stage sampling within small areas. Some further extensions to this model include a multi-level extension in which regression coefficients are assumed to be random and depend on area level auxiliary information (Moura and Holt, 1999) and multivariate models (Kleffe and Rao, 1992 and Datta et al, 1999).

The models considered so far assume that the random area effects are independent between areas, but in practice, it would be reasonable to assume that area effects associated with neighbouring areas by some distance measure (not necessarily geographical) are correlated, and correlation decays to zero as distance increases. Such models are very common in spatial analysis (Cressie, 1993), but are not in wide use in SAE. An improvement in the EBLUP method can be achieved by including spatial structure in the random area effects. See Petrucci and Salvati (2004) and Pratesi and Salvati (2005) for the spatial-EBLUP approach in SAE. Petrucci, Pratesi and Salvati (2005) described SAE under spatially correlated random area effects model using geographic information. Chambers, Pratesi, Salvati and Tzavidis (2006) considered spatially correlated random effects model and defined the spatial M-quantile method of SAE. Pfefferman (2002) noticed that the loss in efficiency from using a model with independent area effect is small unless the correlations between the areas are large. There is a drawback in the spatial model since it depends on how the neighbourhoods are defined which introduces some subjectivity (Marshall, 1991 in Rao, 2003).

As noted earlier, in order to increase the overall sample size in SAE, we borrow the information from other data sets. This information can be borrowed from 'similar' areas or from a previous occasion. In the time series modelling approach, we exploit information in data over time (e.g., repeated surveys) in order to obtain further improvement in efficiency of estimators. In general, empirical studies show that small area estimates that draw upon information across time are more efficient than those that drawn upon information across area since the time series data usually represent the same information about the target variable from the past. Related references are Pfeffermann et al (1998), Pfeffermann and Burck (1990), Tiller (1992), Ghosh et al (1998) and Datta et al (1999).

Sometimes cross sectional and times series data are combined to obtain further improvement in efficiency of the small area estimators. In general, empirical studies show that for repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. See Rao and You (1994), Datta et al (2002) and Rao (2003). Singh et al (2005) used spatial-temporal models in small area estimation. They used spatial models for exploitation of spatial auto-correlation amongst the small area units and a spatial temporal model fitted via Kalman filtering for the time series data. Chambers and Tzavidis (2006) introduced the M-quantile approach to SAE whereas Pratesi et al (2006) considered the nonparametric M-quantile small area estimation via penalized splines.

2.9 Summary

In this chapter we summarized several SAE methods proposed in the literature. The merits and limitations of each approach is discussed. The SAE method based on mixed models, particularly the unit level nested error regression model, is discussed in detail. We notice that the EBLUP based approaches (Prasad and Rao, 1990) are the most popular model-based approach under the unit level random effect model. However, these approaches do not use the unit level survey weights. The pseudo-EBLUP approach (Prasad and Rao, 1999) proposed in the literature uses survey weights, but is complicated to work with, particularly with respect to MSE estimation. The MBD approach of the Chambers (2005) uses calibrated sample weights and the small area estimator is a weighted linear estimator with a simple MSE expression. The motivation of this approach lies in its simplicity. In the next chapter of the thesis we return with details on the MBD approach of small area estimation.

CHAPTER 3

MODEL BASED DIRECT ESTIMATION FOR SMALL AREAS

3.1 Introduction

Unit level random effect models are often used in small area estimation (SAE). The empirical best linear unbiased prediction (EBLUP, Prasad and Rao, 1990) is then the widely used approach for the estimation of small areas under such models. However, this approach does not lead to small area estimators that are a weighted linear function of the sample data from these areas. As a result, several practical advantage of using such weighted estimators are lost, with probably the most important being the relative simplicity of their mean squared error estimation. The calibrated weighting based approach introduced in Chambers (2005) overcomes some of these limitations. This approach uses calibrated sample weights derived from a population level version of the linear mixed effects model to define weighted linear small area estimators and a simple expression for their mean squared error. The associated small area estimators are the model-based direct (MBD) estimators because they depend on area specific sample data. However, the sample weights defining the MBD estimator are function of the data from the entire sample. Therefore, this method 'borrows strength' from other areas via the mixed model that defines the weights. Hereafter we refer to this approach as the MBD method of small area estimation.

In this chapter we evaluate the empirical performance of EBLUP and MBD methods of SAE. Our empirical evidence is based upon data from Australian broadacre farms that participated in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) in the late 1980s. We also examine the robustness of these two methods under wrong model choices. In addition, we study some properties of these two methods of SAE. The rest of the chapter is organised as below.

In the following section we describe the calibrated weighting approach in survey sampling for population estimation. We elaborate both the design-based (Deville and Särndal, 1992) and the model-based (Chambers, 2005) perspective of calibration weighting. In section 3.3 we illustrate the sample weights derived under a linear mixed model for SAE. Then we define the EBLUP and MBD estimators for small area means and their corresponding mean squared error estimators. Empirical results are reported and discussed in section 3.4. Finally, in section 3.5 we present some concluding remarks and further extensions of the MBD methods of SAE.

3.2 Calibrated Sample Weighting for Population Estimation

In this section we briefly review calibrated sample weighting for estimation of population level quantities. Calibration is now a widely used approach for population estimation in survey sampling. This is, basically, a method to improve estimation in survey sampling when auxiliary information is available. Here auxiliary information is included at the estimation stage to produce efficient estimates. In this approach, survey weights are

modified so that known population characteristics, in practice totals (or means), are reproduced from the sample data. Therefore, for variables in the survey correlated with the auxiliary variables, higher precision estimates are obtained by these new weights. The efficiency of the estimate depends on how well the auxiliary variables explain the variability in the survey variable. Kott (2003) described calibration weighting as a methodology under which probability sample weights are adjusted in such a way that when applied to survey data they can produce model unbiased estimators for a number of different target variables.

Let Y_U denote an N-vector of population values of a characteristic Y of interest, where U denote the population of size N. Suppose that we are interested in the estimation of the population total $T_y = \sum_U y_j$ (or population mean $\bar{Y}_U = N^{-1}\sum_U y_j$) of Y. In order to assist us in this objective, we shall assume that we have 'access' to X_U, an $N \times p$ matrix of values of p auxiliary variables that are related, in some sense, to the values in Y_U. In particular, we assume that the individual sample values in X_U are known. The non-sample values in X_U may not be individually known, but are assumed known at some aggregate level. At a minimum, we know the population totals T_x of the columns of X_U. Given this set up, it is standard to estimate the total and mean of the values in Y_U by

$$\hat{T}_y^w = \sum_s w_j y_j \tag{3.1}$$

and

$$\hat{\bar{Y}}_U^w = \sum_s w_j y_j / \sum_s w_j \tag{3.2}$$

respectively. The sample weights $\{w_j; j \in s\}$ reflect the relationship between the values of Y and X, typically via some form of statistical model. Here s is a probability sample of size n from a population of size N with the probability $p(s)$. The inclusion probabilities $\pi_j = \sum_{j \in s} p(s)$ are known for all j $(j = 1, ..., N)$. Further, we assume that the design is such that $\pi_j > 0$ for all elements j. Let $d_j = \pi_j^{-1}$ denote the design weight of element j. The original idea of calibration is to modify the design weights d_j so that known totals are reproduced from the sample data. A set of calibrated sample weights is then produced. More precisely for known total T_x we calibrate by constructing new weight w_j such that

$$\hat{T}_x^w = \sum_{j \in s} w_j x_j = T_x. \tag{3.3}$$

The new weights w_j are as close as possible to the old weights d_j. In other words, we want to replace old weights d_j with more efficient weights w_j determined by using available auxiliary information. There are two basic approaches proposed in the literature to construct the calibration sample weights, design-based and model-based calibration weighting, see Chambers (1996, 1997, 2005), Chambers and Skinner (1999) and Deville and Särndal (1992).

3.2.1 Design Based Calibration Weighting

Design-based calibration weighting is based on the concept of "closest" calibrated weights. Deville and Särndal (1992) first introduced the notation of a calibration

estimator. They proposed the calibration estimator for population total T_y as the linear combination of the observations, $\hat{T}_y = \sum_{j\in s} w_j y_j$, with calibration weights w_j's chosen to minimise their average distance from the basic design weights, $d_j = \pi_j^{-1}$, that are used by the Horvitz-Thompson estimator, $\hat{T}_y^{HT} = \sum_{j\in s} d_j y_j$. Here minimization of average distance is subject to the calibration constraint $\sum_{j\in s} w_j x_j = \hat{T}_x^w = T_x$. Alternative distance measures can also be used. See Deville and Särndal (1992). All resulting estimators are asymptotically equivalent to the one obtained from minimising the chi-squared distance function:

$$Q_s = \sum_{j\in s} (w_j - d_j)^2 / d_j q_j \tag{3.4}$$

where w_j's are known positive weights unrelated to d_j and q_j's are constants. The existence of initial design weights d_j is assumed and these are the inverse of inclusion probabilities of the sample units. These weights do not always have to be the inverse of inclusion probabilities (Chambers, 1996). Minimisation of quadratic distance measure (3.4) leads to new set of weights called calibrated weights:

$$w_j = d_j(1 + q_j x_j' \lambda) \tag{3.5}$$

where $\lambda = T_s^{-1}(T_x - \hat{T}_x^{HT})$ and $\hat{T}_x^{HT} = \sum_{j\in s} d_j x_j$ is the Horvitz-Thompson (HT) estimator for the population total of X. Here existence of the inverse of $T_s = \sum_{j\in s} d_j q_j x_j x_j'$ is assumed. Using the calibrated sample weight (3.5) in (3.1), the calibration estimator of population total T_y is

$$\hat{T}_y^w = \sum_{j\in s} w_j y_j = \hat{T}_y^{HT} + (T_x - \hat{T}_x^{HT})'\hat{B} \tag{3.6}$$

where $\hat{B} = T_s^{-1}(\sum_{i \in s} d_j q_j x_j y_j)$. The calibration estimator given by (3.6) is equivalent to a generalized regression (GREG) estimator, which is derived as model assisted estimator assuming a linear regression model, with variance structure provided by the diagonal matrix with elements $(1/q_j)$. See Deville and Särndal (1992) for examples on the role of the constants given by the q_j's.

In matrix notation, we denote the set of initial weights by $d = \{d_j; j \in s\}$ and then we find a set of calibrated sample weights w that minimises the quadratic distance measure $Q = (w - d)'\Omega(w - d)$, where Ω is a known positive definite matrix. The minimisation of quadratic distance measure, Q subject to (3.3) leads to sample weights of the form

$$w_\Omega(d) = d + H_\Omega'(X_U' 1_N - X_s' d) \tag{3.7}$$

with $H_\Omega = (X_s'\Omega^{-1}X_s)^{-1}X_s'\Omega^{-1}$ and 1_N is a vectors of 1's of order N. This is the design-based interpretation of the calibration approach. The model-based perspective of the calibration approach is described as below.

3.2.2 Model Based Calibration Weighting

In model-based calibration we assume that survey variable Y and auxiliary variable X are related by some model and then the calibrated sample weights are derived under the model to satisfy the calibration constraints (3.3). Let us assume that Y_U and X_U are related by the linear regression model

$$Y_U = X_U\beta + \varepsilon_U \tag{3.8}$$

where β is a $p\times 1$ vector of unknown regression parameters, ε_U is random error vector of dimension N with $E(\varepsilon_U)=0$ and $Var(\varepsilon_U)=\sigma^2 V_U$, where V_U is a known positive definite matrix of order N and σ^2 is some constant. Without loss of generality, we arrange the vector Y_U so that its first n elements correspond to the sample units. We can then partition Y_U, X_U and V_U according to sample and non-sample units as

$$Y_U = \begin{bmatrix} Y_s \\ Y_r \end{bmatrix},\ X_U = \begin{bmatrix} X_s \\ X_r \end{bmatrix} \text{ and } V = \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix}.$$

Here Y_s is the $n\times 1$ vector defined by the sample values in Y_U, X_s is the corresponding $n\times p$ matrix of sample values of the auxiliary variable and V_{ss} is the $n\times n$ component of V associated with Y_s. A subscript of r is used to denote corresponding quantities defined by the $N-n$ non-sample units, e.g. V_{rs} is the $(N-n)\times n$ matrix defined by $Cov(Y_r, Y_s)=\sigma^2 V_{rs}$. We denote 1_N, 1_n and 1_{N-n} as vectors of 1's and I_N, I_n and I_{N-n} as identity matrices of order N, n and $N-n$ respectively.

Given this set-up, and assuming (3.8) holds, the **B**est **L**inear **U**nbiased **P**redictor (BLUP) of population total of Y can be derived from the following Theorem (Royall, 1976).

Theorem 3.1 Among linear prediction unbiased estimators $\hat{T}_y$ of T_y satisfying $E(\hat{T}_y - T_y)=0$, the error variance $E(\hat{T}_y - T_y)^2$ is minimised by

$$\hat{T}_y = 1_n' Y_s + 1_{N-n}' \left\{ X_r\hat{\beta} + V_{rs}V_{ss}^{-1}(Y_s - X_s\hat{\beta}) \right\}, \text{ where } \hat{\beta} = (X_s' V_{ss}^{-1} X_s)^{-1} X_s' V_{ss}^{-1} Y_s.$$

Then error variance of $\hat{T}_y$ is

$$Var(\hat{T}_y - T_y) = 1'_{N-n}(V_{rr} - V_{rs}V_{ss}^{-1}V_{rs}')1_{N-n} + \\ +1'_{N-n}(X_r - V_{rs}V_{ss}^{-1}X_s)(X_s'V_{ss}^{-1}X_s)^{-1}(X_r - V_{rs}V_{ss}^{-1}X_s)'1_{N-n}$$

When sample and non-sample units are uncorrelated (i.e. $V_{rs} = 0$) the BLUP of T_y is obtained simply adding to the sample sum the BLUP $1'_{N-n}X_r\hat{\beta}$ of the expected value of the non-sample sum $T_{y_r} = 1'_{N-n}Y_r$. Further, under a special case of model (3.8), the BLUP of small area mean of Y given in (2.15) can be derived from this result. See section 2.5.1.1.

Proof: Proof of this theorem is given in Royall (1976).

Given this result, it can be seen that the BLUP of population total of Y is given by (3.1) with weights defined by

$$w_{BLUP} = 1_n + H'_{BLUP}\left(X'_U 1_N - X'_s 1_n\right) + \left(I_n - H'_{BLUP}X'_s\right)V_{ss}^{-1}V_{sr}1_{N-n} \tag{3.9}$$

of order $O(Nn^{-1})$. Here $H_{BLUP} = (X_s'V_{ss}^{-1}X_s)^{-1}X_s'V_{ss}^{-1}$. These are the BLUP weights and calibrated on X_U, in the sense that they exactly reproduce the known population totals defined by the columns of X_U. That is $X_s'w_{BLUP} = X'_U 1_N = T_x$. Further, these weights define an unbiased predictor of T_y since

$$E(\hat{T}_y^w - T_y) = E(w'_{BLUP}Y_s - 1'_N Y_U) = E(w'_{BLUP}X_s - 1'_N X_U)\beta = 0. \tag{3.10}$$

Furthermore, any linear estimator with weights that are calibrated on X_U will be unbiased under (3.8), and conversely, any linear estimator that is unbiased under (3.8) will have weights that are calibrated on X_U (Chambers, 2005).

The weights (3.9) implicitly rely on the assumption that the survey variable Y and the auxiliary variables X are linearly related. However, if the underlying regression model is non-linear then these weights can be inefficient. For example, if the variable Y and X are not linear on themselves but they are linear on some transform scale (e.g. in case of skewed data), then the weights (3.9) based on linear model lead to inefficient estimates. In these situations, Wu and Sitter (2001) proposed a model calibration approach as a generalisation of the calibration procedure under a general model. We shall discuss this approach in chapter 5 in context of small area estimation for skewed data.

The reasons for calibration vary. There is the largely intuitive argument that such weights, because they are 'perfect' for key known population quantities, should be good for estimating other population quantities for which only sample data are available. In other words, estimates are 'consistent' with known information. The consistency means that the calibrated weights reproduce exactly the known population total for each auxiliary variable. Further, the variance of a calibrated estimator tends to decrease as more variables and their known totals are brought into the calibration. In fact, the more auxiliary totals we use in the calibration, the 'better' we expect the resulting weight system to be. However, one of the serious problems of the approach is the negative weights which sometimes appear. Modification is possible at the expense of a more complicated procedure. Huang and Fuller (1978), Bardsley and Chambers (1984) and Chambers (1996) described the methods for dealing with negative weights. Park and Fuller (2005) discussed the procedures for constructing the non-negative weights in

which initial weights are the inverse of the approximate conditional inclusion probabilities.

3.3 Small Area Estimation Based on a Linear Mixed Model

3.3.1 The Small Area Models

A commonly used class of models in small area inference is the class of linear mixed models. Let Y_i be the $N_i \times 1$ vector of values of variable of interest in small area i and let X_i be the $N_i \times p$ matrix of associated values of the auxiliary variables. Here a subscript of i denotes restriction to small area i. We consider the following linear mixed model for the distribution of Y_i given X_i:

$$Y_i = X_i\beta + Z_i u_i + e_i . \tag{3.11}$$

Here N_i is the number of the population units in small area i, β is a $p \times 1$ vector of fixed effects, Z_i is a $N_i \times q$ matrix of known covariates characterising differences between small areas, u_i is a $q \times 1$ random area effect associated with the i^{th} small area and e_i is a $N_i \times 1$ vector of individual level random errors. Normality of these two random variables is often assumed. The random vectors u_i and e_i are assumed to be independently distributed, with zero means and with variances $Var(u_i) = \Sigma$ and $Var(e_i) = \sigma_e^2 I_{N_i}$ respectively. The covariance matrix of Y_i is $Var(Y_i) = V_i = \sigma_e^2 I_{N_i} + Z_i \Sigma Z_i'$, depends on a vector of parameters $\theta = (\sigma_e^2, \Sigma)$ usually called the variance components of the model.

Finally, it is usually assumed that sampling is uninformative given the values of the auxiliary variables, so the sample data also follow the population model (3.11).

By aggregating the area-specific models (3.11) over the m small areas, we are led to the population level model

$$Y_U = X_U\beta + Z_U u + e \tag{3.12}$$

where $Y_U = (Y_1', \ldots\ldots, Y_m')'$, $X_U = (X_1', \ldots\ldots, X_m')'$, $Z_U = diag(Z_i; 1 \le i \le m)$, $u = (u_1', \ldots, u_m')'$ and $e = (e_1', \ldots\ldots, e_m')'$. Under (3.12), the covariance matrix of Y_U is $V_U = diag(V_i; 1 \le i \le m)$. This is the general linear mixed model. This model includes most of the small area models used in the literature (Rao, 2003, page 107). As mentioned after equation (3.8) we again consider the sample and non-sample decomposition of Y_U, X_U, Z_U and V_U. We use similar notation at the small area level by introducing an extra subscript i to denote small area. For example, we denote by s_i the set of n_i sample units in area i, r_i the corresponding $N_i - n_i$ non-sampled units in the area and put $V_{iss} = \sigma_e^2 I_{n_i} + Z_{is}\Sigma Z_{is}'$ and $V_{isr} = Z_{is}\Sigma Z_{ir}'$.

In practice the variance components that define V_U are unknown and must be estimated from the sample data using suitable estimation methods such as Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) or methods of moment. We use a "hat" to denote an estimate and put $\hat{V}_U = diag(\hat{V}_i; 1 \le i \le m)$, with $\hat{V}_i = \hat{\sigma}_e^2 I_{N_i} + Z_i\hat{\Sigma}Z_i'$. Empirical studies presented in section 3.4 uses both ML and REML methods for the estimation of variance component parameters.

3.3.2 Sample Weights for Small Area Estimation

The sample weights (3.9) are typically based on models for 'population level variability' and small area effects are assumed to average out over the population. This assumption fails when population level weights are used for small area estimation since small area effects do not average out at small area level. That is sample weights (3.9) are appropriate for estimation of population level quantities, while using these weights for small area estimation can lead to inefficient estimates for small area level quantities. Consequently some form of local weighting is required if weighted estimators are going to be used for small area estimates - i.e. weights must differentiate between the small areas that make up the population. The most common class of models that includes random area effects (i.e. differentiate between areas) are the mixed effect models. In this section, we describe the sample weights (3.9) which are derived via a linear mixed model suitable for small area estimation.

Under the population level version of the linear mixed model (3.12), the sample weights (3.9) that define the BLUP for the population total of Y are

$$w_{BLUP} = 1_n + H'_{BLUP}\left(X'_U 1_N - X'_s 1_n\right) + \left(I_n - H'_{BLUP} X'_s\right) V_{ss}^{-1} V_{sr} 1_{N-n} \tag{3.13}$$

where $H_{BLUP} = \left(X'_s V_{ss}^{-1} X_s\right)^{-1} X'_s V_{ss}^{-1} = \left(\sum_{i=1}^{m} X'_{is} V_{iss}^{-1} X_{is}\right)^{-1} \left(\sum_{i=1}^{m} X'_{is} V_{iss}^{-1}\right)$. The weights (3.13) are special case of the weights (3.9) and so calibrated on the same population level quantities.

Replacing the estimates of unknown variance components in (3.13), the empirical version of the BLUP weights (3.13) that define the EBLUP for the population total of Y are

$$w_{EBLUP} = 1_n + H'_{EBLUP}\left(X'_U 1_N - X'_s 1_n\right) + \left(I_n - H'_{EBLUP} X'_s\right)\hat{V}_{ss}^{-1}\hat{V}_{sr} 1_{N-n} \tag{3.14}$$

where $H_{EBLUP} = \left(X'_s \hat{V}_{ss}^{-1} X_s\right)^{-1} X'_s \hat{V}_{ss}^{-1} = \left(\sum_{i=1}^{m} X'_{is} \hat{V}_{iss}^{-1} X_{is}\right)^{-1} \left(\sum_{i=1}^{m} X'_{is} \hat{V}_{iss}^{-1}\right)$. The EBLUP weights (3.14) are the special case of weights (3.10) and so they are calibrated on X_U, i.e. $X'_s w_{EBLUP} = X'_U 1_N$ and define an unbiased linear predictor of the population total of Y (Royall, 1976). Furthermore, since they only depend on the random area effects structure of the mixed model (3.12) via the covariance structure in the sample/population, extension to more complex covariance structures (e.g. spatial correlation between population units) only requires $\hat{V}_{ss}^{-1}$ and $\hat{V}_{sr}$ to be computed under these more complex models. We do not pursue this extension in this thesis however.

3.3.2.1 *Calibrated Weighting Based Estimator for Small Areas*

The model-based direct (MBD) estimator of the mean of Y for small area i, $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_j$ is the direct estimator of this quantity based on the EBLUP weights (3.14). That is, it is defined as

$$\hat{\bar{Y}}_i^{MBD} = \sum_{s_i} w_j y_j \Big/ \sum_{s_i} w_j \tag{3.15}$$

where the weights used in (3.15) are those associated with the sample units in small area i in (3.14). Note that we refer to (3.15) as a direct estimator because it is a weighted mean of the sample data from the small area of interest. However, this does not mean that it can

be calculated just using these data. The EBLUP sample weights (3.14) will be a function of the data from the entire sample. That is, they 'borrow strength' from other areas through the model (3.12).

3.3.2.2 *Estimation of Mean Squared Error*

An important consideration in small area estimation is estimation of the mean squared error (MSE) of the small area estimators. We can easily adapt straightforward methods of MSE estimation for population level estimators to estimation of the MSE of (3.15). Well known results (see Royall and Cumberland, 1978 and Valliant et al, 2000, chapter 5) indicate that robust MSE estimators are of the form $Var(\hat{\bar{Y}}_U) = \sum_{j \in s} w_j^2 (y_j - \hat{y}_j)^2$ $+lower\ \ order\ terms$, where $\hat{y}_j$ denotes the fitted value for y_j under the linear model implied by the calibration constraints.

In order to estimate the MSE of (3.15), we note that the implied population level model (3.12) includes random area effects and so one needs to consider whether it is appropriate to condition on these effects u_i when estimating this MSE. For example, the rather complicated MSE estimator of the EBLUP does involve this conditioning (Prasad and Rao, 1990). On the other hand, estimation of the MSE of (3.15) is straightforward if we do not condition on random area effects, treat the EBLUP weights (3.14) as fixed and use standard methods for estimating the MSE of a weighted linear estimator of a domain mean under the population model (3.8). See Royall and Cumberland (1978). The choice between these two approaches is largely philosophical and depends on how much one

'believes' the linear mixed model (3.12). In particular, here we treat this model as a vehicle for generating estimation weights, but then base inference on (3.8), which is consistent with the way MSEs are estimated at population level. Following Royall and Cumberland (1978) and Chambers (2005), we can write the prediction variance for the area i weighted mean (3.15) as

$$Var(\hat{\bar{Y}}_i^{MBD} - \bar{Y}_i) = Var\left\{\left(\sum_{s_i} w_j\right)^{-1}\left(\sum_{s_i} w_j y_j\right) - N_i^{-1}\left(\sum_{s_i} y_j + \sum_{r_i} y_j\right)\right\}$$

$$\approx N_i^{-2}\left(\sum_{j\in s_i} a_j^2 Var(y_j) + \sum_{j\in r_i} Var(y_j)\right) \qquad (3.16)$$

where $a_j = \left(N_i w_j - \sum_{g\in s_i} w_g\right)\Big/\left(\sum_{g\in s_i} w_g\right)$.

A robust model-based estimate of prediction variance (3.16) is obtained by substituting the squared residual $(y_j - x_j'\hat{\beta})^2$ for $Var(y_j)$ in the first (leading) term on the right hand side of (3.15). If these squared sample residuals are also used to estimate the second term, the resulting estimator of (3.16) is

$$v(\hat{\bar{Y}}_i^{MBD}) = \sum_{s_i} \lambda_j (y_j - x_j'\hat{\beta})^2 \qquad (3.17)$$

where $\lambda_j = N_i^{-2}\left[a_j^2 + (N_i - n_i)/(n_i - 1)\right]$. Using (3.17) to estimate the prediction MSE of $\hat{\bar{Y}}_i^{MBD}$ implicitly assumes that this weighted mean is unbiased for $\bar{Y}_i$. However, this is not generally the case, since $E(\hat{\bar{Y}}_i^{MBD} - \bar{Y}_i) \approx (\hat{\bar{X}}_i^{MBD} - \bar{X}_i)'\beta$ under (3.12), where $\hat{\bar{X}}_i^{MBD}$ denotes the weighted average of the sample values of the auxiliary variables in area i. Further, calibration on X ensures that this term vanishes at population level, but not necessarily at small area level. In other words, this bias correction arises due to fact that

the sample weights used to define the MBD estimator are not locally calibrated at area level. The result (C.5) in appendix C presents some explicit expression for this bias under a special case of model (3.11). The magnitude and order of this bias in result (C.5) clearly shows this bias cannot be ignored. A simple estimate of this bias is

$$b(\hat{\bar{Y}}_i^{MBD}) = (\hat{\bar{X}}_i^{MBD} - \bar{X}_i)'\hat{\beta}. \tag{3.18}$$

A robust† estimator of the mean squared error of (3.15) is therefore

$$mse(\hat{\bar{Y}}_i^{MBD}) = v(\hat{\bar{Y}}_i^{MBD}) + \left\{b(\hat{\bar{Y}}_i^{MBD})\right\}^2. \tag{3.19}$$

Obviously, one could alternatively 'bias correct' $\hat{\bar{Y}}_i^{MBD}$ directly using $b(\hat{\bar{Y}}_i^{MBD})$. However, this is not recommended since this correction increases the variability of our estimator much more than it reduces its bias. Using it in (3.19) is a more conservative, and safer, approach. Further, use of the square of the unbiased estimator (3.18) of the bias of $\hat{\bar{Y}}_i^{MBD}$ in the MSE estimator (3.19) can be criticised because this term is not itself unbiased for the squared bias term in the MSE. This can be corrected by replacing by $\left\{b(\hat{\bar{Y}}_i^{MBD})\right\}^2$ by $\left\{b(\hat{\bar{Y}}_i^{MBD})\right\}^2 - \hat{Var}\left\{b(\hat{\bar{Y}}_i^{MBD})\right\}$ in (3.19), where $\hat{Var}\left\{b(\hat{\bar{Y}}_i^{MBD})\right\}$ is the estimator of the variance of (3.18). However, small area sample sizes may lead to (3.19) becoming quite unstable, and thus it is preferable to use (3.19) with square of (3.18). The MSE estimation for linear predictors for domains described in Chambers, Chandra and Tzavidis (2007) shows the estimator (3.19) is consistent for the MSE of the MBD estimator (3.15).

† The estimator (3.17) is called a robust model-based estimator because it does not depend on the second order moments assumptions and thus robust to misspecification of the second order moment of the working model. Consequently we referred (3.19) as a robust MSE estimator

3.3.3 Empirical Best Linear Unbiased Predictor

With the above notation, and assuming (3.11) holds, the EBLUP for the mean of Y for small area i, $\bar{Y}_i$ is

$$\hat{\bar{Y}}_i^{EBLUP} = f_i \bar{y}_{is} + (1-f_i)\left\{\bar{X}_{ir}'\hat{\beta} + \bar{Z}_{ir}'\hat{\Sigma}\, Z_{is}'\hat{V}_{iss}^{-1}(Y_{is} - X_{is}'\hat{\beta})\right\} \tag{3.20}$$

where $\hat{\beta} = \left(\sum_{i=1}^{m} X_{is}'\, \hat{V}_{iss}^{-1} X_{is}\right)^{-1}\left(\sum_{i=1}^{m} X_{is}'\, \hat{V}_{iss}^{-1} Y_{is}\right)$, $f_i = n_i/N_i$ is sampling fraction (assumed to be non-negligible) and $\bar{X}_{ir}$ and $\bar{Z}_{ir}$ are vectors of mean values for the $N_i - n_i$ non-sampled units in small area i. In chapter 2 we defined the EBLUP (2.17) which is particular case of (3.20) when underlying model is a random intercept model, a special case of model (3.11). However, in this chapter we are dealing with a general form of the linear mixed model. Therefore, we define the EBLUP and associated MSE estimator under this model. In our empirical studies in section 3.4 we have considered four special cases of model (3.11).

An approximate mean squared error estimator for the EBLUP (3.20) is

$$MSE(\hat{\bar{Y}}_i^{EBLUP}) = (1-f_i)^2\left\{g_{1i}(\theta) + g_{2i}(\theta) + g_{3i}(\theta)\right\} + g_{4i}(\theta) \tag{3.21}$$

where

$$g_{1i}(\theta) = \bar{Z}_{ir}'(\Sigma - \hat{\Sigma} Z_{is}' V_{iss}^{-1} Z_{is}\Sigma)\bar{Z}_{ir},$$

$$g_{2i}(\theta) = \left(\bar{X}_{ir}' - b_i' X_{is}\right)\left(\sum_{i=1}^{m} X_{is}'\, V_{iss}^{-1} X_{is}\right)^{-1}\left(\bar{X}_{ir}' - b_i' X_{is}\right)',$$

$$g_{3i}(\theta) = tr\left\{(\nabla b_i')V_{iss}(\nabla b_i) Var(\hat{\theta})\right\}, \text{ and } g_{4i}(\theta) = N_i^{-1}(1-f_i)\sigma_e^2$$

with $b_i' = \bar{Z}_{ir}' \Sigma Z_{is}' V_{iss}^{-1}$, $\nabla b_i' = \partial b_i' / \partial \theta$ and $Var(\hat{\theta})$ is asymptotic covariance matrix of estimates of variance components $\hat{\theta}$. For Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimates of variance components, $Var(\hat{\theta})$ is given by the inverse of the relevant information matrix, see Rao (2003, page 107-110).

For ML estimates of variance components $\hat{\theta} = \hat{\theta}_{ML}$ the elements of the information matrix are given by

$$I_{jk}(\theta_{ML}) = \frac{1}{2} \sum_{i=1}^{m} tr\left\{ \left(V_{iss}^{-1} \frac{\partial V_{iss}}{\partial \theta_j} \right) \left(V_{iss}^{-1} \frac{\partial V_{iss}}{\partial \theta_k} \right) \right\}.$$

For REML estimates of variance components, $\hat{\theta} = \hat{\theta}_{REML}$ the elements of the information matrix are given by

$$I_{jk}(\theta_{REML}) = \frac{1}{2} \sum_{i=1}^{m} tr\left\{ \left(P_{iss} \frac{\partial V_{iss}}{\partial \theta_j} \right) \left(P_{iss} \frac{\partial V_{iss}}{\partial \theta_k} \right) \right\},$$

where $P_{iss} = \left\{ V_{iss}^{-1} - V_{iss}^{-1} X_{is} \left(X_{is}' V_{iss}^{-1} X_{is} \right)^{-1} X_{is}' V_{iss}^{-1} \right\}$. Asymptotically, $Var(\hat{\theta}_{ML}) \cong Var(\hat{\theta}_{REML})$, provided p is fixed. The neglected terms in this approximation are of the order $0(m^{-1})$, where m is the number of small areas.

An approximately unbiased estimator of the MSE of EBLUP (3.20) is

$$mse(\hat{\bar{Y}}_i^{EBLUP}) \approx (1 - f_i)^2 \left\{ g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2g_{3i}(\hat{\theta}) - B'(\hat{\theta}) \nabla g_{1i}(\hat{\theta}) \right\} + g_{4i}(\hat{\theta}) \qquad (3.22)$$

where $\nabla g_{1i}(\hat{\theta})$ is first derivative of $g_{1i}(\theta)$ with respect to θ at $\theta = \hat{\theta}$, $B'(\hat{\theta})$ is bias in estimating $\hat{\theta}$ and $B'(\hat{\theta}) \nabla g_{1i}(\hat{\theta})$ is the bias correction term. For the method of fitting constant (MFC) and REML estimates of variance components, the bias $B'(\hat{\theta})$ is zero and

consequently the bias correction in (3.22) vanishes. However, for the ML estimates of variance components, the bias is given by

$$B'(\hat{\theta}_{ML}) = \frac{1}{2m}\left\{\left(I^{-1}(\hat{\theta}_{ML})\right) \underset{1\le j\le q}{Col}\, tr\left[\left(\sum_{i=1}^{m} X_{is}' V_{iss}^{-1} X_{is}\right)^{-1}\left(\sum_{i=1}^{m} X_{is}'\left(\frac{\partial V_{iss}^{-1}}{\partial \theta_j}\right) X_{is}\right)\right]\right\} \quad (3.23)$$

where $\frac{\partial V_{iss}^{-1}}{\partial \theta_j} = -V_{iss}^{-1}\left(\frac{\partial V_{iss}}{\partial \theta_j}\right)V_{iss}^{-1}$. Further, replacing θ by $\hat{\theta}$ in $g_{1i}(\theta)$, $g_{2i}(\theta)$, $g_{3i}(\theta)$ and $g_{4i}(\theta)$ we get $g_{1i}(\hat{\theta})$, $g_{2i}(\hat{\theta})$, $g_{3i}(\hat{\theta})$ and $g_{4i}(\hat{\theta})$ respectively. See Datta and Lahiri (2000) and Prasad and Rao (1990) for further detail of the MSE estimator (3.22).

The MBD estimator (3.15) is not the same as the EBLUP (3.20) even though both sum to the same population level EBLUP. This is because there is no unique representation of (3.20) as a weighted means of the sample data values from small area *i*. Appendix C presents some analytical expressions to compare EBLUP, MBD and design-based direct (DBD) estimators. Here we show in general that the MBD (3.15) and the EBLUP (3.20) are not same. However, in certain special cases the two methods are equivalent.

A major advantage of MBD methods (3.15) is the relative simplicity of estimation. In particular, we can calculate an estimator of the mean squared error of (3.15) via a straightforward generalisation of the standard robust estimator of the prediction variance of the EBLUP of the population mean of *Y*. This is in sharp contrast to the rather complicated estimator of the conditional prediction variance of (3.20). However, this does not mean that the MBD estimator (3.15) is superior to the EBLUP (3.20). As noted earlier, both (3.15) and (3.20) sum to the population EBLUP under the linear mixed

model (3.11). Furthermore, under this model it is clear that the EBLUP must be more efficient asymptotically, since it approximates the best linear predictor when (3.11) actually holds. For example, in the special case where $X_U = Z_U = 1_N$, the weight associated with sampled unit j in area i under the MBD approach is

$$w_j = \frac{N}{n}\left\{1+\frac{1}{1+n_i\hat{\phi}}\left[(N_i-n_i)\hat{\phi}+\frac{\bar{N}-\bar{n}}{\bar{n}}\right]\right\}$$

where $\hat{\phi} = \hat{\Sigma}/\hat{\sigma}_e^2$, $\bar{N} = \sum_i N_i(1+n_i\hat{\phi})^{-1} / \sum_i (1+n_i\hat{\phi})^{-1}$ and $\bar{n}$ is defined similarly. That is, MBD (3.15) reduces to the area i sample mean. In contrast, EBLUP (3.20) is then a linear combination of the overall sample mean and the area i sample mean. Appendix C compares the weights used in EBLUP and MBD for the estimation of i^{th} small areas. These results indicate that weights for the EBLUP of small areas are $w_j^{(EBLUP)} \sim O(N_i n^{-1})$ while weights for the MBD are $w_j^{MBD} \sim O(Nn^{-1})$; $j \in s_i, i = 1,....,m$. In other words, the sample weights used to define MBD estimator are of order $O(Nn^{-1})$ and the EBLUP (3.20) is defined as the indirect linear predictor using n-vector of sample weights that are $O(N_i n^{-1})$. This indicates that variance of the EBLUP will be lower order that the MBD. Thus, we expect the EBLUP to be more efficient, if the model holds.

It is sometimes claimed that a disadvantage of any direct estimator (including the MBD estimator) is that it is not defined when there is no sample in small area i. In contrast, the EBLUP (3.20) then equals the synthetic estimator $\bar{X}_i'\hat{\beta}$. However, no sample data in an area also means that the validity of any estimator for that area is completely model-dependent. In particular, we cannot check to see if (3.11) holds. There is also the problem

that different areas are then treated unequally in estimation. Areas with sample data have their means estimated via EBLUP, while those without have their means estimated via synthetic estimators. Furthermore, in such a case the weighted average of these estimates across all small areas does not equal the EBLUP of the population mean. A standard work-around when this occurs is to rescale all the small area estimates to sum to this population estimate (or some other acceptable value). However, this is rather arbitrary. For example, if most of the small areas have no sample, then such a rescaling exercise could substantially change the final predicted value of the area i mean of Y for a 'sample area' relative to its EBLUP value (3.20), in which case one has to wonder about the efficiency of the final result.

In contrast, direct estimators like (3.15) are easy to interpret and to build into survey processing systems. Furthermore, they do not allow the prediction for areas where there are no sample data, which, in light of the discussion in the previous paragraph, may be considered to be a good thing. However, choice of the weights in this approach is very crucial and the wrong choice of the weights can result in an inefficient direct estimator. The model-based direct estimator (3.15) based on the linear mixed model (3.11) with the sample weights (3.14) for small areas that possesses some of the efficiency properties of the EBLUP (3.20) seem to be appropriate. In Appendix D we present some empirical results which contrast the efficiency of the BLUP with direct estimators for small areas. Here we consider both usual design-based direct (DBD) and the model based direct (MBD) estimators of small areas. These results show the MBD provides an improvement over a design-based direct approach and can compete with the BLUP method.

Like the EBLUP itself, the EBLUP sample weights (3.14) used in MBD estimator (3.15) are variable specific since they depend on the estimated variance components of a particular variable (i.e., estimated variance components for Y_U via the matrices $\hat{V}_{sr}$ and $\hat{V}_{ss}$) and efficient for estimation related to the variable on which they are based. This can be a limitation if a true 'multipurpose' approach to small area estimation is required. Development of 'multipurpose' weights (i.e. not variable specific) can be more useful if there is more than one response variable in a survey (which is very common in practice). In chapter 4 we shall return with details on multipurpose sample weighting.

3.4 An Empirical Study

Simulation studies use computer intensive procedures to assess the appropriateness and accuracy of a variety of statistical methods in relation to the known truth. These techniques provide empirical estimation of the sampling distribution of the parameters of interest that could not be achieved from a single study and enable the estimation of accuracy measures, such as the bias in the estimates of interest, as the truth is known. Therefore, simulation studies should be designed with similar rigour to any real data study, since the results are expected to represent the results of simultaneously performing many real studies. See for example Morgan (1984) and Lewis and Orav (1989). In this section we illustrate the design-based simulation studies using real data to contrast the performance of the MBD (3.15) and the EBLUP (3.20) methods of SAE. We also examine the robustness of these methods under model misspecifications. The results from this study are also reported in Chandra and Chambers (2005, 2006c, 2006d) and Chambers and Chandra (2006).

3.4.1 Simulated Data

Our basic data come from the same sample of 1652 Australian broadacre farms that participated in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) in the late 1980s and were used in the simulation study reported in Chambers (1996). This survey was carried out by the Australian Bureau of Agricultural and Resource Economics. Here we use these sample farms to generate a target population of 81982 farms by sampling with replacement from them with probabilities proportional to their sample weights. We then drew 1000 independent stratified random samples from this (fixed) population, with total sample size in each simulation equal to the original sample size (1652) and with strata defined by the 29 different Australian broadacre agricultural regions. Sample sizes within these strata were fixed to be the same as in the original sample. Note that these varied from a low of 6 to a high of 117, allowing an evaluation of the performance of different small area estimation methods across a range of realistic small area sample sizes. Table 3.1 shows the various parameters for this population.

As noted earlier, we considered the 29 regions as small areas. The total cash costs (A$) of the farm business over the surveyed year (TCC), is our variable of interest (y). Our aim is to estimate average total cash costs (A$) in these regions. In doing so, we used the fact that these regions can be grouped into three zones (Pastoral, Wheat-Sheep or Mixed farming, and Coastal or High rainfall), with farm size (hectares) known for each farm in the population. Figure 3.1 shows the map of these 29 farming regions (or small areas) and zones where they are located. The numbers shown in the map are the regions codes. The auxiliary variable total farm size (hectares) is referred to as Size in what follows.

Table 3.1 Regional characteristics of simulation population.

Region	Population size	Sample size	Average farm area	Average farm costs
1	79	6	297958	467964
2	115	10	55731	171414
3	189	30	359383	670926
4	330	25	178355	186984
5	388	36	108038	208142
6	465	19	16717	130316
7	604	36	131544	302583
8	729	40	21976	242836
9	737	30	23083	179112
10	964	30	23712	180467
11	1586	51	2213	116965
12	1778	62	891	114442
13	1984	55	1066	96162
14	2182	47	4398	233171
15	2607	79	1239	97839
16	2683	60	581	93202
17	2689	60	701	84790
18	2847	34	373	36979
19	3056	74	799	101101
20	3139	51	3200	87919
21	3910	73	563	78509
22	4486	117	4635	164889
23	4550	80	960	86218
24	4587	95	1862	184153
25	5368	83	1838	198156
26	5528	103	1013	105151
27	6489	108	1403	134169
28	6980	81	812	95617
29	10933	77	360	66285
Population	81982	1652	5475	118997

Figure 3.1 Map of Australian broadacre zones and farming regions.

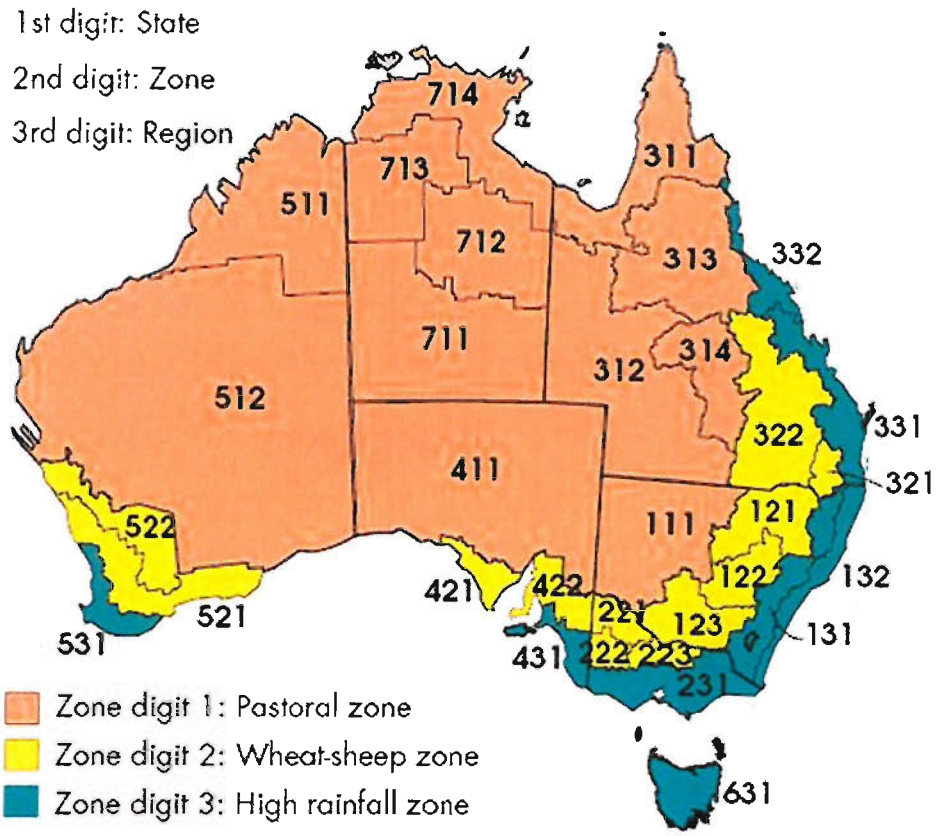
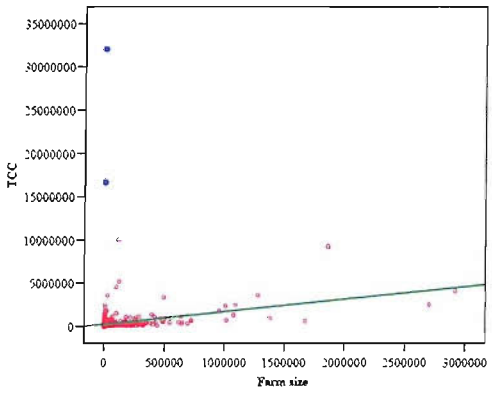
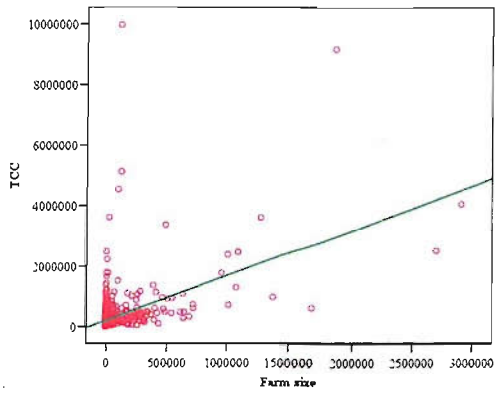


Figure 3.2 Relationship between total cash costs (TTC) and farm sizes in AAGIS data.





TCC = 227843.07 + 1.44432 Size

TCC = 197529.46 + 1.4815 Size

R^2	0.0498	0.236
Root mean square error	970358.4	410043.8
Observations	1652	1650

The overall linear relationship between the total cash costs (TCC) and Size is rather weak in the original sample data, however this improves when separate linear models are fitted within six post-strata as shown in Figure 3.2 and 3.3. These post-strata are defined by splitting each zone into small farms (farm area less than zone median) and large farms (farm area greater than or equal to zone median). These six SizeZone Strata are

1 = Pastoral zone and area of 50000 hectares or less

2 = Pastoral zone and area of more than 50000 hectares

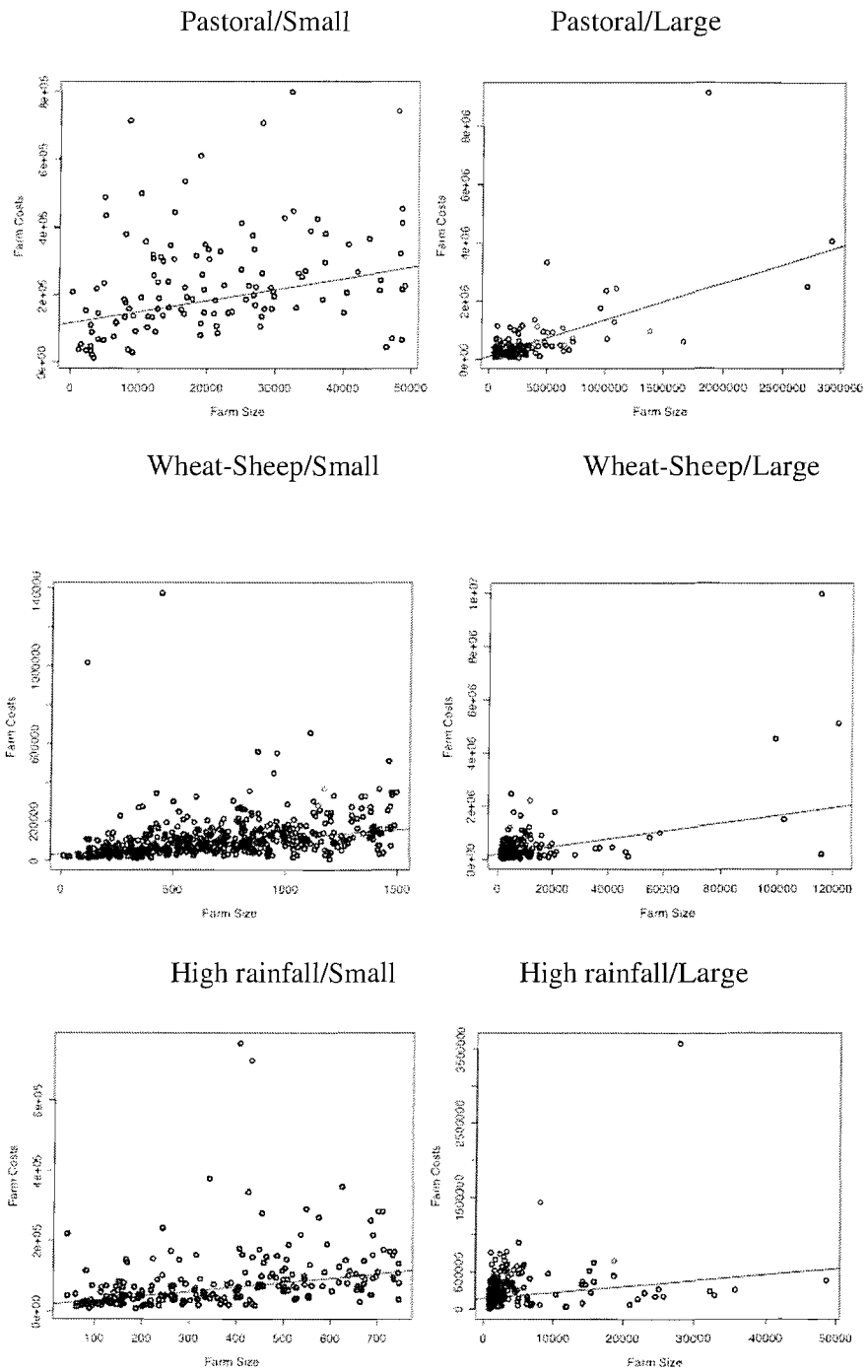
3 = Wheat-sheep zone and area of 1500 hectares or less

4 = Wheat-sheep zone and area of more than 1500 hectares

5 = High rainfall zone and area of 750 hectares or less

6 = High rainfall zone and area of more than 750 hectares

Figure 3.3 Relationship between total cash costs and farm sizes in six post-strata.



In the scatter plot of original sample data in Figure 3.2 we notice the presence of two outlier data points. The linear relationship between TCC and farm size improves if these two points are discarded from the analysis. Further, the values of R^2 (and root mean square error) increases (and decrease) from 0.05 (and 970358.4) to 0.236 (and 410043.8) if we do not include these two data points in the model fitting. In our simulation studies we include these two data points. The purpose is to see the performance of different SAE methods in presence of these outlying points. Anyway, these are the true data values.

Figure 3.4 presents the average total cash costs and average farm size in the 29 regions. Figure 3.5 illustrates the relationship between total cash costs and farm size in each of these six post-strata. This plot indicates the presence of zone effects in the data and shows that the data are extremely heteroskedastic. The matrix X of auxiliary variable values in (3.11) was then defined so as to include an effect for Size, effects for the post-strata and effects for interactions between Size and the post strata. Two different specifications for X (corresponding to whether an intercept was included or not) and two different specifications for Z (corresponding to whether a random slope on farm size was included or not) were then used to specify (3.11) and hence the EBLUP and MBD estimators based on this model. These four special cases of (3.11) are set out in Table 3.2 and shown graphically in Figure 3.6. For the farm data, models I and II are appropriate (with II fitting marginally better, see Appendix A) while models III and IV are badly specified. We use ML and REML estimates of random effects parameters, obtained via the *lme* function in R (Bates and Pinheiro, 1998). For each model, two different estimators (ML and REML) of the 29 regional means are computed, along with corresponding estimators of their mean squared error. These are the EBLUP (3.20) with MSE estimator (3.22), referred to as EBLUP below; the MBD estimator (3.15) based on sample weights (3.14) and with MSE estimator (3.19), referred to as MBD below.

Figure 3.4 Average total cash costs and average farm sizes in different regions.

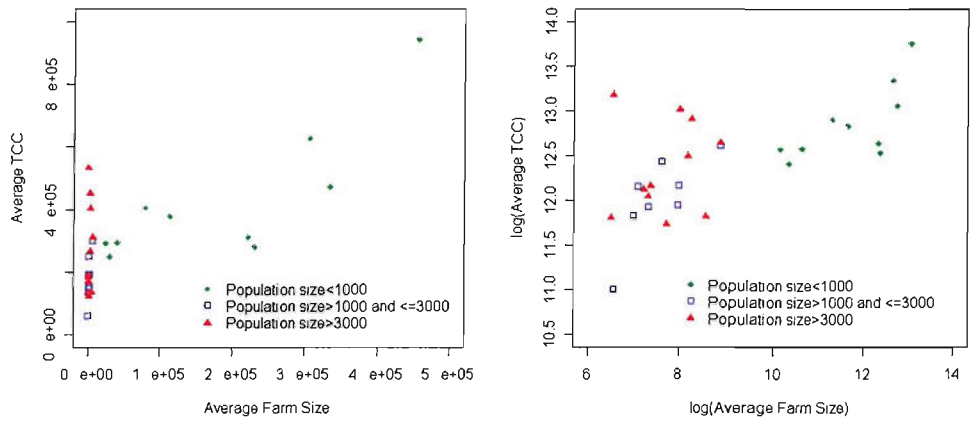


Figure 3.5 Relationship between total cash costs and farm sizes in six post-strata.

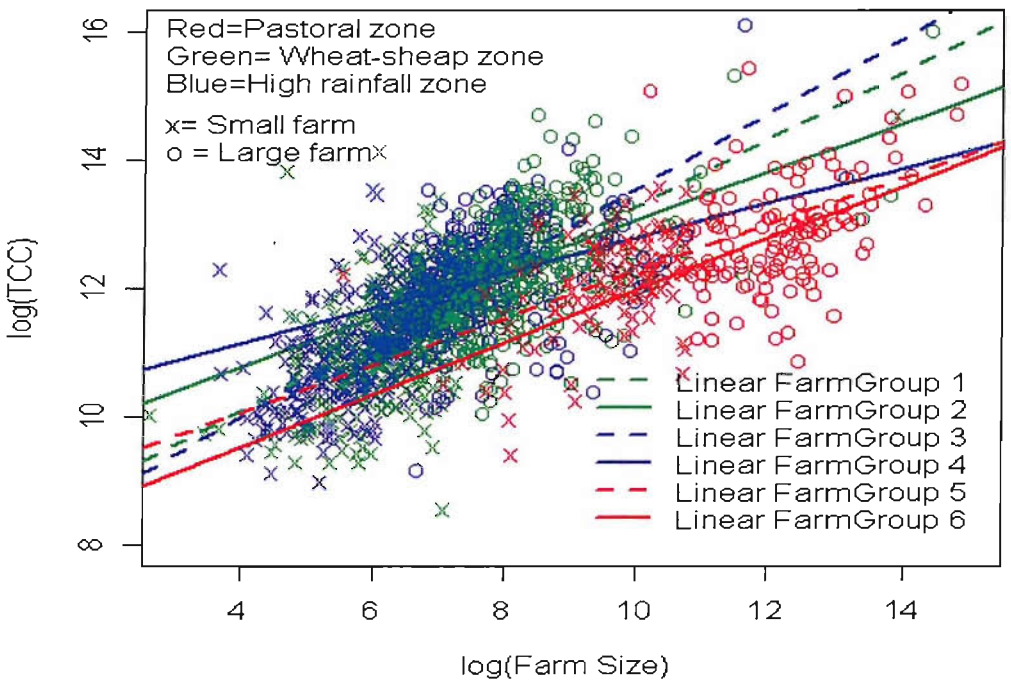
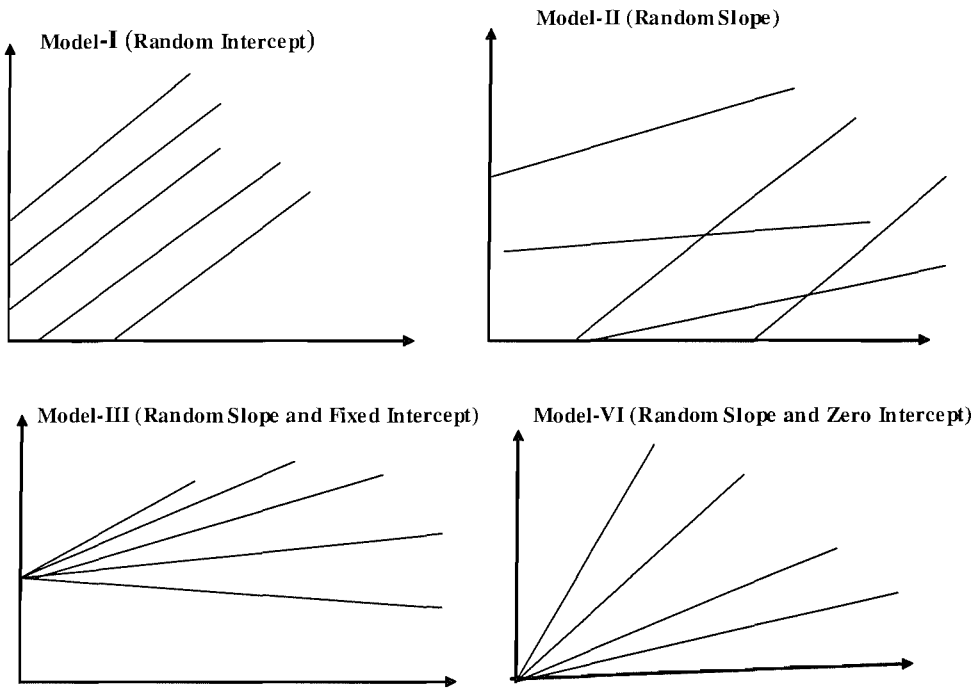


Table 3.2 Different mixed model specifications considered in the simulations.

Model	Model Type	*X*	*Z*
I	Random Intercepts	Intercept included	Intercept only
II	Random Slopes	Intercept included	Intercept + Size
III	Random Slopes with fixed intercept	Intercept included	Size only
IV	Random Slopes with zero intercept	Intercept excluded	Size only

Figure 3.6 Four different model specification considered in the simulation.



3.4.2 Performance Indicators

We use the following criteria to evaluate the performance of different methods :

- *The percentage relative bias* (RB), defined as

$$RB(\hat{T}_i) = T_i^{-1}\left(R^{-1}\sum_{r=1}^{R}\hat{T}_{i(r)} - T_i\right)\times 100,$$

where $\hat{T}_i$ is the estimator (e.g. of the mean) for the $i^{th}(i = 1,.....,m)$ small area for parameter T_i and $\hat{T}_{i(r)}$ is the specific outcome of $\hat{T}_i$ obtained in simulation run r $(r = 1,....,R = 1000)$.

- *The average percentage relative bias* (ARB), averaged over m small areas is

$$ARB = m^{-1}\sum_{i=1}^{m} RB(\hat{T}_i).$$

- *The percentage relative root mean squared error* (RRMSE), defined as

$$RRMSE(\hat{T}_i) = T_i^{-1}\left\{\sqrt{\left(R^{-1}\sum_{r=1}^{R}(\hat{T}_{i(r)} - T_i)^2\right)}\right\}\times 100.$$

- *The average percentage relative root mean squared error* (ARRMSE), averaged over m small areas is

$$ARRMSE = m^{-1}\sum_{i=1}^{m} RRMSE(\hat{T}_i).$$

- *The coverage rate* (CR), defined as

$$CR(\hat{T}_i) = R^{-1}\sum_{r=1}^{R} 1\left(T_i \in \left\{\hat{T}_{i(r)} \pm 2\sqrt{mse\left(\hat{T}_{i(r)}\right)}\right\}\right).$$

Here $mse\left(\hat{T}_{i(r)}\right)$ is the estimate of the MSE of $\hat{T}_{i(r)}$ for the r^{th} simulation.

- *The average coverage rate* (ACR), averaged over m small areas is

$$ACR = m^{-1}\sum_{i=1}^{m} CR(\hat{T}_i).$$

3.4.3 Simulation Results

Three measures of estimation performance define in section 3.4.2 are computed using the estimates generated in the simulation study. These are the relative bias or relative mean errors and the relative root mean squared error (RMSE), both expressed as percentages, of regional mean estimates and the coverage rate of nominal 95 per cent confidence intervals for regional means. Table 3.3 presents the average and median values of these measures (all computed over the 29 regions) generated by EBLUP and MBD under models I-IV for the variable of interest TCC using ML and REML estimates for the random effects.

These results indicate the relative performance of the two SAE methods (EBLUP and MBD) do not change due to ML and REML estimates of variance components (Table 3.3). However, results generated by using REML estimates of variance components provide better performance than those by using ML estimates. Besides REML estimates, we use ML estimates of random effects to see how the MSE estimate of EBLUP (3.22) with a bias correction due to MLE compare with the simple MSE estimate of MBD (3.19). What follows next, we do refer only the results generated by using REML estimate of random effects to compare the EBLUP and the MBD methods.

In Table 3.3 we note that the average relative biases under MBD are smaller than those under EBLUP for all models except model IV. However, the average root mean square errors for MBD are marginally higher than those for EBLUP under models I and II and smaller for models III and IV. Average coverage rates (which should nominally be

around 95 percent) for MBD are relatively higher than those for EBLUP under all models. Although neither approach dominates, it seems clear that MBD is more robust to model misspecification than EBLUP.

Figures 3.7-3.9 show the region-specific performances generated by EBLUP and MBD methods (ordered by increasing population size) under REML estimates of random effects. Similar results generated by ML estimates of random effects are shown in Figures B.1 to B.3 in Appendix B. Figure 3.7 (and Figure B.1) shows the better relative bias performances of both EBLUP and MBD under model I and II and their worse relative bias performance under model IV. Figure 3.8 (also Figure B.2) shows that the relative RMSEs of regional estimates generated by MBD are comparable with those generated under EBLUP, with neither approach dominating. Overall, with the exception of two regions (3 and 21), it seems that MBD under model II performs marginally better overall. As indicated earlier in the AAGIS data, the regional sample sizes vary from 6 to 117 (Table 3.1). However, performances of the two methods (i.e. EBLUP and MBD) have not shown any pattern with sample sizes. That is relative performances of the EBLUP and MBD does not depend on small area sample sizes. See Figures 3.7 to 3.9.

Table 3.3 Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%) and average (ACR) coverage rate generated by MBD and EBLUP using ML and REML estimates of random effects under model I-IV. All averages and medians are over the 29 regions of interest.

	Model	Estimator	ARB	MRB	ARRMSE	MRRMSE	ACR
REML	I	EBLUP	4.24	1.55	19.92	15.74	0.90
		MBD	-2.49	-0.82	20.56	14.45	0.92
	II	EBLUP	2.98	0.61	19.87	16.40	0.85
		MBD	-2.13	-0.47	20.15	13.16	0.93
	III	EBLUP	4.52	1.95	23.89	19.94	0.69
		MBD	-3.84	0.13	21.14	14.44	0.94
	IV	EBLUP	1.17	-2.63	23.38	19.73	0.65
		MBD	2.20	2.06	22.35	20.61	0.97
ML	I	EBLUP	4.58	1.66	20.28	15.93	0.90
		MBD	-2.76	-0.89	20.49	14.38	0.92
	II	EBLUP	3.27	0.91	20.29	16.84	0.85
		MBD	-2.53	-0.61	20.18	13.08	0.93
	III	EBLUP	4.69	1.03	24.05	20.03	0.70
		MBD	-3.94	0.12	21.10	14.48	0.93
	IV	EBLUP	1.34	-2.95	23.50	19.88	0.67
		MBD	2.08	1.85	22.30	20.52	0.97

In the two regions (3 and 21) where MBD fails, inspection of the population and sample data indicated that this is because of a few outlying estimates. In fact, the outlying values of MBD for region 21 are all caused by the presence of a single massive outlier (TCC > A$30,000,000) in the original sample (see Figure 3.2). This outlier was included in the simulation population (twice) and then selected (in one case, twice) in 37 of the 1000 simulation samples. If we discard the outlier driven estimates in regions 3 and 21 then the MBD approach seems the method of choice for regional estimation in our simulation study. This is confirmed when we return to Table 3.3 and now consider the columns containing the median values of relative bias and relative RMSE.

Figure 3.9 (and Figure B.3) summarizes region-specific variation in the nominal 95 percent confidence interval coverage rates generated by EBLUP and MBD. If we ignore the outlier driven results for regions 3 and 21, the results displayed in Figure 3.9 show that MBD approach gives marginally better coverage rates under Models I and II. A close look at these results also indicates that in the event of model misspecification (e.g. under Models III and IV) the MBD coverage rate is more robust.

As mentioned earlier MBD is more robust to model misspecification. We can apply the MBD method of estimation more appropriately in many situations, where the EBLUP approach is not well suited. For example, for estimation of small areas of categorical survey variables, the EBLUP (3.18) based on a linear mixed model (3.11) is not appropriate and in such cases the suitable model is a generalised linear mixed model (GLMM). However, MBD methods still work well for such data. Empirical results (Appendix E) show no efficiency loss by using MBD estimator based on linear

assumption in this case. In chapter 4 we further discuss some other situations when EBLUP is unstable and MBD performs reasonably well.

3.5 Conclusions

Our empirical results indicate that the MBD estimator (3.15) performs well and represents a real alternative to the EBLUP (3.20), with the associated easy to calculate MSE estimator (3.19) providing good coverage performance. The MBD estimator under random slopes model II perform marginally better overall. Further, the MBD approach appears to be more robust than EBLUP in the realistic situation where (3.11) is a working model, rather than the (unknown) true model underpinning the data. However, this does not mean that the MBD is always preferable. Note that EBLUP, which approximates the best linear estimator when (3.11), actually holds, would be expected to dominate MBD in such a case. Further, for SAE of the categorical variables the EBLUP (3.20) based on a linear mixed model (3.11) is not appropriate and the adequate method is based on a generalised linear mixed model (GLMM). However, MBD methods still perform reasonably well in such cases. See Appendix E for some empirical results related to categorical survey variable.

We noticed some issues that influence the utility of the mixed model-based direct estimator (3.15) that remain unresolved. The negative weights, which occurred in some regions in the simulation study reported above, lead to impossible (i.e. negative) estimates. Since such values are easily identified, they should not cause problems in real

life. However, the problem remains of how to modify the weights (3.14) to ensure they are strictly positive. A related issue that has already been noted is the impact of outlier Y-values on (3.15). Certainly, this estimator since it is a linear combination of just the small area data values is more susceptible to outliers in these values than the EBLUP (3.20).

The MBD estimators discussed in this chapter are essentially based on the variable specific weights and efficient for estimation of the variable on which they are based. Development of "multipurpose" weights (i.e. not variable specific) can make the method even more useful. Furthermore, the data used in the simulation studies reported in section 3.4 are heteroskedastic in nature and the relationship between the survey and the auxiliary variables are not linear (Figure 3.4 and 3.5). Thus, the extension of MBD approach for the small area estimation with skewed data seems to be essential. In the proceeding chapters we shall consider these two issues.

Figure 3.7 Region-specific percentage relative biases for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with REML estimates.

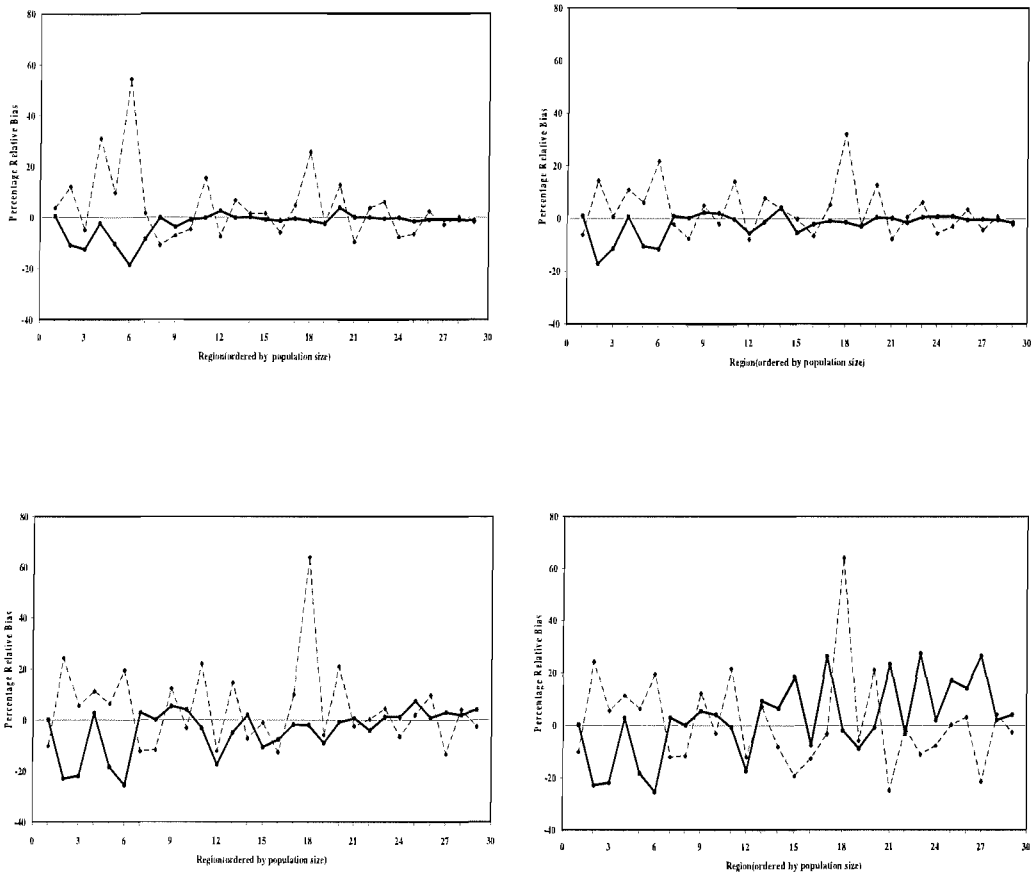


Figure 3.8 Region-specific percentage relative RMSE for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with REML estimates.

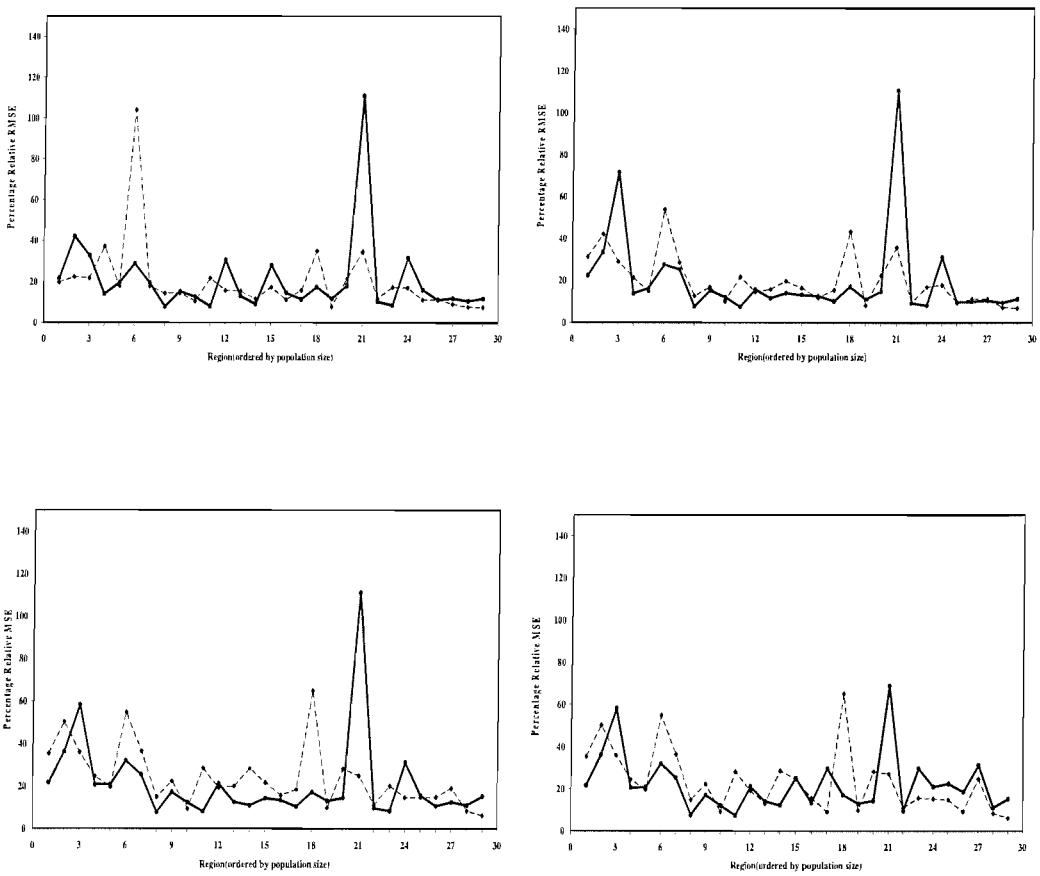
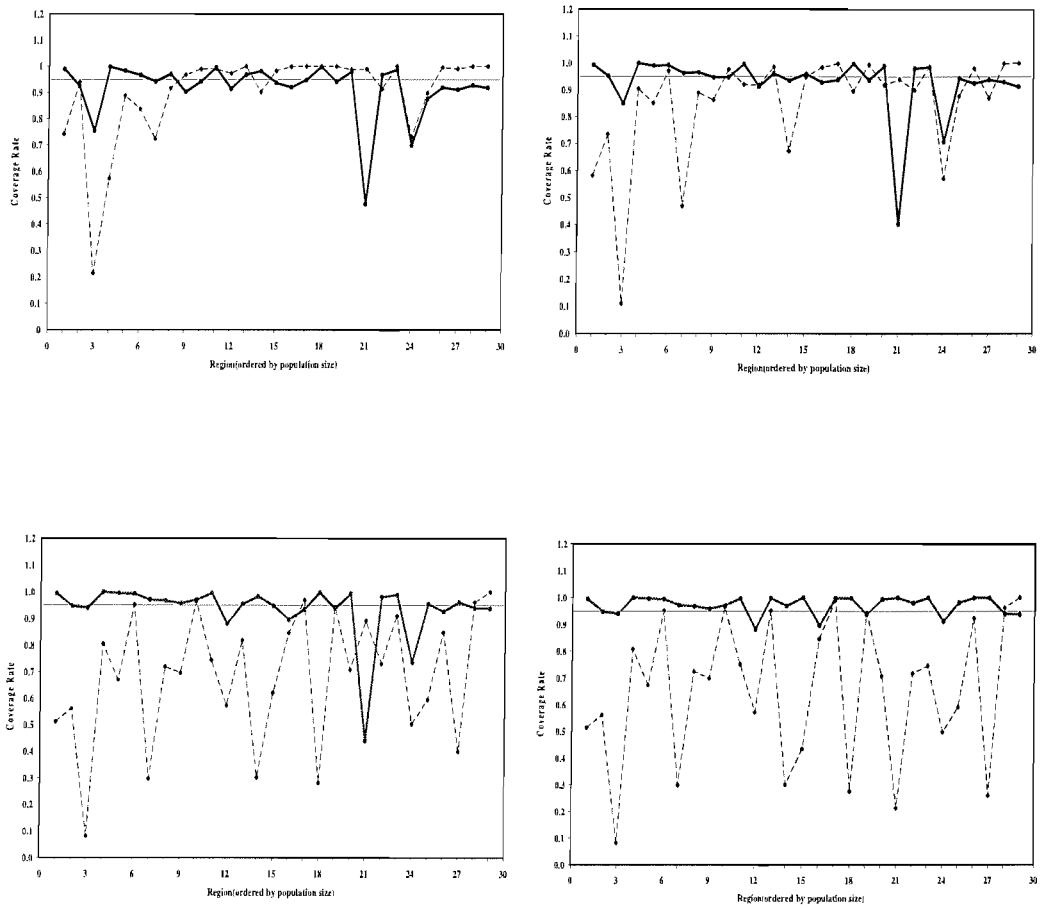


Figure 3.9 Region-specific coverage rate for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with REML estimates.



CHAPTER 4

MULTIPURPOSE SMALL AREA ESTIMATION

4.1 Introduction

The MBD method of small area estimation (SAE) described in chapter 3 uses sample weights derived under a population level linear mixed model to define the estimator for small areas. The weights that define the best linear unbiased predictor (BLUP) for the population total of a variable of interest (see Royall, 1976) depend on the population level conditional variance/covariance matrix for that variable. Unless this matrix is always proportional to a known matrix, this optimality is variable specific (Valliant et al, 2000, chapter 2). However, most surveys are multivariate, and it is often an advantage to have a common weight for all response variables. This is especially true where linear estimates are produced using the survey data. In what follows we refer to such weights as 'multipurpose'. Further, these multipurpose weights facilitate the production of linear estimates using computer software so that we do not need to use a different weight for each variable. Consequently, development of multipurpose sample weight has potential to make the MBD method of SAE more useful.

When a sufficiently rich set of auxiliary variables exist, and response variables can be assumed to be conditionally uncorrelated given these variables, multipurpose weights can be constructed by fitting a linear model for each response variable in terms of the complete set of auxiliary variables, as in Chambers (1996). An essentially equivalent idea is to use a calibrated set of sample weights, where the calibration is with respect to these auxiliary variables, as in Deville and Särndal (1992).

Small area estimation is now widely used in sample surveys. Many of the methods currently in use are variable specific and based on the application of mixed models (Rao, 2003). Weighted direct estimation for small areas based on these models is described in chapter 3, where we refer to this approach as the model-based direct (MBD) method of small area estimation. Since the weights used in MBD estimation are based on the second order properties of linear mixed models fitted to the survey variables, they are variable specific. However, as noted above, there are obvious practical advantages from having a single multipurpose weight that can be used for small area estimation for all the survey variables.

In this chapter we introduce the 'multipurpose' weights, optimal in some sense for a range of variables in multivariate surveys. Then we propose the multipurpose small area estimation. In particular, we extend the MBD approach for SAE for multivariate surveys using 'multipurpose' weights. In section 4.2 of this chapter we replace the variable specific BLUP optimality criterion that underlies the mixed model weights used in the MBD approach by a modified 'total variability' criterion that leads to a single set of

optimal multipurpose weights for use in MBD estimation for small areas. Section 4.3 then presents empirical results on the performance of this approach. Finally, in section 4.4 we summarise our empirical results.

4.2 Optimal Multipurpose Sample Weighting

In section 3.2 of the previous chapter, we define the sample weights for population estimation with single a response variable under the general linear model (3.8). Under the model (3.8) and following the notation given in section 3.2, it is known (see Royall, 1976) that among linear prediction unbiased estimators $\hat{T}_y = w_s' Y_s$ of T_y, the variance of the prediction error $Var(\hat{T}_y - T_y)$ is minimised by weights of the form

$$w_s = 1_n + H'\left(X_U' 1_N - X_s' 1_n\right) + \left(I_n - H'X_s'\right)V_{ss}^{-1}V_{sr}1_{N-n}. \tag{4.1}$$

Here $H = \left(X_s' V_{ss}^{-1} X_s\right)^{-1} X_s' V_{ss}^{-1}$, 1_g is a vectors of ones of order $g(g = n, N, N\text{-}n)$ and I_n is the identity matrix of order n. We refer to the weights (4.1) as the BLUP weights for Y. By definition, these weights are calibrated on the variables in X_U and so exactly reproduce the known population totals defined by the columns of this matrix. In other words $X_s' w_s = X_U' 1_N = T_x$. Furthermore, under the assumption that a linear mixed model can be used to specify the covariance matrix components V_{ss} and V_{sr} in (4.1), the MBD approach to small area estimation introduced in chapter3 uses these weights, with V_{ss} and V_{sr} replaced by suitable estimates, to define direct estimates of small area quantities.

4.2.1 Optimal Multipurpose Weighting for Uncorrelated Variables

Suppose we have K response variables and a common set of auxiliary variables with values defined by the population matrix X_U, and that model (3.8) holds for each of them (although with different parameter values). Suppose further that these variables are mutually uncorrelated. We use an extra subscript k $(k = 1,....,K)$ to denote quantities associated with the k^{th} response variable, for example V_{kss} and w_{ks} denote respectively the $n \times n$ covariance matrix and $n \times 1$ vector of sample weights that are associated with the $n \times 1$ vector Y_{ks} of sample values of the k^{th} response variable. With this notation, our aim is to derive an optimal set of multipurpose weights $w_s = \{w_j ; j \in s\}$ for the K response variables measured in the survey. Let $T_k = 1_N' Y_k$ denote the population total of Y_k, with estimator $\hat{T}_k = w_s' Y_{ks}$ based on these multipurpose weights. The weights w_s are then said to be ϕ-optimal if

(a) $E(\hat{T}_k - T_k) = 0$ for each value of k, and

(b) the ϕ-weighted total prediction variance $\sum_k \phi_k Var(\hat{T}_k - T_k)$ is minimised at w_s.

Here ϕ_k is a user-specified non-negative scalar quantity that reflects the relative importance attached to the k^{th} response variable, with $\sum_k \phi_k = 1$.

We now define the vector $a_s = w_s - 1_s$. Then the estimation error for the estimator $\hat{T}_k = w_s' Y_{ks}$ of $T_k = 1_N' Y_k$ is $\hat{T}_k - T_k = (w_s' Y_{ks} - 1_N' Y_k) = (a_s' Y_{ks} - 1_{N-n}' Y_{kr})$.

In order to derive an explicit expression for the ϕ-optimal multipurpose weights we first note that under (a)

$$E(\hat{T}_k - T_k) = E(a_s' Y_{ks} - 1_{N-n}' Y_{kr}) = E(a_s' X_s - 1_{N-n}' X_r)\beta_k = 0 \Rightarrow \; a_s' X_s = 1_{N-n}' X_r . \quad (4.2)$$

Furthermore, the prediction variance for estimator $\hat{T}_k = w_s' Y_{ks}$ is then

$$Var(\hat{T}_k - T_k) = E(\hat{T}_k - T_k)^2 = E(a_s' Y_{ks} - 1_{N-n}' Y_{kr})^2$$

$$= Var(a_s' Y_{ks} - 1_{N-n}' Y_{kr}) + \{E(a_s' Y_{ks} - 1_{N-n}' Y_{kr})\}^2 .$$

The second term on the right hand side above vanishes under (4.2), so that

$$Var(\hat{T}_k - T_k) = a_s' Var(Y_{ks}) a_s - 2a_s' Cov(Y_{ks}, Y_{kr}) 1_{N-n} + 1_{N-n}' Var(Y_{kr}) 1_{N-n}$$

$$= a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n} . \quad (4.3)$$

We use the method of Lagrange multipliers to minimise (4.3) subject to (4.2). The corresponding Lagrangian loss function is

$$\Phi^{(1)} = \sum_{k=1}^{K} \phi_k \{a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n}\} + 2(a_s' X_s - 1_{N-n}' X_r)\lambda . \quad (4.4)$$

The third term on right hand side of (4.4) is independent of a_s so we discarded it and consider the Lagrange function as

$$\Phi^{(1)} = \sum_{k=1}^{K} \phi_k \{a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n}\} + 2(a_s' X_s - 1_{N-n}' X_r)\lambda \quad (4.5)$$

where λ is a vector of Lagrange multipliers. Differentiating (4.5) with respect to a_s and setting the result equal to zero leads to

$$\frac{\partial \Phi^{(1)}}{\partial a_s} = \sum_{k=1}^{K} \phi_k \{2V_{kss} a_s - 2V_{ksr} 1_{N-n}\} + 2X_s \lambda = 0$$

$$\Rightarrow \qquad X_s \lambda = \sum_{k=1}^{K} \phi_k \{V_{ksr} 1_{N-n} - V_{kss} a_s\}$$

$$\Rightarrow \qquad X_s \lambda = \sum_{k=1}^{K} \phi_k V_{ksr} 1_{N-n} - \sum_{k=1}^{K} \phi_k V_{kss} a_s$$

$$\Rightarrow \qquad a_s = \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} \left\{\sum_{k=1}^{K} \phi_k V_{ksr} 1_{N-n} - X_s \lambda\right\} \tag{4.6}$$

Multiplying both sides of (4.6) on the left by X_s' and using (4.2), we see that

$$X_s' a_s = X_s' \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} \left(\sum_{k=1}^{K} \phi_k V_{ksr} 1_{N-n}\right) - X_s' \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} X_s \lambda$$

$$\Rightarrow \qquad X_r' 1_{N-n} = X_s' U_1^{-1} W_1 1_{N-n} - X_s' U_1^{-1} X_s \lambda$$

$$\Rightarrow \qquad \lambda = \left(X_s' U_1^{-1} X_s\right)^{-1} \left\{X_s' U_1^{-1} W_1 - X_r'\right\} 1_{N-n} \tag{4.7}$$

where $U_1 = \sum_{k=1}^{K} \phi_k V_{kss}$ and $W_1 = \sum_{k=1}^{K} \phi_k V_{ksr}$. Substituting (4.7) in (4.6) then yields the optimal value of a_s:

$$a_s^{(1)} = U_1^{-1} W_1 1_{N-n} - U_1^{-1} X_s \lambda = \left[U_1^{-1} W_1 - U_1^{-1} X_s \left(X_s' U_1^{-1} X_s\right)^{-1} \left\{X_s' U_1^{-1} W_1 - X_r'\right\}\right] 1_{N-n}$$

$$= U_1^{-1} X_s \left(X_s' U_1^{-1} X_s\right)^{-1} \left(X_U' 1_N - X_s' 1_n\right) + \left[I_n - U_1^{-1} X_s \left(X_s' U_1^{-1} X_s\right)^{-1} X_s'\right] U_1^{-1} W_1 1_{N-n}.$$

That is, the optimal multipurpose sample weights are given by

$$w_s^{(1)} = 1_n + H_1' \left(X_U' 1_N - X_s' 1_n\right) + \left[I_n - H_1' X_s'\right] U_1^{-1} W_1 1_{N-n} \tag{4.8}$$

where $H_1 = \left(X_s' U_1^{-1} X_s\right)^{-1} X_s' U_1^{-1} = \left\{X_s' \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} X_s\right\}^{-1} X_s' \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1}$.

Observe that the analytical form of the optimal multipurpose weights (4.8) is similar to the variable specific BLUP weights (4.1) except that V_{kss} and V_{ksr} are replaced by the weighted sums $U_1 = \sum_k \phi_k V_{kss}$ and $W_1 = \sum_k \phi_k V_{ksr}$ respectively. Clearly (4.8) reduces to (4.1) for $K = 1$. The choice of weight or importance ϕ_k attached to k^{th} variable is more or less subjective, and can to chosen depending on the nature of the data. When there is no reason to choose one variable over other then we can give equal weight to all

variables, meaning $\phi_k = K^{-1}; \forall k$. For example, we can assign the weight proportional to residual variances of the response variables.

4.2.2 Optimal Multipurpose Weighting for Correlated Variables

The multipurpose weights (4.8) are derived assuming that variables are mutually uncorrelated. However, in general survey variables are correlated. We now define the sample weights exploiting the correlations among the survey variables. For any two variables Y_k and Y_l $(k,l = 1,...,K)$, let $C_{kl} = C_{lk} = Cov(Y_k, Y_l)$. The obvious generalization of the ϕ-weighted total prediction variance to this case leads to the loss function

$$\left(\sqrt{\phi_1}, \sqrt{\phi_2}, ... \sqrt{\phi_K}\right) \Delta \left(\sqrt{\phi_1}, \sqrt{\phi_2}, ... \sqrt{\phi_K}\right)' \tag{4.9}$$

where elements of the matrix $\Delta = \{\Delta_{kl}\}$ are given by

$$\Delta_{kl} = \begin{cases} Var(\hat{T}_k - T_k) & if \ k = l \\ Cov(\hat{T}_k - T_k, \hat{T}_l - T_l) & if \ k \neq l \end{cases}$$

and we now have

$$\begin{aligned} Cov(\hat{T}_k - T_k, \hat{T}_l - T_l) &= Cov(a_s' Y_{ks} - 1_{N-n}' Y_{kr}, a_s' Y_{ls} - 1_{N-n}' Y_{lr}) \\ &= a_s' Cov(Y_{ks}, Y_{ls}) a_s - 2a_s' Cov(Y_{ks}, Y_{ls}) 1_{N-n} + 1_{N-n}' Cov(Y_{ks}, Y_{ls}) 1_{N-n} \\ &= a_s' C_{klss} a_s - 2a_s' C_{klsr} 1_{N-n} + 1_{N-n}' C_{klrr} 1_{N-n}. \end{aligned}$$

$$Var(\hat{T}_k - T_k) = a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n}$$

$$Var(\hat{T}_l - T_l) = a_s' V_{lss} a_s - 2a_s' V_{lsr} 1_{N-n} + 1_{N-n}' V_{lrr} 1_{N-n}.$$

The Lagrange function to be minimized in this case is

$$\Phi^{(2)} = \left(\sqrt{\phi_1}, \sqrt{\phi_2}, ... \sqrt{\phi_K}\right) \Delta \left(\sqrt{\phi_1}, \sqrt{\phi_2}, ... \sqrt{\phi_K}\right)' + 2(a_s' X_s - 1_{N-n}' X_r)\lambda$$

$$= \sum_k \phi_k Var(\hat{T}_k - T_k) + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} Cov(\hat{T}_k - T_k, \hat{T}_l - T_l) + 2(a_s' X_s - 1_{N-n}' X_r)\lambda$$

$$= \sum_k \phi_k \left\{a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n}\right\}$$

$$+ \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} \left\{a_s' C_{klss} a_s - 2a_s' C_{klsr} 1_{N-n} + 1_{N-n}' C_{klrr} 1_{N-n}\right\} + 2(a_s' X_s - 1_{N-n}' X_r)\lambda . \quad (4.10)$$

Differentiating (4.10) with respect to a_s and setting the result equal to zero yields

$$\left\{\sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}\right\} a_s - \left\{\sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}\right\} 1_{N-n} + X_s \lambda = 0$$

$$\Rightarrow \quad U_2 a_s - W_2 1_{N-n} + X_s \lambda = 0$$

$$\Rightarrow \quad a_s = U_2^{-1}\left(W_2 1_{N-n} - X_s \lambda\right) \quad (4.11)$$

where $U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}$ and $W_2 = \sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}$.

Proceeding as in the uncorrelated case then leads to the optimal multipurpose weights for correlated survey variables as

$$w_s^{(2)} = 1_n + H_2'\left(X_U' 1_N - X_s' 1_n\right) + \left[I_n - H_2' X_s'\right] U_2^{-1} W_2 1_{N-n} \quad (4.12)$$

where $H_2 = \left(X_s' U_2^{-1} X_s\right)^{-1} X_s' U_2^{-1}$. As in the uncorrelated variables case, we note that the weights defined by (4.12) have the same analytic form as the BLUP weights (4.1), except that in this case V_{kss} and V_{ksr} are replaced by $U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}$ and $W_2 = \sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}$ respectively. Between multipurpose weight (4.8) and

(4.12), the weight (4.12) includes extra terms in their variance components that arise due to correlation among the variables given by C_{kl}. When $C_{kl} = 0$, the weight (4.12) reduces to (4.8).

4.3. Application to Small Area Estimation

In section 4.2 we discussed three types of sample weights, the variable specific weights (4.1), the multipurpose weights (4.8) based on uncorrelated and (4.12) based on correlated survey variables respectively. These weights are derived under the model appropriate for population estimation (i.e. using a model that explains population level variability and small area effects are averaged out) and using these weights for SAE can lead to inefficient estimates. The most commonly used class of models in small area inference is the class of linear mixed models (see section 3.3, chapter 3). This is also true when we have more than one variable. In this section we first recall the linear mixed model and MBD estimation and then define them for the multivariate case to derive the multipurpose weights and associated estimators for small areas.

Following the MBD estimation elaborated in chapter 3, we use the multipurpose weights (4.8) and (4.12) to construct model-based direct (MBD) estimates for small area means. In this case we assume that the population can be partitioned into m non-overlapping small areas or domains, indexed by i in what follows. Thus, for example, the population size of area i is denoted by N_i and so on. The variable-specific MBD estimate of the mean of the k^{th} response variable with values y_{kj} in area i is then

$$\hat{\bar{Y}}_{k,i}^{MBD} = \sum_{j \in s_i} w_{kj} y_{kj} \Big/ \sum_{j \in s_i} w_{kj} \tag{4.13}$$

where s_i denotes the sample (of size n_i) in area i and the weights w_{kj} are calculated using (4.1), substituting estimated values $\hat{V}_{kss}$ and $\hat{V}_{ksr}$ for the corresponding components of the covariance matrix of the population values of this variable. Here model (3.11) and interpretation of different terms apply directly with inclusion of one extra subscript k for k^{th} variable, however, we redefined the mixed linear model (3.11) by introducing the subscript k, just for shake of continuity. In order to define these estimates, we assume that these population values follow the linear mixed model

$$Y_{kU} = X_U \beta_k + Z_U u_k + e_{kU} \tag{4.14}$$

where $Y_{kU} = (Y'_{k,1},...,Y'_{k,m})'$, $X_U = (X'_1,..,X'_m)'$, $Z_U = diag(Z_i; 1 \le i \le m)$, $u_k = (u'_{k,1},...,u'_{k,m})'$ and $e_{kU} = (e'_{k,1},...,e'_{k,m})'$ denote partitioning into area 'components'. Here $u_{k,i}$ is a $q \times 1$ vector of the random effect associated with area i, with $Var(u_{k,i}) = \Sigma_{u,k}$, and $e_{k,i}$ is the vector of individual level random effects for small area i, with variance $Var(e_{k,i}) = \Sigma_{e,k} I_{N_i}$. It follows that $Var(Y_{k,i}) = V_{k,i} = \Sigma_{e,k} I_{N_i} + Z_i \Sigma_{u,k} Z'_i$. The variance components $\Sigma_{e,k}$ and $\Sigma_{u,k}$ can be estimated from the sample data using standard methods (maximum likelihood, restricted maximum likelihood, i.e. REML, or method of moments). Substituting these estimated variance components back into the definition of $V_{k,i}$ and noting that $V_{kU} = diag(V_{k,i}; 1 \le i \le m)$ then leads to a corresponding estimate of this population level covariance matrix. This can be appropriately partitioned into sample and non-sample components to give the estimated values $\hat{V}_{kss}$ and $\hat{V}_{ksr}$. We refer to the

weights (4.1) with these estimated values substituted as the (variable specific) EBLUP weights (just as in chapter 3).

In order to use the multipurpose weights (4.8) and (4.12) in MBD estimation, we assume that the survey variables all follow the linear mixed model (4.14), with normal random effects. Furthermore, for any two variables of interest, say the k^{th} and l^{th}, area and individual random effects remain uncorrelated but now

$$\begin{pmatrix} u_{ki} \\ u_{li} \end{pmatrix} \sim MVN(0,\Sigma_u) \text{ with } \Sigma_u = \begin{pmatrix} Var(u_{ki}) & Cov(u_{ki},u_{li}) \\ Cov(u_{li},u_{ki}) & Var(u_{li}) \end{pmatrix} = \begin{pmatrix} \Sigma_{u,kk} & \Sigma_{u,kl} \\ \Sigma_{u,kl} & \Sigma_{u,ll} \end{pmatrix} \tag{4.15}$$

and

$$\begin{pmatrix} e_{kij} \\ e_{lij} \end{pmatrix} \sim MVN(0,\Sigma_e) \text{ with } \Sigma_e = \begin{pmatrix} Var(e_{kij}) & Cov(e_{kij},e_{lij}) \\ Cov(e_{lij},e_{kij}) & Var(e_{lij}) \end{pmatrix} = \begin{pmatrix} \Sigma_{e,kk} & \Sigma_{e,kl} \\ \Sigma_{e,kl} & \Sigma_{e,ll} \end{pmatrix}. \tag{4.16}$$

Hence

$$V_{k,i} = Var(Y_{k,i}) = \Sigma_{e,kk} I_{N_i} + Z_i \Sigma_{u,kk} Z_i'$$

$$V_{l,i} = Var(Y_{l,i}) = \Sigma_{e,ll} I_{N_i} + Z_i \Sigma_{u,ll} Z_i'$$

and

$$C_{kl,i} = Cov(Y_{k,i},Y_{l,i}) = \Sigma_{e,kl} I_{N_i} + Z_i \Sigma_{u,kl} Z_i'.$$

Given these definitions, we put $U_1 = diag(U_{1i};1 \le i \le m)$ and $W_1 = diag(W_{1i};1 \le i \le m)$ in (4.8) and $U_2 = diag(U_{2i};1 \le i \le m)$ and $W_2 = diag(W_{2i};1 \le i \le m)$ in (4.12). Here

$$U_{1i} = \sum_k \phi_k V_{kss,i} = \sum_k \phi_k \left(\Sigma_{e,kk} I_{n_i} + Z_{s,i} \Sigma_{u,kk} Z_{s,i}' \right)$$

$$W_{1i} = \sum_k \phi_k V_{ksr,i} = \sum_k \phi_k \left(Z_{s,i} \Sigma_{u,kk} Z_{r,i}' \right)$$

and

$$U_{2i} = \sum_k \phi_k V_{kss,i} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss,i}$$

$$= \sum_k \phi_k \left(\Sigma_{e,kk} I_{n_i} + Z_{s,i} \Sigma_{u,kk} Z'_{s,i} \right) + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} \left(\Sigma_{e,kl} I_{n_i} + Z_{s,i} \Sigma_{u,kl} Z'_{s,i} \right)$$

$$W_{2i} = \sum_k \phi_k V_{ksr,i} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr,i}$$

$$= \sum_k \phi_k \left(Z_{s,i} \Sigma_{u,kk} Z'_{r,i} \right) + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} \left(Z_{s,i} \Sigma_{u,kl} Z'_{r,i} \right).$$

In practice, the bivariate variance components $\Sigma_{u,kk}, \Sigma_{u,kl}, \Sigma_{e,kk}$ and $\Sigma_{e,kl}$, see (4.15) and (4.16), are unknown and must be estimated from the survey data. For example, in the empirical study described in the next section, these components were estimated using the method of moments. In any case, substituting estimates for these components in the formulae above then enables us to compute U_1, W_1, U_2 and W_2, and hence the multipurpose weights (4.8) and (4.12). Computation of MBD estimates for the small area means of the different survey variables is then straightforward using (4.13), with these multipurpose weights replacing the variable specific EBLUP weights there.

As noted earlier, the multipurpose weights (4.8) and (4.12) are essentially EBLUP type weights based on 'importance averaging' of the variance and covariance components associated with the different survey variables. This motivates us to consider a second approach to deriving multipurpose weights based on corresponding 'importance averaging' of the variable specific EBLUP sample weights (4.1) for these variables. That

is, we simply define our multipurpose weights as the importance-weighted average of the variable specific weights (4.1) across all K survey variables. This leads to weights

$$w_s^{(3)} = \sum_k \phi_k w_{sk} \tag{4.17}$$

where w_{sk} denotes the value of (4.1) for the k^{th} survey variable and ϕ_k denotes the relative importance of this variable, with $\sum_k \phi_k = 1$. The two approaches of deriving the 'multipurpose' weights based on averaging the variance-covariance components and averaging the variable specific weights are referred as the approach I and II respectively.

Mean squared errors for the EBLUP are estimated using the approach of Prasad and Rao (1990), while mean squared errors for the various MBD estimators are estimated using the robust method described in section 3.3.2 in chapter 3, which itself is an application of the heteroskedasticity robust method of prediction variance estimation in Royall and Cumberland (1978). That is, estimation of the mean squared errors of the MBD estimators (4.13) defined via multipurpose weights (4.8), (4.12) and (4.17) follows the approach described in section 3.3.2, and treats these estimators as simple weighted domain mean estimates. Under this approach the sample weights derived under mixed effect model are treated as fixed and the prediction variance of corresponding estimators are estimated using a standard robust variance estimator. In particular, the mean squared error estimator for the MBD estimators $\hat{\bar{Y}}_i^{MBD}$ of $\bar{Y}_i$ given in (4.13) is

$$mse(\hat{\bar{Y}}_i^{MBD}) = v(\hat{\bar{Y}}_i^{MBD}) + b^2(\hat{\bar{Y}}_i^{MBD}) \tag{4.18}$$

where $v(\hat{\bar{Y}}_i^{MBD}) = \sum_{s_i} \lambda_j (y_j - x_j'\hat{\beta})^2$ is the estimate of the prediction variance of $\hat{\bar{Y}}_i^{MBD}$ with

$$\lambda_j = N_i^{-2}\left(a_j^2 + (N_i - n_i)/(n_i - 1)\right),$$

$$a_j = \left(\sum_{s_i} w_j\right)^{-1}\left(N_i w_j - \sum_{s_i} w_j\right),$$

and $b(\hat{\bar{Y}}_i^{MBD}) = (\hat{\bar{X}}_i^{MBD} - \bar{X}_i)'\hat{\beta}$ is the estimate of the bias of $\hat{\bar{Y}}_i^{MBD}$. Here $\hat{\bar{X}}_i^{MBD}$ denotes the weighted average of the sample values of the auxiliary variables in area *i*. The sample weights used in (4.18) are the multipurpose weights (4.8), (4.12) and (4.17), depending the estimators. However, $\hat{\beta}$ used in (4.18) are variable specific and estimated from the variable of interest. Appendix F shows some other options for estimating β in context of multipurpose SAE.

We described two approaches for deriving multipurpose weights based on small area models, the first based on the weighted average (or sum) of the variance-covariance components associated with a select group of variables and the second based on weighted average (or sum) of the variable specific sample weights generated for these variables. Using these two approaches, we defined three types of multipurpose weights and corresponding small area estimators and their mean squared error. In the next section we carry out simulation studies to evaluate the empirical performance of these estimators as well as variable specific weighting based MBD and EBLUP. We also examine the utility of multipurpose weights for SAE of an arbitrary variable from same survey, not included in the definition of the multipurpose weights.

4.4 An Empirical Evaluation

In this section we report on a design-based simulation study that illustrates the performance of small area MBD estimation combined with multipurpose weights. The basis of this study is the same target population of N = 81982 farms, the same 1000 independent replications of a stratified random sampling design with overall sample size $n = 1652$ and the same m = 29 small areas of interest (defined by agricultural regions) that underpin the simulation results reported in chapter 3. Note that regional sample sizes in this design are fixed from simulation to simulation but vary between regions, ranging from a low of 6 to a high of 117, and hence allowing an evaluation of the performance of the different methods considered across a range of realistic small area sample sizes. See section 3.4 for more details. Here we consider K = 8 variables of interest. These are

(i) TCC = total cash costs (A$) of the farm business over the surveyed year,

(ii) TCR = total cash receipts (A$) of the farm business over the surveyed year,

(iii) FCI = farm cash income (A$), defined as TCR – TCC,

(iv) Crops = area under crops (in hectares),

(v) Cattle = number of Cattle on the farm,

(vi) Sheep = number of sheep on the farm,

(vii) Equity = total farm equity (A$), and

(viii) Debt = total farm debt (A$).

Our aim is to estimate the average of these variables in each of the 29 different regions. In doing so, we use the fact that these regions can be grouped into three zones (Pastoral, Mixed Farming, and Coastal), with farm area (hectares) known for each farm in the

population. This auxiliary variable is referred to as Size in what follows. Therefore, we have a single auxiliary variable for all 8 target variables.

Although the linear relationship between the eight target variables and Size is rather weak in the original sample data, this improves when separate linear models are fitted within six post strata. These post-strata are defined by splitting each zone into small farms (farm area less than zone median) and large farms (farm area greater than or equal to zone median). The mixed model (4.14) is therefore specified so that the matrix X of auxiliary variable values included an effect for Size, effects for the post-strata and effects for interactions between Size and the post strata as in chapter 3. Two different specifications for Z (corresponding to whether a random slope on Size was included or not) were considered. We refer to these as model I and as model II respectively below. These are random intercepts and random slopes model as in chapter 3. We use REML estimates of random effects parameters, obtained via the *lme* function in R (Bates and Pinheiro, 1998) when fitting (4.14) to individual survey variables. When fitting the multivariate mixed models defined by (4.15) and (4.16) we use the method of moments (Rao, 2003). See Appendix G for definition and expressions for the method of moment estimation.

4.4.1 Description of Estimators

The simulation study investigates the empirical performance of five different estimators of the 29 regional means, along with corresponding estimators of their mean squared error. These are

(i) the variable specific EBLUP under (4.14), referred to as EBLUP (see section 3.3.3, chapter 3);

(ii) the MBD estimator (4.13) based on variable specific EBLUP weights (4.1) under (4.14), referred to as MBD0;

(iii) the MBD estimator (4.13) based on multipurpose weights (4.8) under (4.14), referred to as MBD1-A;

(iv) the MBD estimator (4.13) based on multipurpose weights (4.12) under (4.14), referred to as MBD1-B, and

(v) the MBD estimator (4.13) based on multipurpose weights (4.17) under (4.14), referred to as MBD2.

4.4.2 Description of Simulation Studies

The simulation study is carried out in five stages. In the first stage, model I is assumed and the performance of the three estimators MBD0, MBD1-A and MBD1-B for two variables (TCC and TCR) is investigated to see if there are gains to be had from exploiting correlations among the survey variables. In this case we use method of moments (Henderson's method 3) to estimate the model parameters (see Appendix G). Results from the first stage of simulation are set out in Table 4.1. For this stage of simulations, we also carried out the model-based simulations to contrast the performance of different estimators for the population generated under the model. A description of the model-based simulations and corresponding results are shown in Appendix H. In the second stage of the study we compare the performance of the four estimation methods

EBLUP, MBD0, MBD1-A and MBD2 under models I and II for the 5 response variables (TCC, TCR, FCI, Cattle and Sheep) where both models can be fitted. Results from this stage are presented in Tables 4.2-4.3 and in Figure 4.1-4.5.

Note that three of the eight target variables in the study (Crops, Equity and Debt) are not suited to linear modeling via (4.14) under model II because of large numbers of zeros, so the multipurpose weights used in MBD1-A and MBD2 are based on the $K = 5$ remaining variables (TCC, TCR, FCI, Cattle and Sheep) in the simulations evaluating the performance of different methods under the model I and II. Consequently, in the third stage of the study, we use the multipurpose weights derived in the second phase (i.e. weights based on the $K = 5$ variables TCC, TCR, FCI, Cattle and Sheep) in MBD1-A to evaluate the performance of this estimator for the three variables Crops, Equity and Debt that are impossible to model using model II. That is we evaluate the performance of different methods for three target variables (Crops, Equity and Debt) that contain a large number of zeros and which are not included in the multipurpose weights. In this stage, our purpose is to investigate the utility of multipurpose weights to the variables not included in the weight. Results from this stage are shown in Table 4.4 and in Figures 4.6 to 4.8.

In the fourth stage we use the fact that model I can be fitted to all eight variables to define multipurpose weights that we then use in MBD1-A. That is we consider all the $K = 8$ response variables under model I. In this stage we investigate the effect of number of target variables included in the multipurpose weights. Results from this stage are

presented in Tables 4.5 and in Figures 4.6-4.8. Note that in all four of these simulation stages, we assign equal importance to all variables included in derivation of the multipurpose weights, i.e. $\phi_k = 1/K, \forall k$. However, in the final simulation studies (stage five) we replicate the stage two simulations for MBD1-A, but this time we assign weights to each variable proportional to its variability, i.e. $\phi_k = 1/\sigma^2_{e,k}$ or $\phi_k = 1/total\ \text{var}iance$. Results from this stage are reported in Tables 4.6. We assign an equal importance to the variables included in defining the multipurpose weights if there no reason to prefer one to the other. However, for a given data set, depending on the nature of the variables we identify some criterion to assign relative importance. For example, in AAGIS data, variability of some of the variables is different from the others. Thus, in simulation set five we decide to assign importance proportional to the variability of different variables.

4.4.3 Results of the Simulation Studies

We computed three measures of estimation performance using the estimates generated by different estimation methods in various simulation studies. These are the relative bias (RB) or relative mean errors and the relative root mean squared error (RRMSE), both expressed as percentages, of regional mean estimates and the coverage rate of nominal 95 per cent confidence intervals for regional means. Further, the average and median values of these performance measures are calculated over all the regions. See section 3.4.2 for the definition of these measures.

4.4.3.1 *First Stage Simulations*

Table 4.1 presents the average and median values of various measures of estimation performance for the first stage simulations (all computed over the 29 regions) generated by four methods MBD0, MBD1-A and MBD1-B under model I for the two variable TCC and TCR. As mentioned earlier, this stage of simulations use the method of moment (Henderson's method-3) for the estimation of random effect parameter (Appendix G).

For the variable TCC, we note that the average and median relative biases under MBD0 are larger than both MBD1-A and MBD1-B. However, with equal average coverage rates, the average and median relative RMSEs for MBD0 are marginally lower than both MBD1-A and MBD1-B. In contrast, for the variable TCR, the average and median relative biases under MBD0 are small than both MBD1-A and MBD1-B. However, with the same average coverage rate, the average and median relative RMSEs for MBD0 are marginally higher than both MBD1-A and MBD1-B. We have not presented the regional estimates generated by these methods since there are no significant differences between them. These results show that neither approach dominates the other. Between MBD1-A and MBD1-B it seems clear that both methods perform equally well. This is evidence that the MBD method based on the multipurpose weights (4.8) is not sensitive to correlations between the target variables. Although not presented here, results from model-based simulations of target variables with different levels of correlation support this conclusion. The results from model-based simulations are presented in Appendix H. Consequently the simulation results presented below focus on MBD1-A.

Table 4.1 Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate generated by MBD0, MBD1-A and MBD1-B for TCC and TCR under model I. All averages and medians are over the 29 regions of interest.

Variables	Criterion	MBD0	MBD1-A	MBD1-B
TCC	ARB	-2.99	-2.67	-2.71
	ARRMSE	20.32	20.39	20.39
	ACR	0.92	0.92	0.92
	MRB	-0.92	-0.85	-0.86
	MRRMSE	14.29	14.36	14.35
TCR	ARB	-2.38	-2.62	-2.67
	ARRMSE	21.21	21.13	21.12
	ACR	0.92	0.92	0.92
	MRB	-0.52	-0.56	-0.57
	MRRMSE	13.28	13.27	13.27

4.4.3.2 *Second Stage Simulations*

In the second stage of the simulation study, we compared the two variable specific methods EBLUP and MBD0 with the two multipurpose methods MBD1-A and MBD2. Tables 4.2 and 4.3 show the summary performances generated by these four methods for the five variables TCC, TCR, FCI, Cattle and Sheep under the 'reasonably specified' models I and II respectively. These results show that under the better fitting Model II (Table 4.3), there is little, if any, difference in the average relative biases of the multipurpose methods MBD1-A and MBD2 compared with the average relative bias of the variable specific estimator MBD0, with all three often substantially better than EBLUP (Table 4.2-4.3). Under Model I, the two multipurpose estimators MBD1-A and MBD2 are substantially better than MBD0 and EBLUP. In terms of relative RMSE, the results are more equivocal. Under Model I there is little to choose between MBD0, MBD1-A and MBD2 in terms of average relative RMSE, with the corresponding performance of EBLUP rather more fragile. When one turns to the better fitting Model II, however, it is clear that the better multipurpose approach is MBD1-A. By considering median, rather than average, values of relative bias and relative RMSE, we also see that the estimation performances of the multipurpose estimators MBD1-A and MBD2 appear to be more robust than those of the variable specific estimators MBD0 and EBLUP. Finally, we note that the average coverage rates of all three direct estimators are quite similar under both Models I and II and dominate the corresponding average coverage performance of EBLUP. Overall it seems clear that the multipurpose estimator MBD1-A is the estimator of choice for these five variables.

Table 4.2 Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate for the five variables best suited to linear mixed modelling. All averages and medians over the 29 regions of interest. Model I is assumed.

Criterion	Method	TCC	TCR	FCI	Cattle	Sheep
ARB	EBLUP	4.24	5.48	6.93	138.48	304.24
	MBD0	-2.49	-9.25	-13.80	-15.05	-7.33
	MBD1-A	-1.54	-1.30	-0.50	-1.78	0.69
	MBD2	-1.29	-1.02	-0.04	-1.35	0.98
MRB	EBLUP	1.55	0.55	-2.08	0.95	-0.23
	MBD0	-0.82	-3.87	-2.83	-4.79	-4.48
	MBD1-A	-0.61	-0.42	-0.56	-0.97	-0.35
	MBD2	-0.52	-0.39	-0.54	-0.75	-0.30
ARRMSE	EBLUP	19.92	21.76	63.93	304.74	906.18
	MBD0	20.56	23.34	54.42	37.45	24.88
	MBD1-A	20.86	21.77	59.72	33.29	30.24
	MBD2	20.85	21.77	60.07	33.36	30.64
MRRMSE	EBLUP	15.74	14.83	40.41	25.97	13.00
	MBD0	14.45	16.20	35.85	30.34	15.50
	MBD1-A	14.69	13.41	42.09	30.55	14.67
	MBD2	14.74	13.46	42.45	30.56	14.67
ACR	EBLUP	0.90	0.88	0.87	0.86	0.91
	MBD0	0.92	0.91	0.94	0.93	0.94
	MBD1-A	0.92	0.92	0.94	0.95	0.96
	MBD2	0.92	0.92	0.94	0.95	0.96

Table 4.3 Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate for the five variables best suited to linear mixed modelling. All averages and medians over the 29 regions of interest. Model II is assumed.

Criterion	Method	TCC	TCR	FCI	Cattle	Sheep
ARB	EBLUP	2.98	2.85	16.70	131.66	2.63
	MBD0	-2.13	-1.25	0.50	-0.29	3.66
	MBD1-A	-1.67	-1.29	0.74	-1.95	1.10
	MBD2	-1.30	-0.72	3.17	-1.29	0.93
MRB	EBLUP	0.61	1.37	3.98	0.62	0.00
	MBD0	-0.47	-0.51	0.35	-0.31	0.00
	MBD1-A	-0.65	-0.50	0.24	-0.30	-0.15
	MBD2	-0.52	0.01	0.53	-0.22	-0.09
ARRMSE	EBLUP	19.87	20.28	68.85	231.08	630.01
	MBD0	20.15	21.46	65.43	30.80	37.82
	MBD1-A	19.06	21.03	64.03	30.09	32.04
	MBD2	27.13	34.84	129.29	45.16	34.99
MRRMSE	EBLUP	16.40	15.61	33.89	22.64	11.73
	MBD0	13.16	12.39	37.64	28.79	14.68
	MBD1-A	12.84	12.18	37.92	24.84	14.77
	MBD2	12.84	12.71	37.62	24.93	14.72
ACR	EBLUP	0.85	0.86	0.84	0.86	0.89
	MBD0	0.93	0.93	0.90	0.95	0.96
	MBD1-A	0.93	0.93	0.94	0.95	0.96
	MBD2	0.93	0.93	0.94	0.95	0.96

Figure 4.1 to 4.5 show the regional level performances of EBLUP, MBD0, MBD1-A and MBD2 for the five variables TCC, TCR, FCI, Cattle and Sheep respectively under model I and model II. Note the relatively better performance of all methods under model II. A considerable reduction in relative biases under multipurpose weighting can also be seen in most regions for all variables. These results further show significant gain in efficiency due to multipurpose approach in terms of relative RMSEs as well as coverage rates in different regions and for the different variables.

Figure 4.1, which shows the region-specific performance for the variable TCC, indicates that in two regions (region 3 and 21) the weighting methods (MBD0, MBD1-A and MBD2) fail, in general. Inspection of data indicates that this is the consequence of a few outlying estimates as noted in chapter 3. When we discard these outlying estimates as in chapter 3, the weighting methods, particularly MBD1-A and MBD2, perform well for TCC across all regions.

Figure 4.2 indicates that the root mean squared errors for the variable TCR under weighting methods are relatively higher in two regions (region 3 and 21). Again similar to the TCC, in these two regions results are contaminated by a few outlying estimates. The outlying estimates for region 21 are all caused by presence of a single massive outlier (TCR=A$ 33,031,486) from the original sample that was included in the simulation population (twice). This also affects the coverage rate of TCR under weighting methods in region 21 (Figure 4.2). If we discard the outlier driven estimate in region 21 (i.e. see the median performance measures generated by these methods, Table

4.2 and 4.3) then weighting methods, particularly the MBD1-A and MBD2 seems to be appropriate for regional estimation for the variable TCC and TCR under models I and II. Similarly, the results generated by all methods (EBLUP, MBD0, MBD1-A and MBD2) for FCI are influenced by outlier contaminated estimates in two regions (3 and 15). See Figures 4.3.

The unstable performance of EBLUP for the Cattle and Sheep variables in Table 4.2 and 4.3 is noteworthy. Upon investigation we found that the anomalous results for Cattle are caused by the presence of negative estimates (a negative estimate is really unexpected and surprising) for this variable in two regions (11 and 14), which are themselves the result of zero values in the data (Figure 4.4). In particular, in region 11 there are 1283 zeros in the simulated population of 1586 values (in original sample of size 51, there are 39 zeros). This resulted in 185 negative estimates out of the 1000 simulated for this region. Similarly in the region 14, there are 1972 zeros in the 2182 values in the simulated population (there are 43 zeros out of 47 in original sample), leading to 354 negative estimates. However, in region 6 the MBD0 is affected due to presence of one massive outlier (cattle= 33154) which was selected four times in the simulated population and the EBLUP is affected due to repetition of zero value. In region 6, in a sample of size 19, there are two zero observations.

Figure 4.1 Region-specific performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for TCC under model I (left) and model II (right).

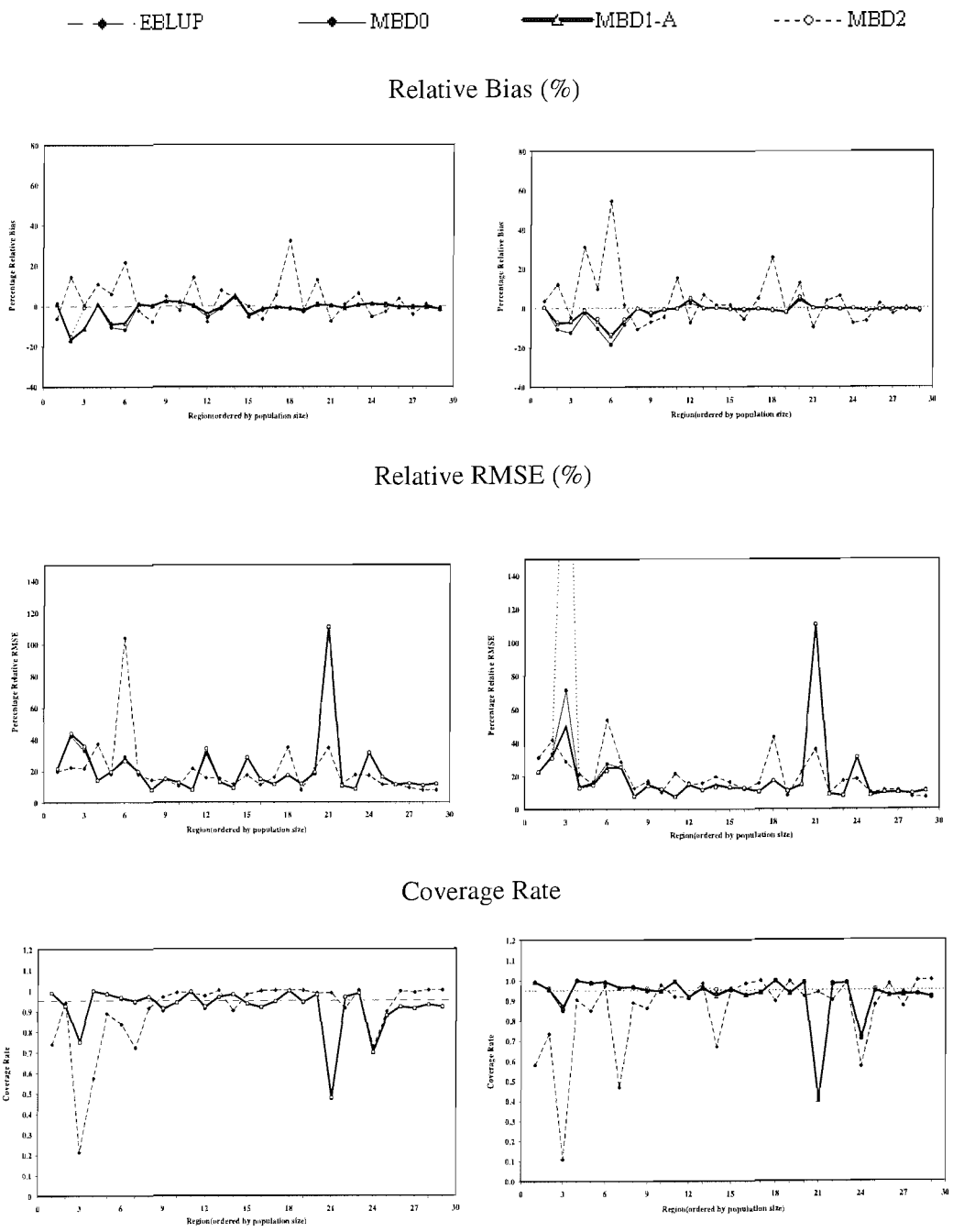


Figure 4.2 Region-specific performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for TCR under model I (left) and model II (right).

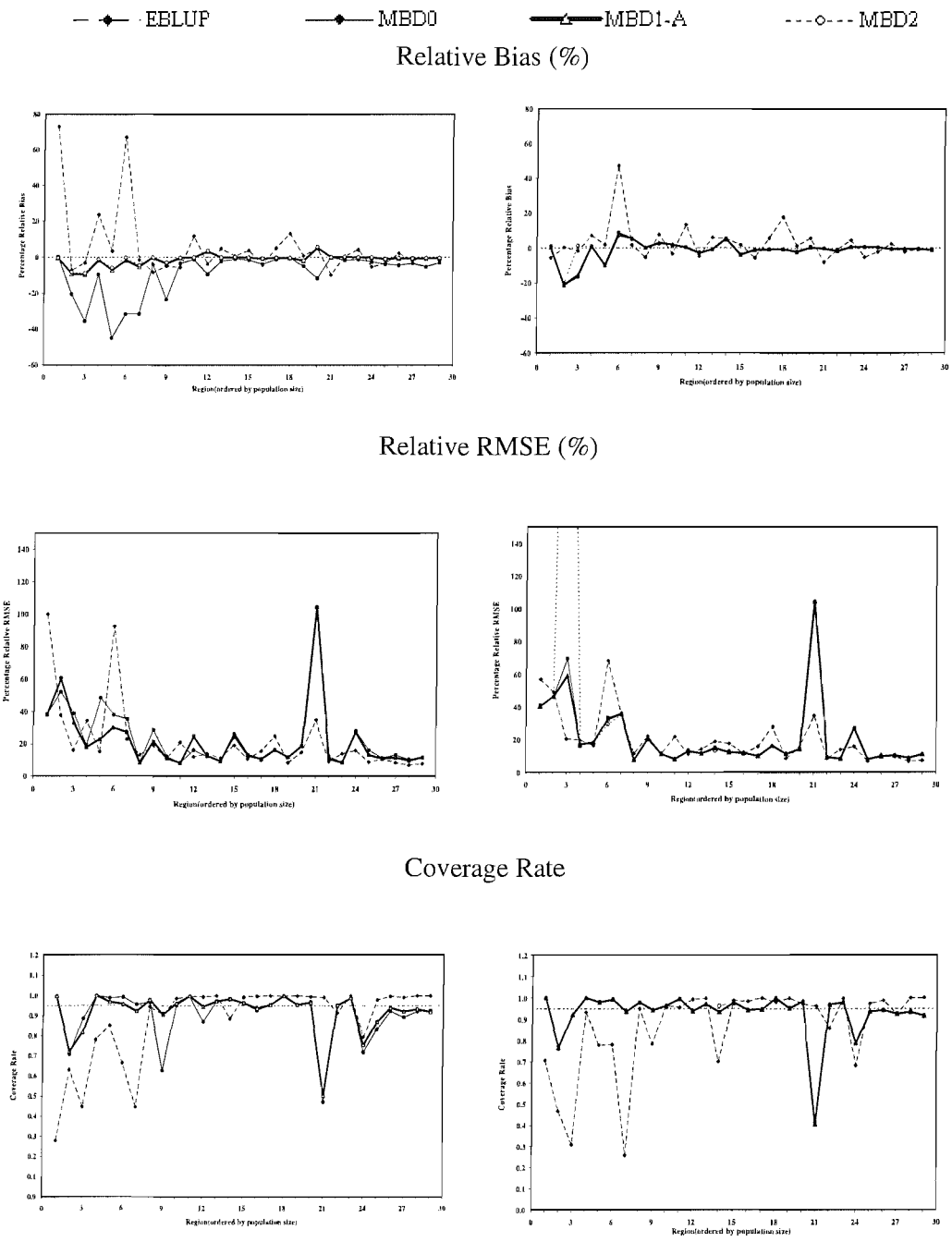


Figure 4.3 Region-specific performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for FCI under model I (left) and model II (right).

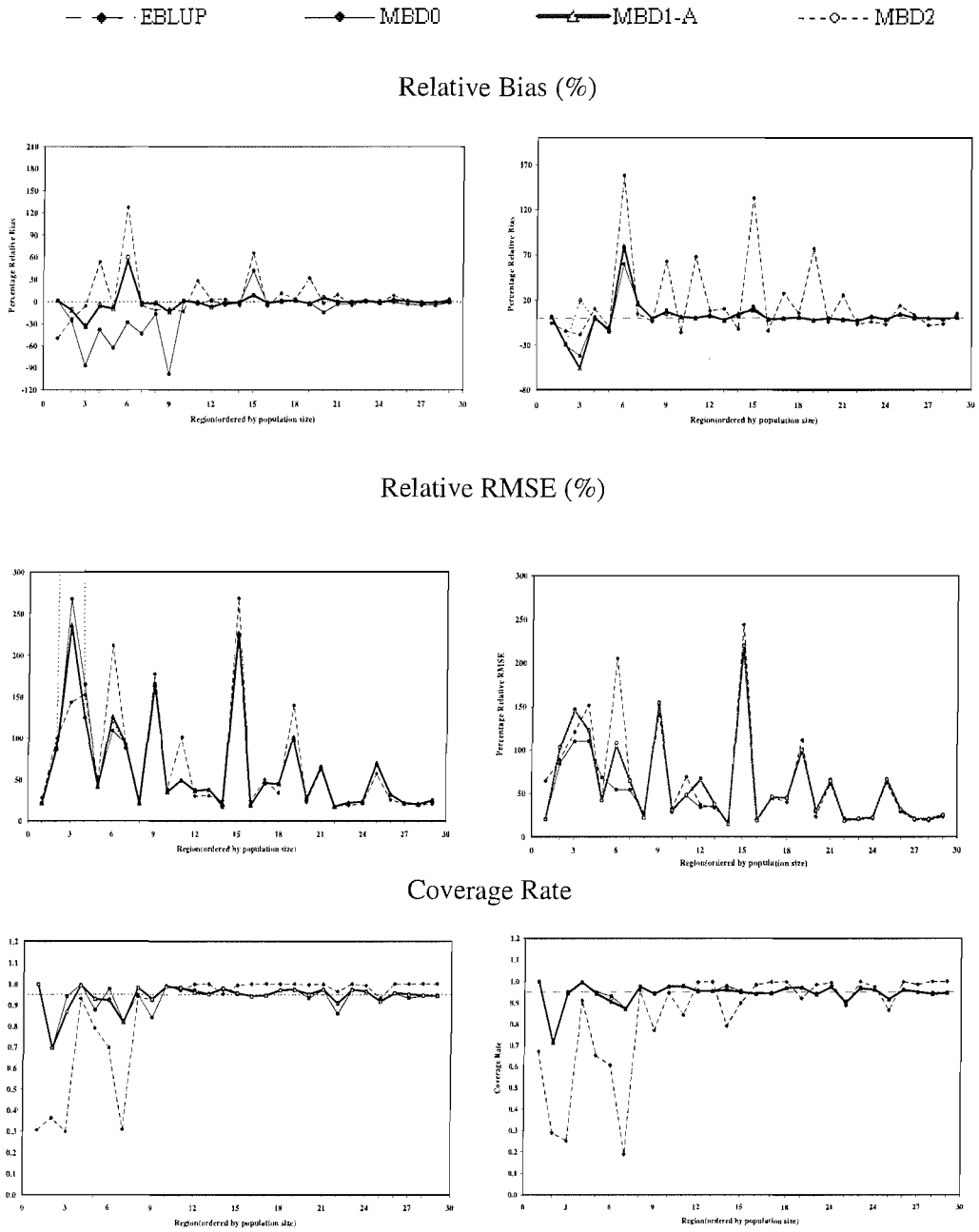


Figure 4.4 Region-specific performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for Cattle under model I (left) and model II (right).

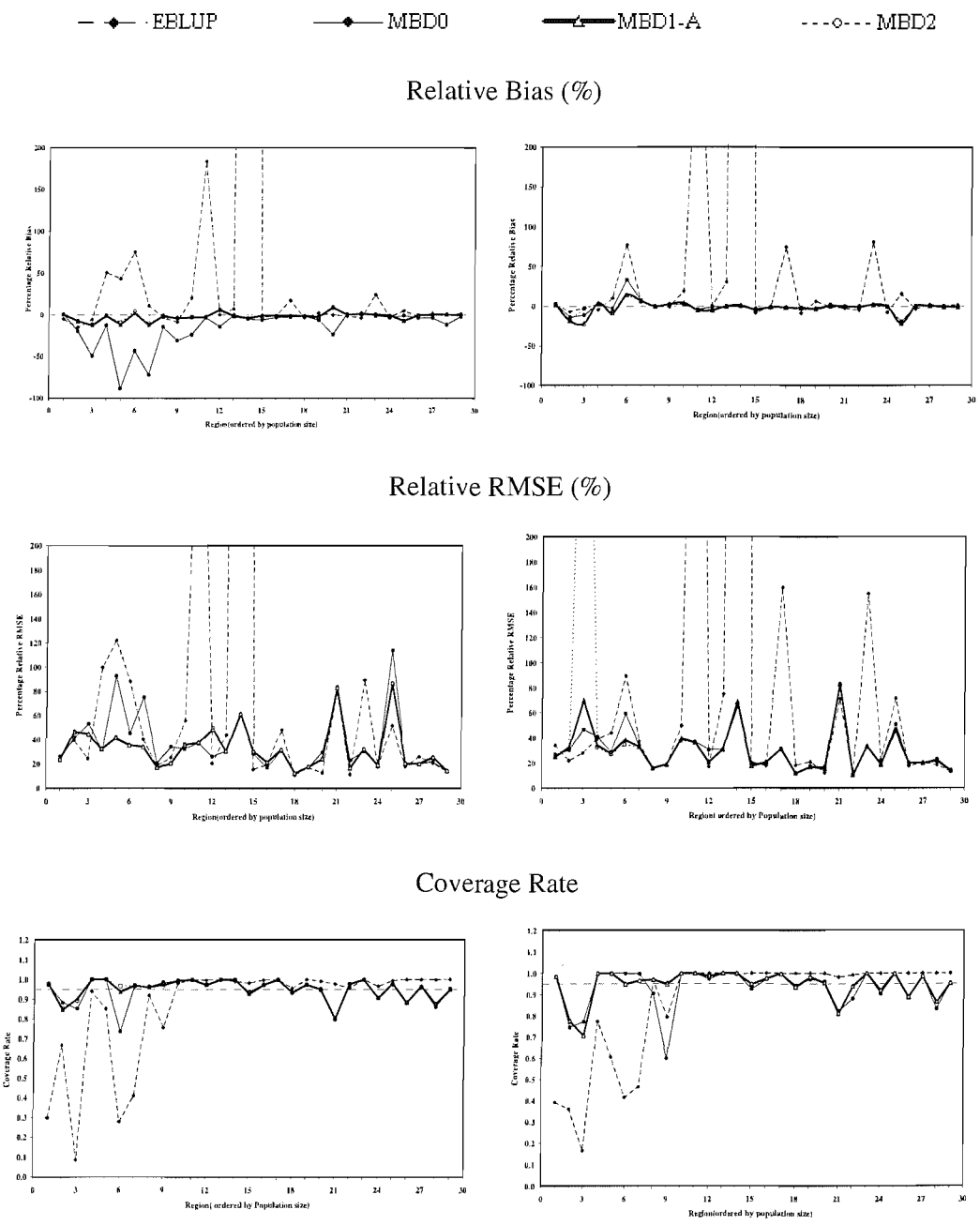
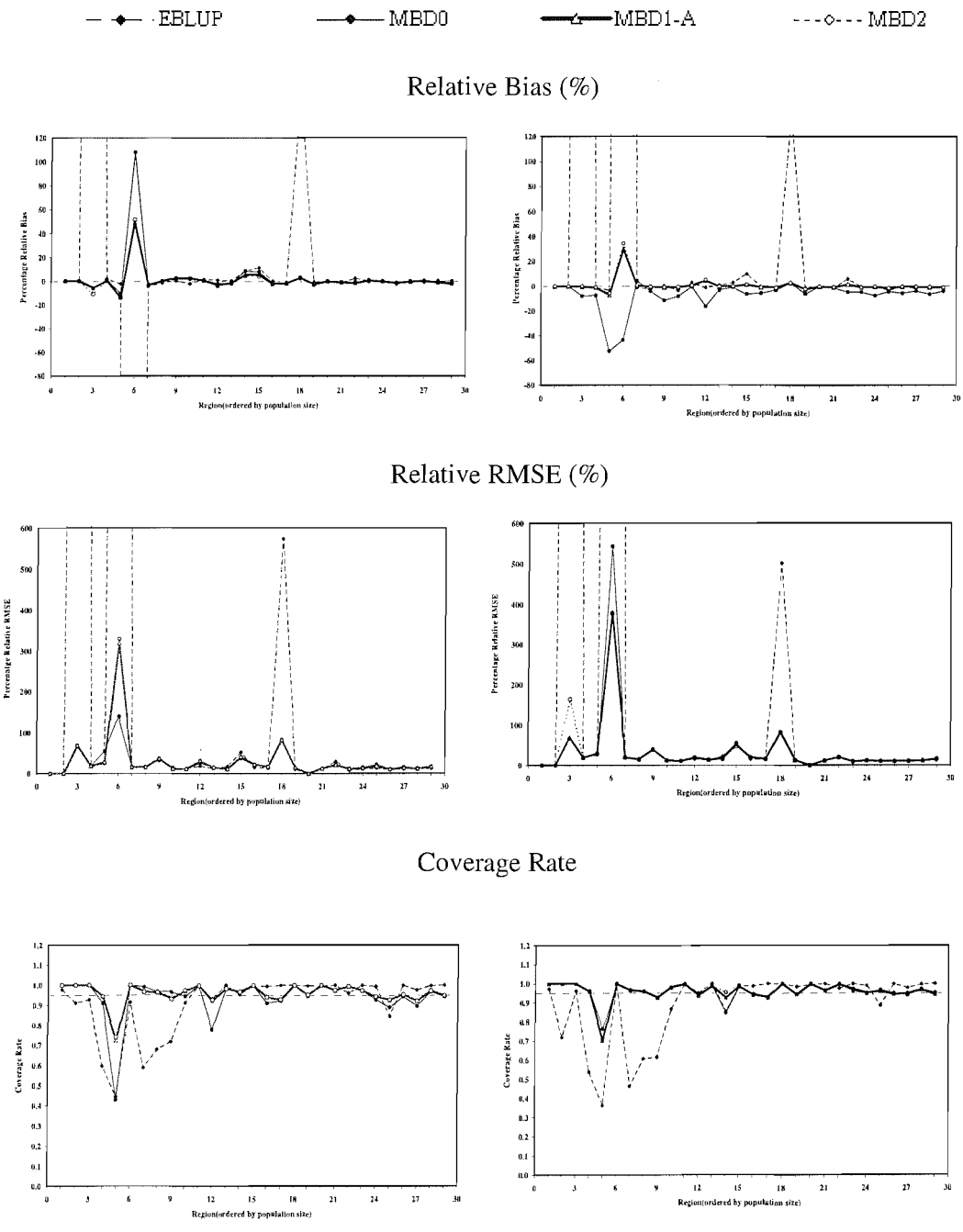


Figure 4.5 Region-specific performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for Sheep under model I (left) and model II (right).



A similar reason lay behind the EBLUP results for Sheep (Figure 4.5). In this case, however, the regions where the zeros occurred were 3 and 18. In particular, in region 3 there were only 11 non-zero values for Sheep in a simulated population of size 189, leading to 223 negative estimates, while in region 18 a majority of zero values for Sheep lead to 323 negative estimates. Further, we noticed that in region 6 all methods are unstable for estimation of Sheep. In this region (region 6) where all procedure fails, in a sample of size 19 all observations are zero expect one which is selected five times in the simulated population of size 465. This non-zero observation is of order 1200, which is like an outlying value. This results in 494 negative estimates for the EBLUP out of 1000 samples and several outlying estimates with weighting methods (MBD0 and MBD1-A).

4.4.3.3 *Third Stage Simulations*

As noted earlier, our results suggest that multipurpose estimation based on MBD1-A is preferable to that based on MBD2. Consequently, in the third stage of simulations we contrast the performances of the variable specific estimators EBLUP and MBD0 with the multipurpose estimator MBD1-A for the three variables (Crops, Equity and Debt) that contain a large number of zeros and are not included in calculation of multipurpose weights (Table 4.4). Note that the results in this simulation stage are based on model I, since model II cannot be fitted to these variables (and also in the stage second we notice no difference in performance of the different methods under model I and II). In this stage of the simulation our purpose is to investigate the applicability of multipurpose weights to variables not included in defining the multipurpose sample weights. Here we examine

how much efficiency will be lost if we apply these multipurpose weights to arbitrary variables from the same survey not included in the definition of the multipurpose weights.

Table 4.4 sets out the average and median relative bias (%), average and median relative root mean squared error and average coverage rate generated by three estimators (EBLUP, MBD0 and MBD1-A) for the three target variables (Crops, Equity and Debt), not included in the multipurpose weights (averaged over the 29 areas). As indicated earlier, the MBD1-A for Crops, Equity and Debt is based on applied multipurpose weights derived using five other variables (TCC, TCR, FCI, Cattle and Sheep).

From Table 4.4 we see that MBD1-A performs marginally better overall. The superior performance of MBD1-A is obvious, as is the poor performance of EBLUP for these variables. The average relative biases under MBD1-A are smaller than MBD0 and EBLUP for Equity and Debt while it is small under MBD0 for Crops. However, the average relative root mean squared error under MBD1-A are lower for the Crops and Equity while for Debt it is lower under MBD0. For Crops and Debt, the average coverage rates of the MBD1-A and MBD0 are same (96 and 93 per cent) but higher than the EBLUP. However, for Equity MBD1-A has the highest coverage rate (94 per cent) overall. These results clearly indicate that the multipurpose weighting based method for small area estimation is the obvious choice for regional estimation, even though the variable is ill-suited for other methods (like EBLUP).

Table 4.4 Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate for EBLUP, MBD0 and MBD1-A for variables with many zeros (Crops, Equity and Debt) under model I. All averages are over the 29 regions of interest.

Criterion	Methods	Crops	Equity	Debt
ARB	EBLUP	90.31	4.36	8.39
	MBD0	0.00	-9.32	-4.94
	MBD1-A	-0.21	-1.20	-0.96
MRB	EBLUP	0.00	-0.28	1.16
	MBD0	-0.84	-3.51	-2.36
	MBD1-A	0.00	-0.32	-0.61
ARRMSE	EBLUP	123.96	18.51	29.02
	MBD0	23.53	19.14	27.71
	MBD1-A	22.92	17.05	28.57
MRRMSE	EBLUP	15.10	12.32	21.49
	MBD0	15.76	16.18	23.70
	MBD1-A	15.80	13.52	24.88
ACR	EBLUP	0.95	0.88	0.91
	MBD0	0.96	0.92	0.93
	MBD1-A	0.96	0.94	0.93

Figure 4.6-4.8 shows the region-specific performance measure for Crops, Equity and Debt respectively generated by three methods (EBLUP, MBD0 and MBD1-A). These region-specific results show some abnormalities in the estimates in few regions. For example, for Crops, the EBLUP method seems to fail in four regions (2, 6, 9 and 18). In these regions we observed the presence of large number of zeros, which gives the negative estimates or under estimates for these regions. As noted earlier, in such cases the EBLUP method is very unstable. In contrast, weighting based methods work reasonably well.

Note that Equity and Debt variables take negative values (also was the case with FCI), and our simulation results examine the application and suitability of different methods with such type of data. For Equity, in three regions (4, 6 and 14) the EBLUP procedure fails, inspection of data indicate the presence of negative values in these regions. For example, in region 4, there are two negative values in original sample and repeated 68 times (first observation 4 times and second 62 times) in the simulated population, which results in negative and under estimates for some of the samples. For Debt, in two regions (3 and 17) only EBLUP and in one region (region 1) all methods are worst, observation of result shows that these are due to under estimation in these regions for most of the sample, due to presence of zero values and outlying. In region 1, where all estimation procedures are affected, in original sample of size 6, there are 5 zeros and one non-zero (y = 19928), which seems to be outlier. In simulated population of size 79, this point was repeated 15 times. Thus, EBLUP method was affected by presence of zeros and the weighting based methods due to outlier. The median relative biases and median relative

RMSE (Table 4.4) show that the dominance of weighting approach and gain due to multipurpose weighting. In regional estimation, multipurpose weighting approach seems to perform well.

4.4.3.4 *Fourth Stage Simulations*

In the results presented so far, the multipurpose weights used in the MBD1-A method have been based on the $K = 5$ target variables that are 'suited' to linear mixed modeling with the model II specification. However, if a model I specification is used, we can use all $K = 8$ target variables to define these weights via (4.8). The aim of this stage is to examine the effect of number of variables in the weight multipurpose. We derived the multipurpose weights based on $K = 8$ variables and compare with those based of $K = 5$ variables. Let us denote by MBD1-A-8Y and MBD1-A-5Y, the MBD1-A estimators based on based on eight and five variables respectively.

Note that the MBD1-A-5Y based on K =5 variables (TCC, TCR, FCI, Cattle, Sheep) are already evaluated for these five variables and also for the rest three variables (Crops, Equity, Debt) with applied weights. At this end, we calculate MBD1-A-8Y estimator based on variables for the entire $K = 8$ variable set (TCC, TCR, FCI, Cattle, Sheep, Crops, Equity, Debt). In Table 4.5 therefore we compare the performance of the MBD1-A method under model I with weights obtained by using both the limited ($K = 5$) and full ($K = 8$) set of target variables in (4.8). Table 4.5 indicates that the relative biases of the MBD1-A-8Y are marginally smaller than the MBD1-A-5Y for all variables except

Sheep. However, the average relative root mean squared errors of the MBD1-A-5Y are marginally lower than the MBD1-A-8Y. The average coverage rates of both the estimators are same.

Table 4.5 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate for multipurpose weighting (MBD1-A) based on original $K = 5$ and extended $K = 8$ variable sets under model I.

Variable	$K = 5$			$K = 8$		
	ARB	ARRMSE	ACR	ARB	ARRMSE	ACR
TCC	-1.54	20.86	0.92	-1.08	20.91	0.92
TCR	-1.30	21.77	0.92	-0.80	21.83	0.92
FCI	-0.50	59.72	0.94	0.21	60.22	0.94
Cattle	-1.78	33.29	0.95	-1.05	33.49	0.95
Sheep	0.69	30.24	0.96	1.24	31.06	0.96
Crops	-0.21	22.92	0.96	-0.20	22.97	0.96
Equity	-1.20	17.05	0.94	-0.72	17.14	0.94
Debt	-0.96	28.57	0.93	-0.68	28.74	0.93

These results in Table 4.5 show that there is little change in the average performance of MBD1-A when the set of variables determining the multipurpose weights used by this estimator is extended from the original $K = 5$ variable set (TCC, TCR, FCI, Cattle, Sheep) to the entire $K = 8$ variable set (TCC, TCR, FCI, Cattle, Sheep, Crops, Equity, Debt). Again, note that this extension is only possible under Model I. Moreover, it is worth noting that for last three variables (Crops, Equity and Debt), not included in the

weight under MBD1-A-5Y and included in weights under MBD1-A-8Y. Overall, this result shows that these weights are quite insensitive to this choice. The almost imperceptible regional difference between the estimates defined by these two sets of weights (see Figure 4.6-4.8) reinforces this observation for these variables (Crops, Equity and Debt).

Figure 4.6-4.8 shows that region-specific performance measure generated by EBLUP, MBD0, MBD1-A-5Y and MBD1-A-8Y methods for Crops, Equity and Debt respectively. As indicated earlier, in Figure 4.8 we show the overall region-specific superior performance of MBD1-A (under either K = 5 or K = 8) for the variable Debt. Similar region-specific performances were observed for Crops and Equity as well.

4.4.3.5 *Fifth Stage Simulations*

So far, when computing the multipurpose weights, we have assigned equal importance to all K target variables that are used to define them. However, a reasonable alternative approach would be to assign importance factors based on the intrinsic variability of these variables (see section 4.4.2, page 100). Two natural options in this regard are $\phi_k = 1/\Sigma_{e,k}$ and $\phi_k = 1/V_k$, where $\Sigma_{e,k}$ and V_k are the individual and total variability of the k^{th} target variable. In this stage, we examine the effect of assigning relative importance of the variables included in the multipurpose weights. Here we denote by MBD1-A ($\phi_k = 1/\Sigma_{e,k}$) and MBD1-A ($\phi_k = 1/V_k$) as the MBD1-A methods with relative weight $\phi_k = 1/\Sigma_{e,k}$ and $\phi_k = 1/V_k$ respectively.

Table 4.6 provides summary details of the performance of the MBD1-A method when the multipurpose weights (based on TCC, TCR, FCI, Cattle and Sheep) are computed using these alternative importance weighting factors. These results show that the average relative bias increases for all variables except Sheep, the average relative root mean squared error reduces and the average coverage rate remains same for all variables by incorporating the variability of the target variables in the multipurpose weights. Overall we see that, for the population considered in the simulation study, there is little to choose between these different importance weighting factors.

Table 4.6 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%), and average (ACR) coverage rate for multipurpose weighting (MBD1-A) under $\phi_k = 1/K$, $\phi_k = 1/\sigma^2_{e,k}$ and $\phi_k = 1/V_k$ for K = 5 target variables (TCC, TCR, FCI, Cattle, Sheep) under model I.

Criterion	ϕ_k^{-1}	TCC	TCR	FCI	Cattle	Sheep
ARB	K	-1.54	-1.30	-0.50	-1.78	0.69
	$\sigma^2_{e,k}$	-1.69	-1.48	-0.82	-2.03	0.52
	V_k	-1.64	-1.42	-0.70	-1.95	0.57
ARMSE	K	20.86	21.77	59.72	33.29	30.24
	$\sigma^2_{e,k}$	20.83	21.71	58.00	33.19	29.99
	V_k	20.85	21.75	58.15	33.25	30.11
ACR	K	0.92	0.92	0.94	0.95	0.96
	$\sigma^2_{e,k}$	0.92	0.92	0.94	0.95	0.96
	V_k	0.92	0.92	0.94	0.95	0.96

Figure 4.6 Regional performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A under $K = 5$ (thick line) and MBD1-A under $K = 8$ (dotted line) for Crops under model I.

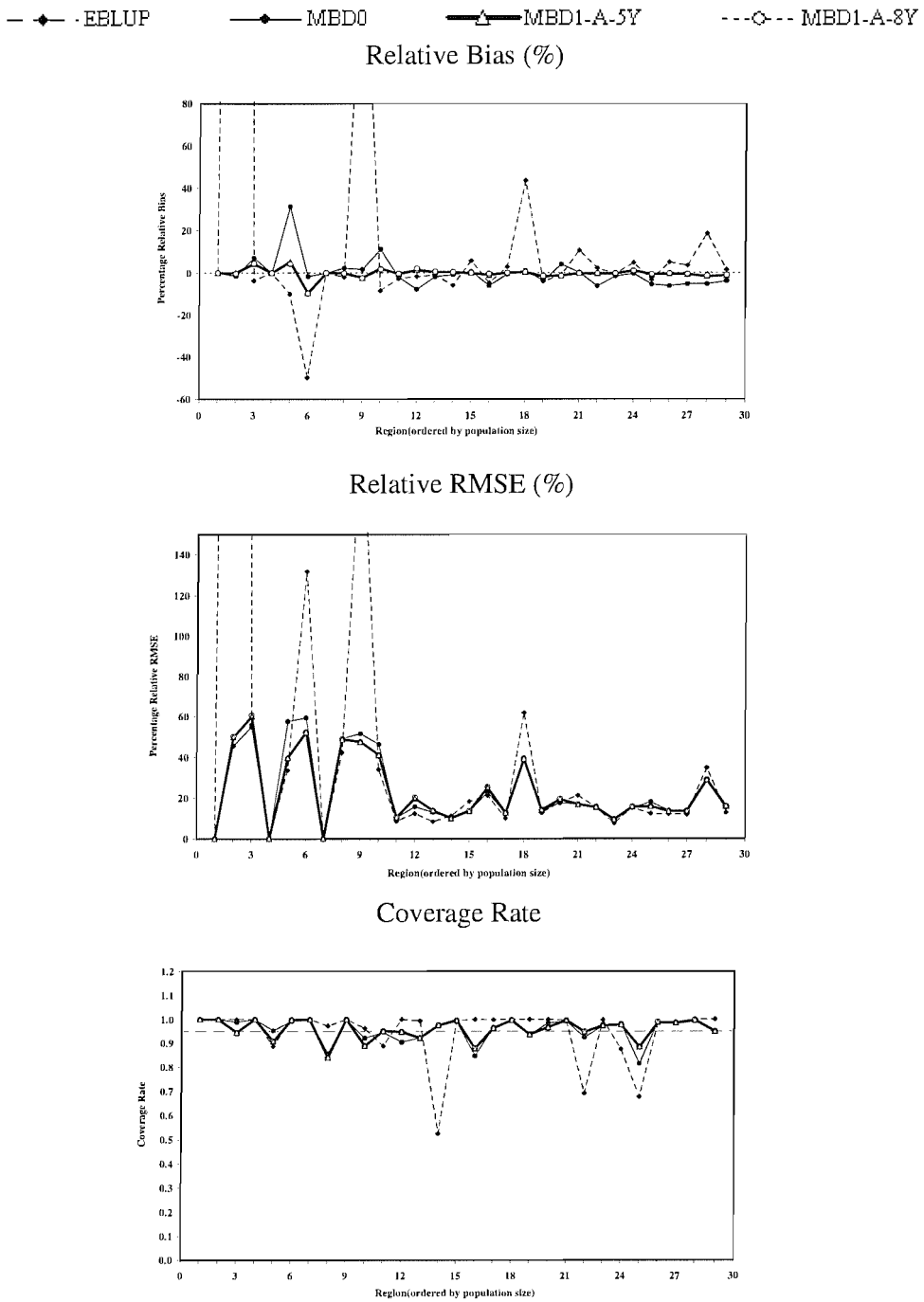


Figure 4.7 Regional performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A under $K = 5$ (thick line) and MBD1-A under $K = 8$ (dotted line) for Equity under model I.

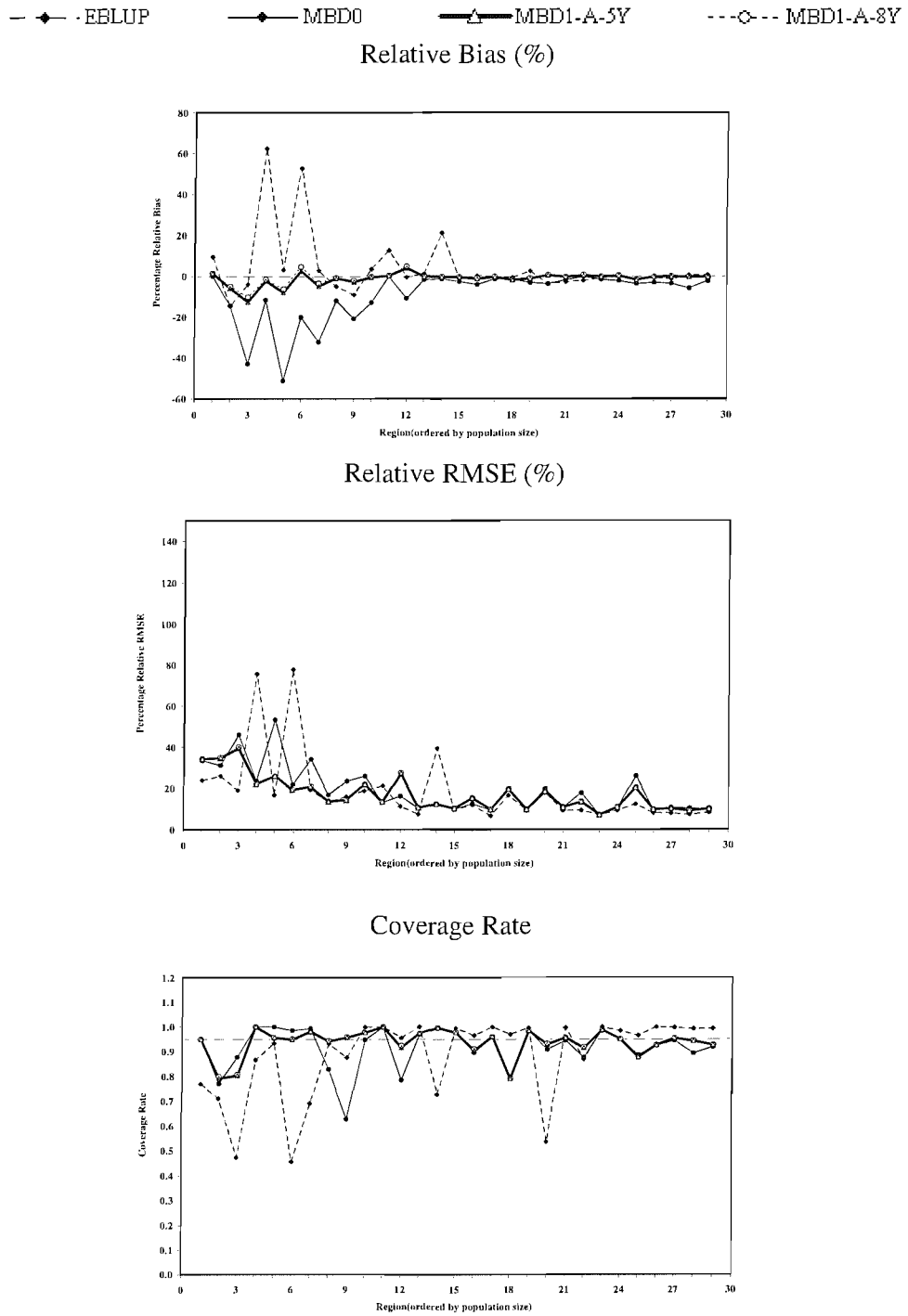
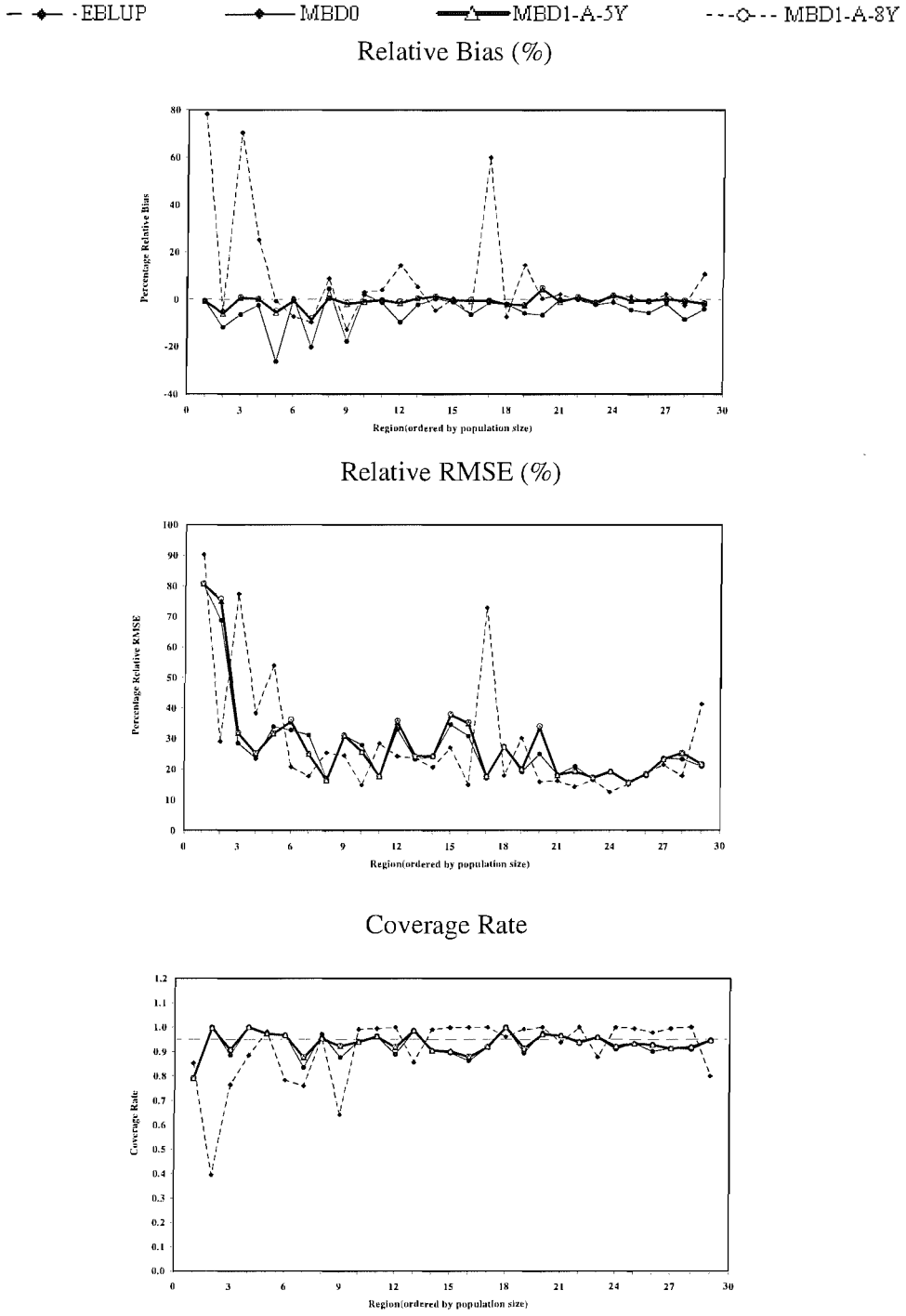


Figure 4.8 Regional performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A under $K = 5$ (thick line) and MBD1-A under $K = 8$ (dotted line) for Debt under model I.



4.5 Conclusions

In this chapter we develop two loss functions that can be used to compute optimal multipurpose weights suitable for use in SAE using MBD estimators. The first (4.8) ignores the correlations between the survey variables, while the second (4.12) takes these into account. For the population considered in our simulation studies the performance of the corresponding multipurpose weighting based MBD1-A and MBD1-B estimators are almost identical, i.e. there are no real gains from taking account of the correlations between the survey variables when constructing the multipurpose weights. We also investigated an alternative approach to constructing multipurpose weights for use in MBD methods of SAE by suitably averaging the variable specific EBLUP weights. Here again, our empirical results demonstrate that this method is somewhat less efficient than the loss function based MBD1-A method. We also show that these multipurpose weights remain efficient across a wide range of variables, even variables that have not been used in the definition of the multipurpose weights. This can be important in some situations (e.g. where variables have many zero values) where standard mixed models cannot be fitted and the usual EBLUP methods do not work. An alternative in such cases is extend the EBLUP approach to mixtures of linear mixed models.

CHAPTER 5

SMALL AREA ESTIMATION FOR SKEWED DATA

5.1 Introduction

Commonly used methods for small area estimation (SAE) assume that a linear mixed model (3.11) can be used to characterize the relationship between the survey variable Y and an auxiliary variable X in the small areas of interest. In particular, empirical best linear unbiased prediction (EBLUP, see Prasad and Rao, 1990) and model-based direct (MBD, see chapter 3-4 and Chandra and Chambers, 2005) estimation are typically based on the linear model assumptions. However, when the data are skewed, as is often the case in business surveys, the relationship between Y and X may not be linear in the original (or raw) scale, but can be linear in a transformed scale, e.g. the logarithmic scale. In such cases we would expect estimation based on a linear model for Y to be inefficient, and an appropriate technique for SAE should then be based on a linear mixed model for a transformed version of Y. See Hidiroglou and Smith (2005). Choice of an appropriate transformation function plays an important role in the transformed variable based estimation methods. Practically, it should be selected by examining the data for possible model relationship. The use of transformed variables for survey estimation with skewed data has been investigated by Carroll and Ruppert (1988), Chen and Chen (1996), Karlberg (2000) and Chambers and Dorfman (2003).

In this chapter we explore the use of transformed variable based estimation when carrying out small area estimation for skewed data, focussing on the widely used log-log transformation. Implementation of the EBLUP approach under transformation to a linear mixed model is complicated. However, this is not the case with the MBD approach. In particular, we extend the MBD approach described in chapter 3 to small area estimation for skewed data using sample weights derived via model calibration (Wu and Sitter, 2001). Our approach assumes a log-log transform linear model with random area effects.

In the next section we summarize the model calibration approach for the estimation of population level quantities. In section 5.3 we introduce the concept of an 'expected value' (or 'fitted value') model derived from a transformed linear mixed model. In section 5.4 we derive optimal model-based survey weights based on this 'expected value' model and use them in an MBD estimator for SAE. A simple MSE estimator for weighted SAE is also described. We also relax the usual normality assumption for the random area effects in order to examine robustness with respect to this assumption. Finally, section 5.5 presents some concluding remarks.

5.2 Model Calibration Weighting for Population Estimation

The calibrated sample weights described in section 3.2 of chapter 3 implicitly assume that the survey variable Y and auxiliary variables X are linearly related. If the underlying model is non-linear, the calibration estimator derived under linearity assumption can be inefficient. In such cases, Wu and Sitter (2001) proposed the model calibration approach

which generalizes the calibration procedure under a general model. In this section we briefly review the model calibration for the estimation of population level quantities.

To start, we fix our notation. Let U denote a population of size N and let Y_U denote the N-vector of population values of a characteristic Y of interest. Suppose that our primary aim is estimation of the total $T_y = \sum_U y_j$ of these population values (or their mean $\bar{Y}_U = N^{-1}\sum_U y_j$). Let X_U denote the $N \times p$ matrix of population values of p-auxiliary variables X that are related, in some sense, to the values in Y_U. We assume that the individual sample values of X_U are known. The non-sample values of X_U may not be individually known, but are assumed known at some aggregate level. At a minimum, we know the vector of population totals T_x of the columns of X_U. Given this set up, Deville and Särndal (1992) define an X_U-calibrated linear estimator of T_y as $\hat{T}_y = \sum_{j\in s} w_j y_j$, where s denotes the n sample units, and the calibrated weights $\{w_j; j \in s\}$ satisfy $\sum_{j\in s} w_j x_j = T_x$. Assuming that s is a probability sample based on first order inclusion probabilities π_j, they recommend that the vector of calibrated weights be chosen so as to minimise an appropriate measure of its distance from the corresponding vector of design weights $d_j = \pi_j^{-1}$, subject to the constraint $\sum_{j\in s} w_j x_j = T_x$. Their justification for this approach is based on an implicit assumption that the population values of Y and X are linearly related, in which case the calibration constraint is equivalent to ensuring that the estimator $\hat{T}_y$ is an unbiased predictor of T_y under a linear model for the regression of Y on X in the population.

If the underlying population model is non-linear, the calibration estimator $\hat{T}_y$ can be model-biased, and hence inefficient. In particular, suppose that the relationship between Y and X in the population is of form

$$E(y_j \mid x_j) = h(x_j;\eta) \text{ and } Var(y_j \mid x_j) = \sigma_j^2 ; j = 1,...,N \tag{5.1}$$

where η (typically vector-valued) and σ_j^2 are unknown model parameters and the mean function $h(x_j;\eta)$ is a known function of x_j and η. Let Y_U denote the N-vector of population values of Y and suppose that the population units are mutually uncorrelated. We can then express (5.1) in matrix form as

$$E(Y_U \mid X_U) = h(X_U;\eta) \text{ and } Var(Y_U \mid X_U) = \Sigma = diag(\sigma_j^2; j = 1,\ldots,N) \tag{5.2}$$

where it is understood that $h(X_U;\eta)$ is the N-vector with components $h(x_j;\eta)$. The model (5.2) is quite general and includes linear, non-linear, and generalized linear models as special cases. In this context, Wu and Sitter (2001) propose the use of sample weights derived via model calibration, where they define the model-calibrated estimator of the population mean of Y as $\hat{\bar{Y}}_U^{mc} = N^{-1}\left(\sum_{j\in s} w_j^{mc} y_j\right)$ with the vector of weights w_j^{mc} again minimising distance from the vector of design weights, but this time subject to the constraints

$$\sum_{j\in s} w_j^{mc} = N \text{ and } \sum_{j\in s} w_j^{mc} h(x_j;\hat{\eta}_\pi) = \sum_{j\in U} h(x_j;\hat{\eta}_\pi) \tag{5.3}$$

where $\hat{\eta}_\pi$ is a design consistent estimator of η. Note that unlike standard calibration, the model calibration constraints (5.3) typically require that we know the individual population values of X. The calibration is performed with respect to the population total of the fitted values $h(x_j;\hat{\eta}_\pi) = \hat{h}_j$ of $h(x_j;\eta)$. The key idea behind this approach is that

provided the model (5.2) is a reasonable one, y_j is then (at least approximately) a linear function of its 'fitted values' $h(x_j;\hat{\eta}_\pi)$ under this model and so we can carry out linear estimation using the population values of these fitted values as auxiliary information. The calibration constraints (3.3) consist of p-equations, where p is the number of components in X_U, whereas constraint (5.3) has only one equation involving the single data reduction variable $h(x_j;\eta)$. Under this set-up, the model calibration estimator for the population mean $\bar{Y}_U = N^{-1}\sum_U y_j$ is

$$\hat{\bar{Y}}_U^{mc} = N^{-1}\sum_{j\in s} w_j^{mc} y_j = \hat{\bar{Y}}_U^{HT} + N^{-1}\left\{\sum_{j\in U} h(x_j;\hat{\eta}_\pi) - \sum_{j\in s} d_j h(x_j;\hat{\eta}_\pi)\right\}\hat{B}_1 \qquad (5.4)$$

where

$$\hat{B}_1 = \left(\sum_{j\in s} d_j q_j (\hat{h}_j - \bar{h})^2\right)^{-1}\left\{\sum_{j\in s} d_j q_j (\hat{h}_j - \bar{h})(y_j - \bar{y})\right\} \text{ with}$$

$$\bar{y} = \left(\sum_{j\in s} d_j q_j\right)^{-1}\left(\sum_{j\in s} d_j q_j y_j\right),\ \bar{h} = \left(\sum_{j\in s} d_j q_j\right)^{-1}\left(\sum_{j\in s} d_j q_j \hat{h}_j\right),$$

and $\hat{\bar{Y}}_U^{HT} = N^{-1}\sum_{j\in s} d_j y_j$ is the Horvitz-Thompson (HT) estimator for the population mean $\bar{Y}_U$, q_j's are known positive weights unrelated to d_j.

If the constraint $\sum_{j\in s} w_j^{mc} = N$ is dropped, with single calibration constraint $\sum_{j\in s} w_j^{mc} h(x_j;\hat{\eta}_\pi) = \sum_{j\in U} h(x_j;\hat{\eta}_\pi)$, the calibration estimator for population mean $\bar{Y}_U$ is

$$\hat{\bar{Y}}_U^{mc} = \hat{\bar{Y}}_U^{HT} + N^{-1}\left\{\sum_{j\in U} h(x_j;\hat{\eta}_\pi) - \sum_{j\in s} d_j h(x_j;\hat{\eta}_\pi)\right\}\hat{B}_2 \qquad (5.5)$$

where

$$\hat{B}_2 = \left(\sum_{j\in s} d_j q_j \hat{h}_j^{\,2}\right)^{-1}\left(\sum_{j\in s} d_j q_j \hat{h}_j y_j\right).$$

The above discussion represents what might be referred to a design-based interpretation of model calibration. A corresponding model-based perspective on the model calibration follows directly. See Chambers (2005). Let $\hat{\eta}$ denote a 'model-efficient' estimator of η in (5.2) with associated fitted values $h(x_j;\hat{\eta})$. In general, these fitted values will not be unbiased. However, there will still be a systematic relationship between the actual values of Y and their corresponding fitted values that we can approximate. Although there is nothing to stop us looking at more complex approximations, a linear model for the relationship between the population values y_j of Y and the fitted values $\hat{y}_j = h(x_j;\hat{\eta})$ seems a reasonable starting point. We therefore replace (5.1) by a linear model of the form

$$E(y_j \mid \hat{y}_j) = \alpha_0 + \alpha_1 \hat{y}_j \text{ and } Cov(y_j, y_k \mid \hat{y}_j, \hat{y}_k) = \omega_{jk} \tag{5.6}$$

We refer to (5.6) the 'fitted value' or the 'expected value' (interchangeably used) linear model defined by (5.1). Setting $\alpha_0 = 0$ in (5.6) corresponds to a *ratio* specification for this fitted value linear model. Generally, estimation bias implies $\alpha_0 \neq 0$, in which case (5.6) corresponds to a *regression* specification for this model. Let J_U denote the population 'design matrix' defined by (5.6) under either of these specifications. Without loss of generality, we arrange the vector Y_U so that its first n elements correspond to the sample units, and then partition Y_U, J_U and $\Omega_U = [\omega_{jk}]$ according to sample and non-sample units as

$$Y_U = \begin{bmatrix} Y_s \\ Y_r \end{bmatrix},\ J_U = \begin{bmatrix} J_s \\ J_r \end{bmatrix} \text{ and } \Omega_U = \begin{bmatrix} \Omega_{ss} & \Omega_{sr} \\ \Omega_{rs} & \Omega_{rr} \end{bmatrix}.$$

Here a subscript of s denotes components defined by the n sample units while a subscript of r is used to denote corresponding components defined by the remaining $N-n$ non-sample units. In practice the variance components that define Ω are unknown and so need to be estimated from the sample data. We use a 'hat' to denote such an estimate below. Also, we use 1_U, 1_s and 1_r to denote vectors of 1's of the appropriate size, and I_U, I_s and I_r to denote identity matrices of order N, n and $N-n$ respectively. We also assume that sampling is uninformative, so the sample data follow the population model.

Given this notation, the sample weights that define the Empirical Best Linear Unbiased Predictor (EBLUP) for population total of Y under the general *linear* 'fitted value' model (5.6) are

$$w^{mc,EBLUP} = (w_j^{mc,EBLUP}) = 1_s + H'_{mc}(J'_U 1_U - J'_s 1_s) + (I_s - H'_{mc}J'_s)\hat{\Omega}_{ss}^{-1}\hat{\Omega}_{sr}1_r \qquad (5.7)$$

where $H_{mc} = (J'_s\hat{\Omega}_{ss}^{-1}J_s)^{-1}J'_s\hat{\Omega}_{ss}^{-1}$. See Royall (1976). It is easy to see that the weights (5.7) are model-calibrated under (5.6) since $J'_s w^{mc,EBLUP} = J'_U 1_U$. That is, if a regression specification is used for (5.6) then

$$\sum_{j\in s} w_j^{mc,EBLUP} = N \text{ and } \sum_{j\in s} w_j^{mc,EBLUP}\hat{y}_j = \sum_{j\in U}\hat{y}_j .$$

Note that the weights (5.7) are *not* the same as the weights that define the standard EBLUP for the population total of Y under a linear model for the regression of Y_U on X_U in the population described in chapter 3. These weights are given by

$$w^{EBLUP} = (w_j^{EBLUP}) = 1_s + H'(X'_U 1_U - X'_s 1_s) + (I_s - H'X'_s)\hat{V}_{raw,ss}^{-1}\hat{V}_{raw,sr}1_r \qquad (5.8)$$

where $H = (X'_s\hat{V}_{raw,ss}^{-1}X_s)^{-1}X'_s\hat{V}_{raw,ss}^{-1}$ and $\hat{V}_{raw}$ denotes an estimate of $Var(Y_U \mid X_U) = V_{raw}$. Here a subscript 'raw' is used to denote the variance matrix related to raw-scale linear

model as defined in chapter 3. The sample weights (5.8) define the EBLUP for the population total of Y and calibrated on X, in the sense that they exactly reproduce the known population totals defined by the columns of X_U. That is $X_s' w^{EBLUP} = X_U' 1_U = T_x$.

5.3 Small Area Estimation under Transformation

Direct linear estimators for small areas, i.e. estimators that are defined as weighted sums of the sample data from the small areas of interest, have a number of practical advantages, including simplicity of construction and aggregation consistency. In chapter 3, we used the EBLUP weights (5.8) to construct the model-based direct (MBD) estimators for small areas when a linear model assumption is appropriate for the population as a whole. Unlike the design-based weights used in more conventional direct estimators, the weights used in an MBD estimator are based on assuming that a linear mixed model with random area effects holds in the small areas of interest. In this section we extend this approach, exploring the use of MBD estimators based on the model-calibrated EBLUP weights (5.7) for SAE, given that the population data are skewed, but can be transformed to linearity.

5.3.1 A Log-Scale Linear Mixed Model

Linear mixed models (3.11) are popular in SAE. Here we consider the situation where such a model is inappropriate for Y in its original scale, but is appropriate for a suitably transformed version of this variable. In particular where both Y and X are scalar and

strictly positive, with highly skewed population marginal distributions and clear evidence of non-linearity in their relationship, e.g. as in many business surveys applications, but where a linear mixed model holds for the regression of $\log(Y)$ on $\log(X)$. That is, we assume that

$$l_{ij} = \log(y_{ij}) = \beta_0 + \beta_1 \log(x_{ij}) + G'_{ij} u_i + e_{ij} \tag{5.9}$$

where y_{ij} and x_{ij} are the values of Y and X respectively for population unit $j(j = 1,...,N_i)$ in small area $i(i = 1,...,m)$, G_{ij} denotes a covariate of dimension q, u_i denotes a random effect for area i also of dimension q and e_{ij} is a scalar individual random effect. Here N_i is the population size for area i and m is the total number of areas. As usual with this type of model, we assume that all random effects are normally distributed and mutually uncorrelated, with zero expected values, $Var(u_i) = \Sigma(\theta)$ and $Var(e_{ij}) = \sigma_e^2$. Here $\Sigma(\theta)$ is a known matrix-valued function of an unknown vector-valued parameter θ. It follows that $Cov(l_{ij}, l_{ik} \mid x_{ij}, x_{ik}) = G'_{ij}\Sigma(\theta)G_{ik} + I(j = k)\sigma_e^2$ and so the covariance matrix of the vector $l_i = (l_{ij})$ defined by the N_i values of l_{ij} in area i is $V_i = G_i\Sigma(\theta)G'_i + \sigma_e^2 I_{N_i}$, where G_i is the $N_i \times q$ matrix defined by the covariates G_{ij} in area i and I_{N_i} is the identity matrix of order N_i. The model (5.9) is identical to the model (3.11) defined in section 3.3 of chapter 3. However, the model (5.9) is defined on transformed scale and used in slightly different context (e.g. derivation of bias adjustment due to transformation etc) so to maintain continuity we sometimes repeat some of these expressions.

Let $\log(X_i)$ denote the vector of N_i values of $\log(X)$ in area i. Put $W_i = [1_{N_i} \log(X_i)]$, where 1_{N_i} denotes a vector of 1s of dimension N_i, and denote $e_i = (e_{ij})$. By aggregating the area-specific model (5.9) over the m small areas that make up the population, we are led to the population level linear mixed model on log-scale

$$l_U = W_U \beta + G_U u + e \tag{5.10}$$

$u = (u_1', \ldots, u_m')'$ and $e = (e_1', \ldots, e_m')'$. Note that under (5.10), the covariance matrix of l_U is $V_U = diag(V_i; 1 \le i \le m)$.

The model (5.10) includes most of the small area models used in the literature (Rao, 2003, page 107). In practice the variance components θ and σ_e^2 that define the covariance matrix V_U are unknown and have to be estimated from the sample data, e.g. via maximum likelihood (ML), restricted maximum likelihood (REML) or method of moments (Harville, 1977). Using a 'hat' to denote such estimates, we can then estimate V_U by $\hat{V}_U = diag(\hat{V}_i; 1 \le i \le m)$ with $\hat{V}_i = [\hat{v}_{ijk}] = G_i \Sigma(\hat{\theta}) G_i' + \hat{\sigma}_e^2 I_{N_i}$. We can also decompose l_U, W_U, G_U and $\hat{V}_U$ into sample and non-sample components within each small area. If we introduce an extra subscript of i to indicate small area (e.g. we denote by s_i the set of n_i sample units in area i, r_i the corresponding $N_i - n_i$ non-sampled units in the area), the empirical best linear unbiased estimator (EBLUE) of β under (5.10) is then

$$\hat{\beta} = \left(\sum_{i=1}^{m} W_{is}' \hat{V}_{iss}^{-1} W_{is} \right)^{-1} \left(\sum_{i=1}^{m} W_{is}' \hat{V}_{iss}^{-1} l_{is} \right) \tag{5.11}$$

where $\hat{V}_{iss} = G_{is} \Sigma(\hat{\theta}) G_{is}' + \hat{\sigma}_e^2 I_{n_i}$ and $\hat{V}_{isr} = G_{is} \Sigma(\hat{\theta}) G_{ir}'$. Here I_{n_i} is the identity matrix of order n_i, the number of sample units in area i. Note that when the variance components

θ and σ_e^2 are known, the EBLUE (5.11) becomes the BLUE for β. Consequently, for large sample sizes we can write $E(\hat{\beta}) \approx \beta$ and $Var(\hat{\beta}) \approx \left(\sum_{i=1}^{m} W_{is}' \hat{V}_{iss}^{-1} W_{is}\right)^{-1}$. The $\hat{\beta}$ in (5.11) is different from one used in chapter 3 since here underlying model (5.10) is linear on transformed scale. Put $\hat{\phi}_i = (\hat{\phi}_{ij}) = W_i \hat{\beta}$. Then $E(\hat{\phi}_i) \approx W_i \beta$ and

$$Var(\hat{\phi}_i) = A_i = [a_{ijk}] \approx W_i \left(\sum_{g=1}^{m} W_{gs}' \hat{V}_{gss}^{-1} W_{gs}\right)^{-1} W_i' = W_i Var(\hat{\beta}) W_i'$$

where, as $n \to \infty$, $a_{ijk} = W_{ij}' Var(\hat{\beta}) W_{ik} \to 0$ $(i = 1,...,m;\ \ j,k = 1,...,N_i)$. We denote by $a_i = (a_{i11}....a_{iN_iN_i})'$ and $v_i = (v_{i11}...v_{iN_iN_i})'$, the vectors of diagonal elements of the covariance matrices $Var(\hat{\phi}_i)$ and $Var(l_i)$ respectively.

5.3.2 An Expected Value Model for Small Area Estimation

In order to use the MBD method for SAE we require sample weights that reflect the population heterogeneity induced by the small area effects. For skewed data that follow a non-linear mixed model, these weights can be derived via (model-based) model calibration. Consequently, we first define an appropriate fitted value model for our data (see section 5.2). From the development in the previous section it is clear that such a model should be based on fitted values derived from the log-scale linear mixed model (5.10). In particular, we need the first and second order moments of these fitted values before we can use (5.7) to define an appropriate set of model-calibrated weights.

A simple method of defining the fitted values under (5.10) is one where we use the parameter estimates derived under this model to obtain predicted values on the log scale and then back-transform to get the predicted values of *Y*. Unfortunately, this approach is biased (Chambers and Dorfman, 2003). We therefore now develop the first and second order moments of an appropriate bias-corrected fitted value model based on (5.10). To derive the 'fitted value' model from transform scale linear mixed model (5.9) it is important to specify the distribution of random errors. Here we consider both normal and non-normal distributions for these random errors.

5.3.2.1 *Normal Distribution for Random Errors*

Assuming that random errors are normally distributed, we note that under (5.9)

$$E(y_{ij} \mid x_{ij}) = E\left\{\exp(l_{ij}) \mid x_{ij}\right\} = e^{W_{ij}'\beta + v_{ijj}/2} \neq E\left(e^{\hat{\phi}_{ij} + \hat{v}_{ijj}/2}\right) \tag{5.12}$$

which shows that a simple bias correction based on the marginal lognormal distribution of Y is inadequate. That is the naïve-lognormal predictor is biased. We need a more sophisticated bias correction procedure. Let $\hat{\eta}_{ij} = (\hat{\beta}, \hat{v}_{ijj})'$ be an estimate of $\eta_{ij} = (\beta, v_{ijj})'$ such that $E(\hat{\eta}_{ij} - \eta_{ij}) \approx 0$ for large n. Put $z(\eta_{ij}) = e^{W_{ij}'\beta + v_{ijj}/2}$. Using a second order Taylor series approximation we can write

$$z(\hat{\eta}_{ij}) \approx z(\eta_{ij}) + (\hat{\eta}_{ij} - \eta_{ij})' z^{(1)}(\eta_{ij}) + \frac{1}{2}(\hat{\eta}_{ij} - \eta_{ij})' z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})$$

and so

$$E\left\{z(\hat{\eta}_{ij})\right\} \approx z(\eta_{ij}) + \frac{1}{2} tr\left[E\left\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\right\}\right].$$

Here

$$z^{(1)}(\eta_{ij}) = \frac{\partial z(\eta_{ij})}{\partial \eta_{ij}} = \left(\frac{\partial z(\eta_{ij})}{\partial \beta} \quad \frac{\partial z(\eta_{ij})}{\partial v_{ijj}} \right)' = \left(W_{ij}' e^{W_{ij}'\beta + v_{ijj}/2} \quad \frac{1}{2} e^{W_{ij}'\beta + v_{ijj}/2} \right)'$$

and

$$z^{(2)}(\eta_{ij}) = \frac{\partial^2 z(\eta_{ij})}{\partial \eta_{ij} \partial \eta_{ij}'} = \begin{pmatrix} W_{ij} W_{ij}' e^{W_{ij}'\beta + v_{ijj}/2} & \frac{1}{2} W_{ij} e^{W_{ij}'\beta + v_{ijj}/2} \\ \frac{1}{2} W_{ij}' e^{W_{ij}'\beta + v_{ijj}/2} & \frac{1}{4} e^{W_{ij}'\beta + v_{ijj}/2} \end{pmatrix}$$

are the vector and matrix respectively containing the first and second order derivatives of $z(\eta_{ij})$ with respect to η_{ij}. Since $\hat{\beta}$ and $\hat{v}_{ijj}$ are independent (McCulloch and Searle, 2001), we have

$$tr\left[E\left\{ z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})' \right\} \right] = tr\left[z^{(2)}(\eta_{ij}) E\left\{ (\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})' \right\} \right]$$

$$= tr\left\{ \begin{pmatrix} W_{ij} W_{ij}' e^{W_{ij}'\beta + v_{ijj}/2} & \frac{1}{2} W_{ij} e^{W_{ij}'\beta + v_{ijj}/2} \\ \frac{1}{2} W_{ij}' e^{W_{ij}'\beta + v_{ijj}/2} & \frac{1}{4} e^{W_{ij}'\beta + v_{ijj}/2} \end{pmatrix} \begin{pmatrix} V(\hat{\beta}) & 0 \\ 0 & V(\hat{v}_{ijj}) \end{pmatrix} \right\}$$

$$\approx e^{W_{ij}'\beta + \frac{v_{ijj}}{2}} \left[W_{ij}' \left(\sum_{g=1}^{m} W_{gs}' \hat{V}_{gss}^{-1} W_{gs} \right)^{-1} W_{ij} + \frac{1}{4} Var(\hat{v}_{ijj}) \right]$$

$$= E(y_{ij} \mid x_{ij}) \left[\hat{a}_{ijj} + \frac{1}{4} Var(\hat{v}_{ijj}) \right]$$

where $\hat{a}_{ijj} = W_{ij}' \hat{V}ar(\hat{\beta}) W_{ij}$ and $\hat{V}ar(\hat{\beta}) = \left(\sum_{i=1}^{m} W_{is}' \hat{V}_{iss}^{-1} W_{is} \right)^{-1}$ is the usual estimator of $Var(\hat{\beta})$. Collecting these expressions we can see

$$E\left[z(\hat{\eta}_{ij}) \right] \approx e^{W_{ij}'\beta + \frac{v_{ijj}}{2}} + 0 + \frac{1}{2} e^{W_{ij}'\beta + \frac{v_{ijj}}{2}} \left[W_{ij}' \left(\sum_{g=1}^{m} W_{gs}' \hat{V}_{gss}^{-1} W_{gs} \right)^{-1} W_{ij} + \frac{1}{4} Var(\hat{v}_{ijj}) \right]$$

$$= e^{W_{ij}'\beta+\frac{v_{ijj}}{2}}\left\{1+\frac{1}{2}\left[W_{ij}'\left(\sum_{g=1}^{m}W_{gs}'\hat{V}_{gss}^{-1}W_{gs}\right)^{-1}W_{ij}+\frac{1}{4}Var(\hat{v}_{ijj})\right]\right\}$$

$$\neq e^{W_{ij}'\beta+\frac{v_{ijj}}{2}} = E\left[z(\eta_{ij})\right]$$

Our fitted values are therefore defined by the second order bias corrected estimator of $E(y_{ij} \mid x_{ij})$ as

$$\hat{y}_{ij} = h(W_{ij};\hat{\eta}_{ij}) = \hat{k}_{ij}^{-1}e^{W_{ij}'\hat{\beta}+\hat{v}_{ijj}/2} \tag{5.13}$$

where $\hat{k}_{ij} = 1+\frac{1}{2}\left\{\hat{a}_{ijj}+\frac{1}{4}\hat{V}(\hat{v}_{ijj})\right\}$ is the bias correction and $\hat{V}ar(\hat{v}_{ijj})$ is the estimated asymptotic variance of $\hat{v}_{ijj}$. Under the ML and REML estimation of the variance components of (5.10), this estimated asymptotic variance can be obtained from the inverse of the relevant information matrix. Note that the bias adjustment described in Karlberg (2000) is a special case of (5.13). Appendix J elaborates the evaluation of $\hat{V}ar(\hat{v}_{ijj})$ under a random slope specification of model (5.9).

In order to use (5.6) and (5.7) to define the model-calibrated sample weights, we also need an expression for the second order moments, under a log scale linear mixed model (5.10), of the population values of *Y* given these fitted values. A first order approximation to these moments is defined by the conditional moments of *Y* given *X* under (5.10). In particular, assuming normality of the random effects vectors u_i and e_i, the covariance between y_{ij} and y_{ik} in small area *i* is

$$Cov(y_{ij}, y_{ik} \mid x_{ij}, x_{ik}) = Cov(e^{W_{ij}'\beta+G_{ij}'u_i+e_{ij}}, e^{W_{ik}'\beta+G_{ik}'u_i+e_{ik}})$$

$$= e^{(W_{ij}+W_{ik})'\beta}\left\{E(e^{G'_{ij}u_i+e_{ij}}e^{G'_{ik}u_i+e_{ik}}) - E(e^{G'_{ij}u_i+e_{ij}})E(e^{G'_{ik}u_i+e_{ik}})\right\}$$

$$= \begin{cases} e^{(W_{ij}+W_{ik})'\beta}[e^{\frac{1}{2}(v_{ijj}+v_{ikk})}(e^{v_{ijk}}-1)] & if \quad j \neq k \\ e^{2W'_{ij}\beta}[e^{v_{ijj}}(e^{v_{ijj}}-1)] & if \quad j = k. \end{cases} \qquad (5.14)$$

The expression (5.14) uses a well known result that for a normal random variable t, $E(e^t) = e^{E(t)+\frac{1}{2}Var(t)}$. See Casella and Berger (1990), page 628.

We therefore define our estimate $\hat{\omega}_{ijk}$ of $Cov(y_{ij}, y_{ik} \mid \hat{y}_{ij}, \hat{y}_{ik})$ by substituting estimates for unknown quantities in (5.14) as

$$\hat{\omega}_{ijk} = \begin{cases} e^{(W_{ij}+W_{ik})'\hat{\beta}}[e^{\frac{1}{2}(\hat{v}_{ijj}+\hat{v}_{ikk})}(e^{\hat{v}_{ijk}}-1)] & if \quad j \neq k \\ e^{2W'_{ij}\hat{\beta}}[e^{\hat{v}_{ijj}}(e^{\hat{v}_{ijj}}-1)] & if \quad j = k. \end{cases} \qquad (5.15)$$

Note that we can then write $\hat{\Omega}_i = [\hat{\omega}_{ijk}] = E_i \Delta_i E_i'$, where $E_i = diag\left\{e^{W'_{ij}\hat{\beta}}; 1 \leq j \leq N_i\right\}$ and $\Delta_i = [\delta_{ijk}]$ is the $N_i \times N_i$ positive definite matrix with $\delta_{ijk} = e^{(\hat{v}_{ijj}+\hat{v}_{ikk})/2}(e^{\hat{v}_{ijk}}-1)$.

Under the random intercept specification of model (5.9), we have:

$Var(u_i) = \sigma_u^2$, $Var(e_{ij}) = \sigma_e^2$ and $V_i = [v_{ijk}] = \sigma_u^2 1_{N_i} 1'_{N_i} + \sigma_e^2 I_{N_i}$ with $v_{ijj} = \sigma_e^2 + \sigma_u^2$, $v_{ijk} = \sigma_u^2$. This leads to $\Omega_i = E_i\left[e^{(\sigma_e^2+\sigma_u^2)}\left\{\exp(\sigma_u^2 1_{N_i} 1'_{N_i} + \sigma_e^2 I_{N_i}) - 1_{N_i} 1'_{N_i}\right\}\right] E_i'$. A similar analytical expression under the random slope specification of model (5.9) is presented in Appendix I.

In order to compute the model-calibrated weights (5.7) under the log-scale linear mixed model (5.10) we finally need to define the design matrix J_U and the estimated covariance matrices $\hat{\Omega}_{ss}$ and $\hat{\Omega}_{sr}$. Under a ratio specification J_U is just the population vector $\hat{Y}_U$ of fitted values (5.13), while under a regression specification $J_U = [1_N \ \hat{Y}_U]$. In both cases, $\hat{\Omega}_{ss} = diag\left\{\hat{\Omega}_{iss}; 1 \le i \le m\right\}$ and $\hat{\Omega}_{sr} = diag\left\{\hat{\Omega}_{isr}; 1 \le i \le m\right\}$, where $\hat{\Omega}_{iss}$ and $\hat{\Omega}_{isr}$ are defined by the sample/non-sample decomposition of $\hat{\Omega}_i$, where $\hat{\Omega}_i$ is defined below (5.15).

5.3.2.2 *Non-Normal Distribution for Random Errors*

The bias corrected predictor (5.13) and the covariance (5.14) are derived assuming normality of log-scale random effects. However, there is no good reason (beyond convenience) to assume that with skewed data these random area effects should be normal. In such cases, random effects with non-normal (non-symmetric) distribution may describe the data well. One alternative, given a scalar area effect in (5.9), is to assume that the random effects in (5.9) are drawn from the *gamma* family of distributions. We consider gamma family of distribution since most of the skewed distributions (e.g. exponential, chi-square etc) are special case of this family. Similar to normal distribution (see section 5.3.2.1) we shall derive first and second moments under gamma distribution of random effects. Before deriving these moments, we recall some common results to be used.

If a random variable t follows a gamma distribution with shape parameter a and rate parameter b (or scale parameter $1/b$), which is denoted by $t \sim Gamma(a,b)$ then their mean, variance and moment generating function are defined as: $E(t) = ab^{-1}$, $V(t) = ab^{-2}$ and $M_t(x) = E(e^{xt}) = (1 - xb^{-1})^{-a}, b > x$ respectively. Further, sum of two independent gamma variables is also a gamma variable. See Casella and Berger (1990).

We first consider two independent gamma distributed random variables as: $u_i^* \sim Gamma(a,b)$ and $e_i^* \sim Gamma(c,d)$ with means $E(u_i^*) = ab^{-1}$ and $E(e_i^*) = cd^{-1}$ and variances $Var(u_i^*) = ab^{-2} = \Sigma$ and $Var(e_i^*) = cd^{-2} = \sigma_e^2$ respectively. Then we define two centred mean gamma distributed random variables $u_i = u_i^* - E(u_i^*)$ and $e_i = e_i^* - E(e_i^*)$ such that

$$E(u_i) = E[u_i^* - E(u_i^*)] = E(u_i^* - ab^{-1}) = 0 \text{ and } Var(u_i) = Var(u_i^* - ab^{-1}) = \Sigma,$$

$$E(e_i) = E[e_i^* - E(e_i^*)] = E(e_i^* - cd^{-1}) = 0 \text{ and } Var(e_i) = Var(e_i^* - cd^{-1}) = \sigma_e^2.$$

That is we defined two random errors: $u_i \sim Gamma(0,\Sigma)$ and $e_i \sim Gamma(0,\sigma_e^2)$. These two random errors are independent. There is no loss of generality in taking $E(u_i) = 0 = E(e_i)$ that is in making adjustment for zero mean (McCulloch and Searle, 2001, page 157) in defining the model (5.9). Let us consider model (5.9) assuming that two random effects u_i and e_i follow gamma distribution. In particular, we consider the random intercept specification of model (5.9).

From the properties of gamma distribution and using binomial and exponential expansions (ignoring higher order terms) we then have

$$E(y_{ij} \mid x_{ij}) = E\left(e^{W_{ij}'\beta + u_i + e_{ij}}\right) = e^{W_{ij}'\beta}\left\{E\left(e^{u_i^* - \frac{a}{\lambda}}\right)E\left(e^{e_{ij}^* - \frac{c}{d}}\right)\right\}$$

$$= e^{W_{ij}'\beta} e^{-\frac{a}{b}}\, e^{-\frac{c}{d}}\left\{E(e^{u_i^*})E(e^{e_{ij}^*})\right\} = e^{W_{ij}'\beta} e^{-\frac{a}{b}}\, e^{-\frac{c}{d}}(1-b^{-1})^{-a}(1-d^{-1})^{-c}$$

$$\approx e^{W_{ij}'\beta}\left(1-\frac{a}{b}+\frac{1}{2!}\frac{a^2}{b^2}\right)\left(1+\frac{a}{b}+\frac{1}{2!}\frac{a}{b^2}+\frac{1}{2!}\frac{a^2}{b^2}\right)\times$$

$$\times\left(1-\frac{c}{d}+\frac{1}{2!}\frac{c^2}{d^2}\right)\left(1+\frac{c}{d}+\frac{1}{2!}\frac{c}{d^2}+\frac{1}{2!}\frac{c^2}{d^2}\right)$$

$$\approx e^{W_{ij}'\beta}\left(1+\frac{1}{2}\frac{a}{b^2}\right)\left(1+\frac{1}{2}\frac{c}{d^2}\right) \approx e^{W_{ij}'\beta} e^{\frac{a}{2b^2}} e^{\frac{c}{2d^2}}$$

$$= e^{W_{ij}'\beta} e^{\frac{1}{2}\left(\frac{a}{b^2}+\frac{c}{d^2}\right)} = e^{W_{ij}'\beta} e^{\frac{1}{2}Var(l_{ij})} \approx e^{W_{ij}'\beta + v_{ijj}/2} = z(\eta_{ij}).$$

This expression is identical to the (5.13) derived under the normal distribution of the random effects. Therefore rest of derivation follows from (5.13). This indicates that MBD estimators based on the normal theory fitted value model defined by (5.13) and (5.14) can be expected to possess some robustness with respect to the distribution of the random effects in (5.9).

5.4 Small Area Estimation under Model-Calibration

Given an appropriate design matrix J_U defined by the fitted values (5.13) and estimated covariance matrices $\hat{\Omega}_{ss}$ and $\hat{\Omega}_{sr}$ defined by (5.15), we can compute a set of model-calibrated weights (5.7). These weights depend on the random area effects in the log-scale linear mixed model (5.9) and are thus suited to SAE. Here we use them to define

MBD estimators for small area means (see chapter 3). In particular, we consider two forms of the MBD estimator for a small area mean. The first is the Hájek form of the MBD estimator, defined as a weighted mean of the sample data from the small area of interest. Given a set of weights $w_s = (w'_{s1}\ w'_{s2} \ldots w'_{sm})'$, where $w_{si} = (w_{ij}; j \in s_i)$ are the model-calibrated weights for the n_i units making up the sample s_i from small area i, this estimator is

$$\hat{\bar{Y}}_i^{Hajek} = \sum_{j \in s_i} w_{ij} y_{ij} \Big/ \sum_{j \in s_i} w_{ij} . \tag{5.16}$$

An alternative MBD estimator when the population size N_i of the small area is known is the Horvitz-Thompson form

$$\hat{\bar{Y}}_i^{HT} = N_i^{-1} \sum_{j \in s_i} w_{ij} y_{ij} . \tag{5.17}$$

In chapter 3 we only considered the Hájek form of the MBD estimator for small areas using the sample weights (5.8) derived via a linear mixed model (3.11). However, the sample weights (5.7) are derived via model calibration where estimator is defined as the HT form (see section 5.2). Therefore, we consider both forms of the MBD estimators.

Both estimators (5.16) and (5.17) depend on how the 'fitted value' model (5.6) underpinning the model calibration weights (5.7) is specified. In particular, we consider two different specifications for the fitted value model (5.6) that is two types of specification for J_U, the ratio and regression specifications for this model (see below equation 5.6). This leads to four different MBD estimators set out in Table 5.1. Note that all four use the same predicted values (5.13) and the same estimated covariance structure (5.15).

Table 5.1 Different MBD estimator configurations

Estimator	Estimator type	Model specification
TrMBD1	Hájek type	Ratio specification
TrMBD2	Horvitz-Thompson type	Ratio specification
TrMBD3	Hájek type	Regression specification
TrMBD4	Horvitz-Thompson type	Regression specification

Estimation of mean squared error of (5.16) and (5.17) follows the approach described in section 3.3.2 of chapter 3, which treats these estimators as simple weighted estimators of a domain mean. Under this approach the sample weights derived from (5.7) are considered as fixed and the prediction variance of (5.16) and (5.17) is estimated using a standard heteroskedasticity robust variance estimator that only assumes the first order moments defined by (5.6). See Royall and Cumberland (1978). A "plug-in" estimate of the squared bias of (5.16) and (5.17) under (5.6) is added to this estimated prediction variance to finally define a simple estimate of the mean squared error of these estimators. Under this approach the sample weights underlying (5.16) and (5.17) "borrow strength" via the log-scale linear mixed model (5.9), but this model is not used in inference. In particular, since the mean squared error estimators for small area means only assume the first order moments specified by (5.6), we ensure consistency with the way mean squared errors are estimated at population level. See Chandra and Chambers (2005). In particular, the mean squared error of the weighted estimator $\hat{\bar{Y}}_i^w$ ($\hat{\bar{Y}}_i^{Hajek}$ or $\hat{\bar{Y}}_i^{HT}$) for the population mean of Y in small area i, $\bar{Y}_i$ is

$$MSE(\hat{\bar{Y}}_i^w) = Var(\hat{\bar{Y}}_i^w - \bar{Y}_i) + B^2(\hat{\bar{Y}}_i^w) \tag{5.18}$$

where $Var(\hat{\bar{Y}}_i^w - \bar{Y}_i) = N_i^{-2}\left\{\sum_{j\in s_i} a_j^2 Var(y_j) + \sum_{j\in r_i} Var(y_j)\right\}$ is the prediction variance of weighted estimators (5.16) and (5.17) with

$$a_j = \begin{cases} (N_i w_j - \sum_{g\in s_i} w_g)/(\sum_{g\in s_i} w_g) & for\ Hájek\ form \\ w_j - 1 & for\ HT\ form \end{cases}$$

and $B(\hat{\bar{Y}}_i^w) = E(\bar{h}_{iw}) - \bar{h}_i$ is the bias of (5.16) and (5.17). Here $\bar{h}_i$ and $\bar{h}_{iw}$ denotes the population mean and weighted average of the fitted values $\hat{h}_{ij} = h(W_{ij};\hat{\eta}_{ij})$ in area i respectively.

A robust estimator of the mean squared error of (5.16) and (5.17) is

$$mse(\hat{\bar{Y}}_i^w) = v(\hat{\bar{Y}}_i^w) + \hat{B}^2(\hat{\bar{Y}}_i^w) \tag{5.19}$$

where $v(\hat{\bar{Y}}_i^w) = \sum_{j\in s_i} c_j (y_{ij} - \hat{h}_{ij})^2$, with $c_j = N_i^{-2}\left\{a_j^2 + (N_i - n_i)(n_i - 1)^{-1}\right\}$, and

$\hat{B}(\hat{\bar{Y}}_i^w) = \hat{\bar{h}}_{iw} - \hat{\bar{h}}_i$, with $\hat{\bar{h}}_i = N_i^{-1}\sum_{j=1}^{N_i} \hat{h}_{ij}$, and

$$\hat{\bar{h}}_{iw} = \begin{cases} (\sum_{j\in s_i} w_{ij}\hat{h}_{ij})/(\sum_{j\in s_i} w_{ij}) & for\ Hájek\ form \\ \sum_{j\in s_i} (w_{ij}\hat{h}_{ij})/N_i & for\ HT\ form \end{cases}.$$

Besides these four MBD estimators (TrMBD1-TrMBD4, Table 5.1) defined by (5.16) and (5.17), we also define an Empirical Best Predictor (EBP) for the mean of Y for small area i (denoted by TrEBP) under the 'fitted value' model define by (5.13) as

$$\begin{aligned} \hat{\bar{Y}}_i^{EBP} &= N_i^{-1}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}\right\} \\ &= N_i^{-1}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \left(\hat{k}_{ij}^{-1} e^{W_{ij}'\hat{\beta} + \hat{v}_{ij}/2}\right)\right\} \end{aligned} \tag{5.20}$$

where $\hat{k}_{ij}; j \in s_i$ is define below equation (5.13). Unlike MBD estimators, the MSE estimation of the EBP (5.20) is not straightforward. We do not pursue the MSE estimation of (5.20). See Appendix L.

5.5 Conclusions

In this chapter we developed the SAE techniques for skewed data when standard methods for the SAE based on linearity assumption are inappropriate. In particular, we derived the SAE methods for the survey variables which are linear on log-log scale. We defined the MBD estimators for small area means based on normality assumption of random errors. However, for skewed data random effects are not always normal and the estimation procedure based on non-normal random effects seem suitable. We also consider the *gamma* distribution for random effects. Our results show method is robust with respect to distribution of random effects.

In this chapter we proposed four different types of MBD estimators for small means for skewed data and their mean squared error estimate. However, it remains to evaluate the empirical performance of these estimators. In chapter 6 we shall examine the performance of these methods using a Monte Carlo simulation study and application to real population data. We also study an empirical best predictor (EPB) for small means under the 'fitted value' model. Some empirical results related the EBP (5.20) are presented in Appendix L.

CHAPTER 6

MONTE CARLO EVALUATIONS

6.1 Introduction

In chapter 5 we proposed the small area estimation (SAE) techniques for skewed data. However, it remains to assess the performance these methods of SAE. In this chapter we evaluate these techniques of SAE by designing a series of Monte Carlo simulation experiments. The use of simulation techniques in statistics has its origins in the beginning of the 20^{th} century (Morgan, 1984). Lewis and Orav (1989) define simulation as a controlled statistical procedure (experiment) based on repeated sampling carried out on a computer. We present in this chapter the characteristics and results of simulation studies, which has the main objective of evaluating the comparative performance of different methods of SAE.

In the next section we introduce the different estimators investigated in the simulation studies. In section 6.3 and 6.4 we provide illustrative information on how these simulation studies is implemented, associated population and criterion used to assess the performance of different estimators. Section 6.5 is devoted to reporting the results and their explanations. Finally, section 6.6 presents summary of the major findings from the empirical studies and a discussion of some outstanding issues.

6.2 Description of Simulation Studies

In this section we illustrate different methods of SAE considered and the population data used in the empirical studies.

6.2.1 Estimators Investigated in Simulation Studies

In our empirical evaluations, we investigate the comparative performance of the seven different estimators for SAE. These are

I. The proposed model-based direct estimators for skewed data based on the model-calibrated EBLUP weights for skewed data calculated via (5.7) under a fitted value model derived from the log-scale linear mixed model (5.9) (section 5.4, Table 5.1)
    1. Hájek type estimator under ratio specification: TrMBD1
    2. Horvitz-Thompson (HT) type estimator under ratio specification: TrMBD2
    3. Hájek type estimator under regression specification: TrMBD3
    4. HT type estimator under regression specification: TrMBD4

II. The MBD estimators based on the sample weights (5.8) derived under 'standard' raw-scale linear mixed model (3.11) (section 3.3.2)
    5. Hájek type estimator: MBD1
    6. HT type estimator: MBD2

III. The EBLUP derived under the same raw-scale linear mixed model as that used to calculate the weight (5.8) (Prasad and Rao, 1990)

7. EBLUP

We note that first six estimators are the model-based direct (MBD) estimator defined as weighted linear estimator of the form either given by $\hat{\bar{Y}}_i^w = (\sum_{j\in s_i} w_j y_j)/(\sum_{j\in s_i} w_j)$ for Hájek type estimator, or $\hat{\bar{Y}}_i^w = (\sum_{j\in s_i} w_j y_j)/N_i$ for HT type estimator. In the first four estimators, the sample weights used (corresponding to small areas) to define the estimators for small areas are derived under the population version of 'expected value' model via 'model calibration' approach. In the next two estimators, the sample weights are derived from a population version of raw-scale linear mixed model, referred as the sample weights via 'standard calibration' approach. The seventh estimator is the standard EBLUP, an indirect estimator under the raw-scale linear mixed model. Besides these seven estimators we also examine the performance of an empirical best predictor (EBP) for small areas (5.20) under a log scale linear mixed model (5.9), denoted by TrEBP. See chapter 5. We do not pursue this estimator in details. Appendix L presents some of the empirical results related to the TrEBP method of SAE.

The mean squared errors for the MBD estimators (that is for first six estimators) are estimated using the method described in chapter 3 and 5, while the mean squared error of the EBLUP is estimated using the method described in Prasad and Rao (1990), discussed in chapter 3.

6.2.2 Types of Simulation Studies

We consider two types of simulation studies. The first type of study uses the model-based simulation to generate artificial population and sample data. These data are then used to compare the performances of the different estimators. We carry out three sets of model-based simulations, labelled by sets A, B and C respectively. In the first set of simulations (denoted by Set-A), we investigate the performance of these estimators given population data generated using the log-scale linear mixed model (5.9). In second set of simulations (denote by Set-B), we examine the robustness of these estimators to misspecification of this model. In simulation Set A and B we assume that the random effects have normal distribution. In the third set of the simulation (denoted by Set-C) we study the performance of these estimators given population data generated under the same log-scale linear mixed model (5.9) identical to set-A except that random effects have non-normal distribution. We consider a *gamma* distribution for these random effects. The second type of simulation study is the design-based. Here we evaluate the empirical performance of these estimators in the context of repeated sampling from a real population using realistic sampling methods.

6.3 The Model Based Simulation Study

In this section we describe the model-based simulations to contrast the performance of different estimators used for SAE with skewed population data.

6.3.1 Simulated Data

In our model-based simulations we set a population size of $N = 15,000$ with $m = 30$ small areas and randomly generated the small area population sizes N_i, $i = 1,...,30$ from a chi-square distribution with 750 degree of freedom so that $\sum_i N_i = N$. We used an overall sample size of $n = 600$ with small area sample sizes set so that they were proportional to the corresponding small area population sizes. That is $n_i = N_i(n/N)$ so that $\sum_i n_i = n$. The average small area population and sample sizes are 500 and 20 respectively. These area-specific sample sizes were kept fixed in all our simulations (Set-A, B and C).

6.3.1.1 *Simulation Set-A*

In Set A of our model-based simulations the population values y_{ij} are generated using the multiplicative model $y_{ij} = 5.0 x_{ij}^{\beta} u_i e_{ij}$, with random samples then taken from each small area. The generated population is skewed on raw scale and linear on log-scale. We used six different values of parameter β (0.5, 0.8, 1.0, 1.3, 1.5 and 2.0). These are denoted by ParA1 to ParA6. Here the values of covariate x_{ij} are independently drawn from the log-normal distribution LN(6, σ_x), while the individual effects e_{ij} and the area effects u_i are independently drawn from the LN(0, σ_e) and LN(0, σ_u) distributions respectively. The values of σ_e and σ_u are chosen so that the intra-area correlation $\left(Rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)\right)$

in the population varied between 0.20 and 0.25. Table 6.1.a sets out the six different sets of parameter values that are used in the simulation Set A. These ensured that the simulated populations contained a wide range of variation.

Using the sample data in each case, parameter values are estimated using the *lme* function in R (Bates and Pinheiro, 1998), and estimates for the small area means then calculated, along with appropriate nominal 95% confidence intervals. The process of generating population and sample data, estimation of parameters and calculation of small area estimates are independently replicated 1000 times. The results from this part of the simulation study are set out in Table 6.2.

Table 6.1.a Parameters of the simulation set-A.

Parameter	β	σ_u	σ_e	σ_x
ParA1	0.5	0.30	0.50	3.00
ParA2	0.8	0.35	0.60	2.50
ParA3	1.0	0.40	0.70	2.25
ParA4	1.3	0.45	0.80	1.75
ParA5	1.5	0.50	0.90	1.50
ParA6	2.0	0.60	1.00	1.20

6.3.1.2 *Simulation Set-B*

In Set B of the model-based simulations, population data are generated using the model $y_{ij} = 5.0 x_{ij} [\exp(\log^2(x_{ij}))]^{\gamma} u_i e_{ij}$. The generated population is non-linear on raw scale and quadratic on log-scale. Here the individual effects e_{ij} and the area effects u_i are independently drawn from the LN (0, 1.0) and LN (0, 0.5) distributions respectively,

while the covariate values x_{ij} are drawn from a LN (3, 0.2) distribution. Five different values for the parameter γ (-1.0, -0.5, 0.0, 0.5 and 1.0) are investigated, thus generating population data with different degrees of curvature. These parameter sets are denoted by ParB1-ParB5. Table 6.1.b shows different parameters of this simulated population. All other aspects of these simulations, including the estimators considered, are the same as in Set A. Table 6.3 presents results from this component of the simulation study.

Table 6.1.b Parameters of the simulation set-B.

Set	γ	σ_u	σ_e	σ_x
ParB1	-1.0	0.5	1.0	0.2
ParB2	-0.5	0.5	1.0	0.2
ParB3	0.0	0.5	1.0	0.2
ParB4	0.5	0.5	1.0	0.2
ParB5	1.0	0.5	1.0	0.2

6.3.1.3 *Simulation Set-C*

In Set C of the model based simulations, the model $y_{ij} = \exp\{\alpha + \beta \log x_{ij} + u_i + e_{ij}\}$ is used to generate the population data. This population data is skewed on raw scale and linear on log-scale. Here random effects are generated from gamma distribution. We fixed $\alpha = 5.0$ and chosen six different values of the parameter β (0.5, 0.8, 1.0, 1.3, 1.5 and 2.0) which corresponds to six different parameter sets denoted by ParC1 to ParC6, shown in Table 6.1.c. We first generate independent random errors e_{ij}^* from a gamma distribution with shape parameter a and rate parameter b (scale parameter $1/b$), that is $e_{ij}^* \sim Gamma(a,b)$ with mean ab^{-1} and variance ab^{-2} and then get the independent

random effects $e_{ij} = \left(e_{ij}^{*} - E(e_{ij}^{*})\right) = \left(e_{ij}^{*} - ab^{-1}\right)$ with mean zero and variance $\sigma_e^2 = ab^{-2}$. Similarly we first generated random errors $u_i^{*} \sim Gamma(c,d)$ then get the random area effects $u_i = \left(u_i^{*} - cd^{-1}\right)$ with mean zero and variance $\sigma_u^2 = cd^{-2}$. The covariate values x_{ij} are generated from LN $(6, \sigma_x)$ distribution. The values of parameter $\sigma_e^2 = ab^{-2}$ and $\sigma_u^2 = cd^{-2}$ are fixed up so that intra-area correlation varies between 0.20-0.25. The rest of the process is identical to the Set-A. The results from this set of the simulation study are presented in Table 6.4.

Table 6.1.c Parameters of the simulation set-C.

Parameter	α	β	σ_u	σ_e	σ_x
ParC1	5.0	0.5	0.30	0.50	3.00
ParC2	5.0	0.8	0.35	0.60	2.50
ParC3	5.0	1.0	0.40	0.70	2.25
ParC4	5.0	1.3	0.45	0.80	1.75
ParC5	5.0	1.5	0.50	0.90	1.50
ParC6	5.0	2.0	0.60	1.00	1.20

6.3.2 Performance Indicators

We use following measures to assess the performance of different estimators for SAE:

- *The percentage relative bias*, defined as

$$RB(\hat{T}_i) = \left(R^{-1}\sum_{r=1}^{R} T_{i(r)}\right)^{-1} \left\{\left(R^{-1}\sum_{r=1}^{R} \hat{T}_{i(r)}\right) - \left(R^{-1}\sum_{r=1}^{R} T_{i(r)}\right)\right\} \times 100 \qquad (6.1)$$

where $\hat{T}_i$ is the estimator (e.g. for the mean or total) for the i^{th} $(i=1,...,m)$ small area for parameter T_i and $\hat{T}_{i(r)}$ is the specific outcome of $\hat{T}_i$ obtained in the simulation r $(r=1,.....,R=1000)$.

- *The average percentage relative bias* (averaged over m small areas), defined as

$$ARB = m^{-1}\sum_{i=1}^{m} RB(\hat{T}_i) \tag{6.2}$$

- *The percentage relative root mean squared error*, defined as

$$RRMSE(\hat{T}_i) = \left(R^{-1}\sum_{r=1}^{R} T_{i(r)}\right)^{-1}\left\{\sqrt{R^{-1}\sum_{r=1}^{R}\left(\hat{T}_{i(r)} - T_{i(r)}\right)^2}\right\}\times 100 \tag{6.3}$$

- *The average percentage relative root mean squared error* (averaged over m small areas), defined as

$$ARRMSE = m^{-1}\sum_{i=1}^{m} RRMSE(\hat{T}_i) \tag{6.4}$$

- *The coverage rate*, defined as

$$CR(\hat{T}_i) = R^{-1}\sum_{r=1}^{R} 1\left\{T_i \in \left(\hat{T}_{i(r)} \pm 2\sqrt{mse(\hat{T}_{i(r)})}\right)\right\}. \tag{6.5}$$

Here $mse(\hat{T}_{i(r)})$ is the estimate of the MSE of $\hat{T}_{i(r)}$.

- *The average coverage rate* (averaged over m small areas), defined as

$$ACR = m^{-1}\sum_{i=1}^{m} CR(\hat{T}_i) \tag{6.6}$$

- *The 2-sigma confidence interval width*, defined as

$$wd(\hat{T}_i) = R^{-1}\sum_{r=1}^{R}\left\{4\sqrt{mse(\hat{T}_{i(r)})}\right\} \tag{6.7}$$

- *The average 2-sigma confidence interval width* (averaged over m areas), defined as

$$Awd = m^{-1}\sum_{i=1}^{m} wd(\hat{T}_i) \tag{6.8}$$

This section is similar to section 3.4.2 of chapter 3 where we have already described these performance criterions. However, section 3.4.2 defines various performance indicators in context of design-based simulations where population is fixed. In contrast, this section defines these criterions for the model-based simulations where population is not fixed and changes over the simulations (i.e. population is random over the simulation and drawn under the model). Further, the equations defining the averages over small areas are same as in section 3.4.2 but these are repeated just to bring continuity.

6.4 The Design Based Simulation Study

In this section we describe the design-based simulations to test the different methods of SAE using real data. That is an application of the proposed SAE methods to real population data.

6.4.1 Simulated Data

In design-based simulations, our basic data come from the same sample of 1652 Australian broadacre farms from the Australian Agricultural and Grazing Industries Survey (AAGIS) data that were used for the empirical evaluations reported in chapter 3 and 4 and also used in simulation study reported in Chambers and Chandra (2006) and Chandra and Chambers (2005). In particular, we use the same target population of 81982 farms (obtained by sampling with replacement from the original sample of 1652 farms

with probabilities proportional to their sample weights). The same 1000 independent stratified random samples as in chapter 3 were then drawn from this (fixed) population, with total sample size in each draw equal to the original sample size (1652) and with the small areas of interest defined by the 29 Australian agricultural regions represented in this population. Sample sizes within these regions were fixed to be the same as in the original sample (varied from a low of 6 to a high of 117). Various characteristics of this simulated population are described in Table 3.1 in chapter 3. The aim is to estimate average annual farm costs (TCC, measured in A$) in each region using farm size (hectares) as the auxiliary variable. The same mixed model specification as in chapter 3 and Chandra and Chambers (2005) is used. This includes an interaction term (zone by size) in the fixed effects and a random slope specification for the area effect. In its linear form the model does not fit the AAGIS sample data terribly well. This fit is improved (albeit marginally) when a log-scale linear specification is used. Our results are summarized in Table 6.5.

6.4.2 Performance Indicators

To evaluate the comparative performance of different estimators in design based simulation studies we use the criteria of percentage relative bias, percentage relative root mean squared error and coverage rate defined in section 3.4.2 in chapter 3 for design based simulations.

6.5 Results of the Simulation Studies

6.5.1 Model Based Simulations

Table 6.2 sets out the average relative biases (%), average relative root mean squared errors (%), average coverage rates and average width of 2-sigma confidence intervals generated by different estimators for the Set-A.

The most striking feature of Table 6.2 is the extremely large values of the average relative bias of the Hájek-type estimators (TrMBD1 and TrMBD3) under model-calibrated weighting. In contrast, the HT-type MBD estimators based on model-calibrated weights (TrMBD2 and TrMBD4) are almost identical in their performance, which improves markedly on that of the Hájek type estimators. An investigation of the reason for this anomaly revealed that summing the model-calibrated EBLUP weights (5.7) within small areas produced extremely variable estimates of the small area population sizes, implying that these weights cannot be considered as 'multipurpose' – they function well when used with variables that are reasonably correlated with the variable that defines the fitted value model, but can fail with other, less well correlated, variables (e.g. the indicator variable for small area inclusion). We further note that this problem does not arise with the 'standard' EBLUP weights (5.8), as the Hájek type (MBD1) and HT type (MBD2) MBD estimators derived under a raw-scale linear mixed model are very close in their performances across all six of the scenarios explored in Table 6.2. From now on we therefore focus our discussion on the three estimators, TrMBD2, MBD1 and EBLUP.

Table 6.2 shows that the average relative biases and the average relative RMSEs for TrMBD2 are consistently lower than those generated by MBD1 and EBLUP. Furthermore, average coverage rates and interval widths for TrMBD2 are better than those generated by MBD1 and EBLUP. In comparison, for same order of RB, the RRMSE of EBLUP is smaller than that of MBD1, and, although both estimators generate very similar coverage rates, confidence intervals generated via EBLUP tend to have smaller average widths than those generated via MBD1. The plots in Figure 6.1 and 6.2 display the region-specific performance measures generated by these three estimators (TrMBD2, MBD1 and EBLUP) for the Set A simulations. These show that the RB and the RRMSE values generated by TrMBD2 are smaller than corresponding values for MBD1 and EBLUP in all regions (Figure 6.1). Further, the RB and the RRMSE of MBD1 and EBLUP increase as the non-linearity in the data increases (ParA1 to ParA6). We also see that TrMBD2 generates better coverage rates across all regions compared with the coverage rates generated by EBLUP and MBD1(Figure 6.2).

Overall, these results show that when the model for the underlying population is non-linear there can be significant gains from the use of HT-type MBD estimators for small area means (TrMBD2) based on the model-calibrated weights (5.7) compared with standard linear mixed model-based estimators like MBD1 and EBLUP. They also show that EBLUP performs relatively better than MBD1 in these situations.

Table 6.2 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%), average (ACR) coverage rate and average (AW) 2-sigma confidence interval width for simulation set-A.

Criterion	Estimator	ParA1	ParA2	ParA3	ParA4	ParA5	ParA6
ARB (%)	TrMBD1	-86.02	-96.54	-98.43	-98.58	-98.45	-99.06
	TrMBD2	-0.01	-0.05	0.27	0.09	-0.43	0.76
	TrMBD3	-75.2	-95.97	-97.97	-98.55	-98.12	-98.66
	TrMBD4	0.02	-0.07	0.28	0.11	-0.39	0.75
	MBD1	10.98	4.11	-0.29	-6.28	-7.81	-9.59
	MBD2	12.63	5.47	0.48	-5.91	-7.58	-9.5
	EBLUP	12.65	5.44	0.49	-5.85	-7.68	-9.32
ARRMSE (%)	TrMBD1	0.92	1.13	1.2	1.29	1.43	1.56
	TrMBD2	0.15	0.29	0.39	0.52	0.7	0.88
	TrMBD3	7.98	1.25	1.22	1.3	1.44	1.59
	TrMBD4	0.15	0.29	0.39	0.52	0.7	0.88
	MBD1	1.03	1.47	1.79	1.89	1.98	2.78
	MBD2	1.16	1.6	1.83	1.91	1.99	2.79
	EBLUP	0.76	0.69	0.61	0.75	0.98	1.29
ACR	TrMBD1	0.99	0.98	0.96	0.95	0.94	0.92
	TrMBD2	0.94	0.91	0.89	0.89	0.89	0.89
	TrMBD3	0.99	0.98	0.96	0.95	0.94	0.92
	TrMBD4	0.94	0.91	0.89	0.89	0.89	0.89
	MBD1	0.87	0.85	0.85	0.87	0.88	0.87
	MBD2	0.87	0.85	0.85	0.87	0.88	0.87
	EBLUP	0.85	0.85	0.85	0.87	0.87	0.87
AW	TrMBD1	1265	22389	140563	27×10^4	35×10^5	44×10^6
	TrMBD2	208	4326	33228	7.0×10^4	11×10^5	15×10^6
	TrMBD3	1753	22487	141001	27×10^4	35×10^5	43×10^6
	TrMBD4	220	4426	33722	8.0×10^4	11×10^5	16×10^6
	MBD1	1007	19318	139346	28×10^4	38×10^5	56×10^6
	MBD2	1033	19677	140626	28×10^4	38×10^5	56×10^6
	EBLUP	380	7253	55498	13×10^4	20×10^5	31×10^6

Figure 6.1 Region-specific percentage relative biases and percentage relative RMSEs for simulation set-A.

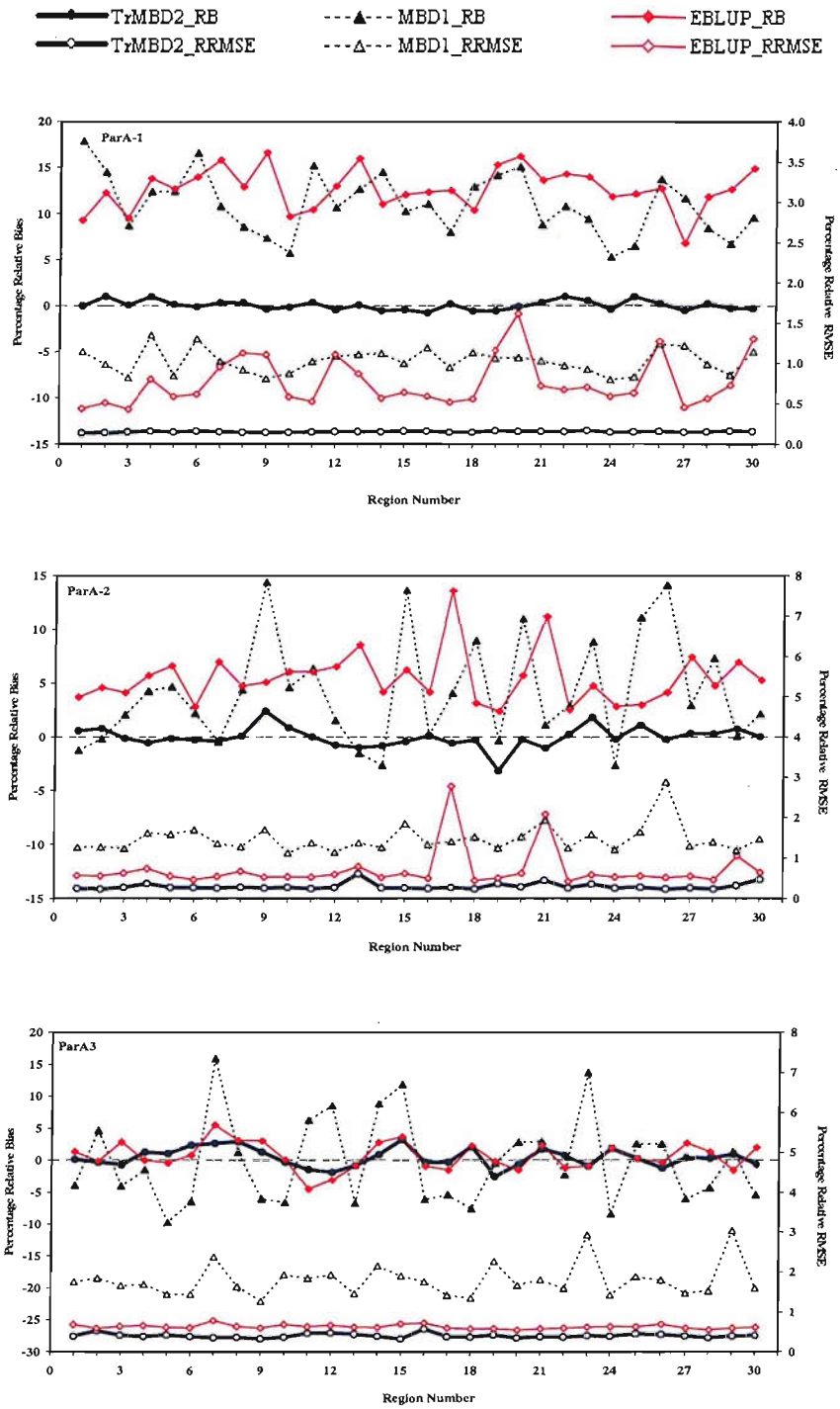


Figure 6.1 (Continued) Region-specific percentage relative biases and percentage relative RMSEs for simulation set-A.

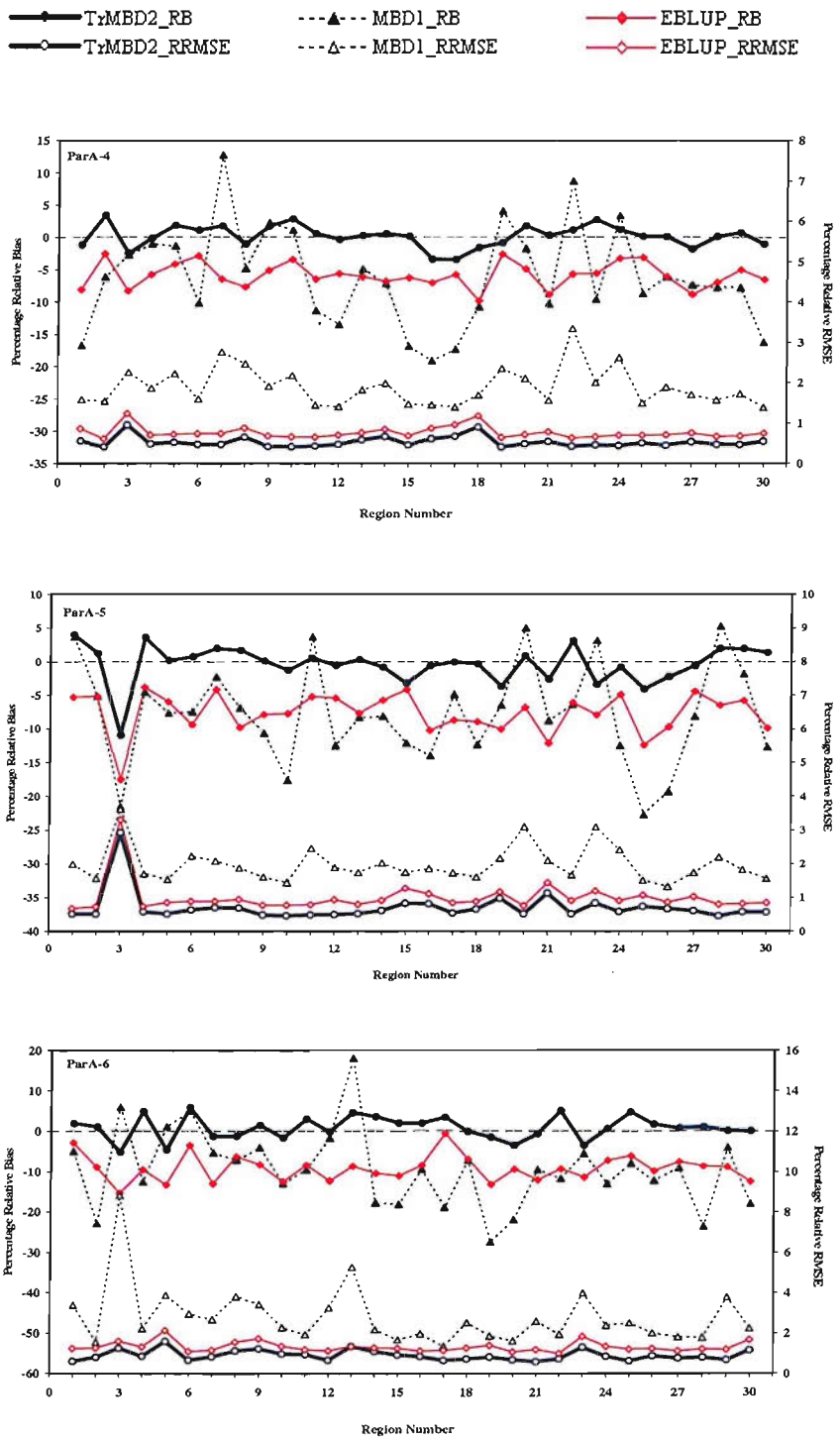


Figure 6.2 Region-specific coverage rates and confidence interval widths for simulation set-A.

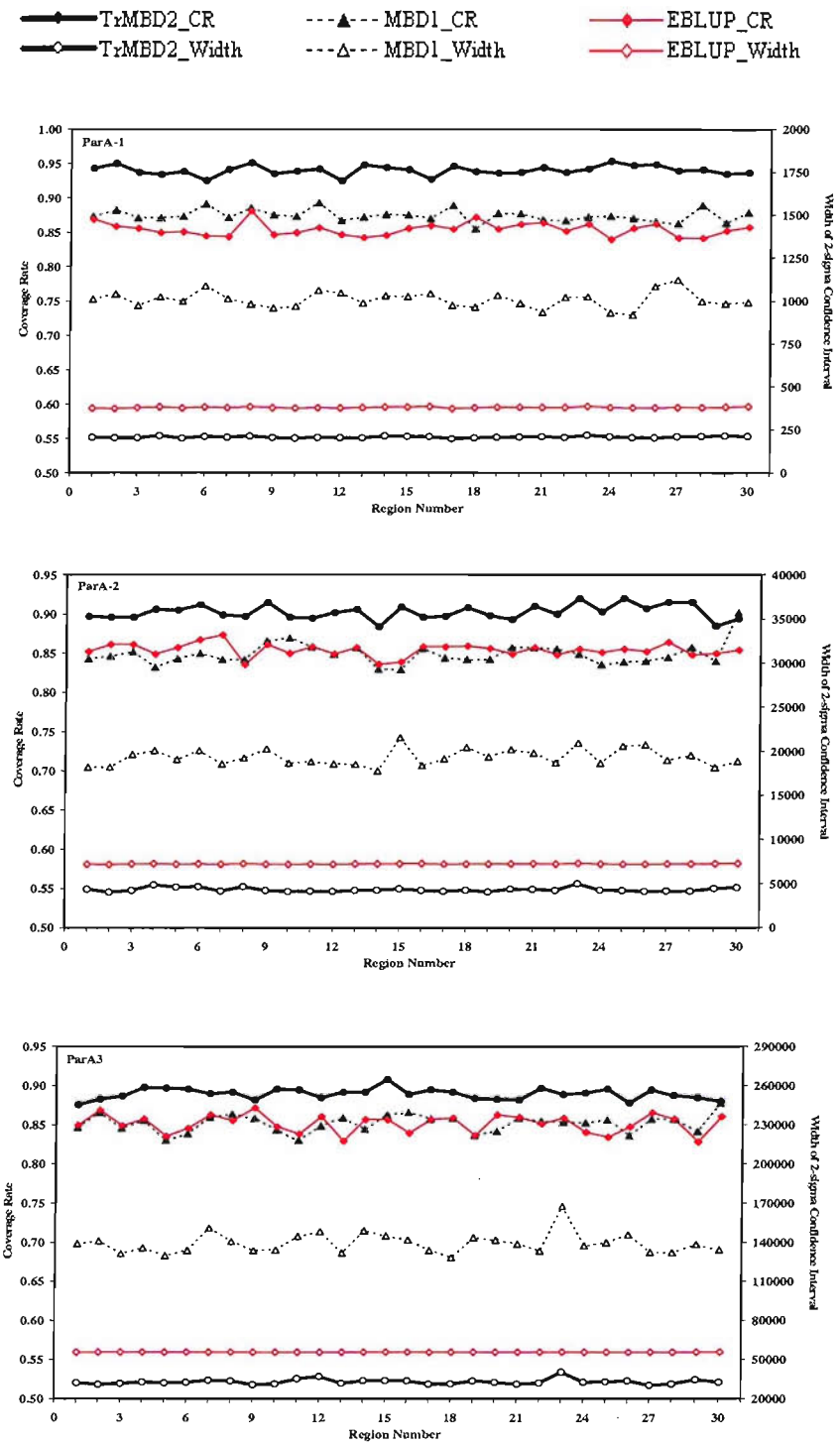
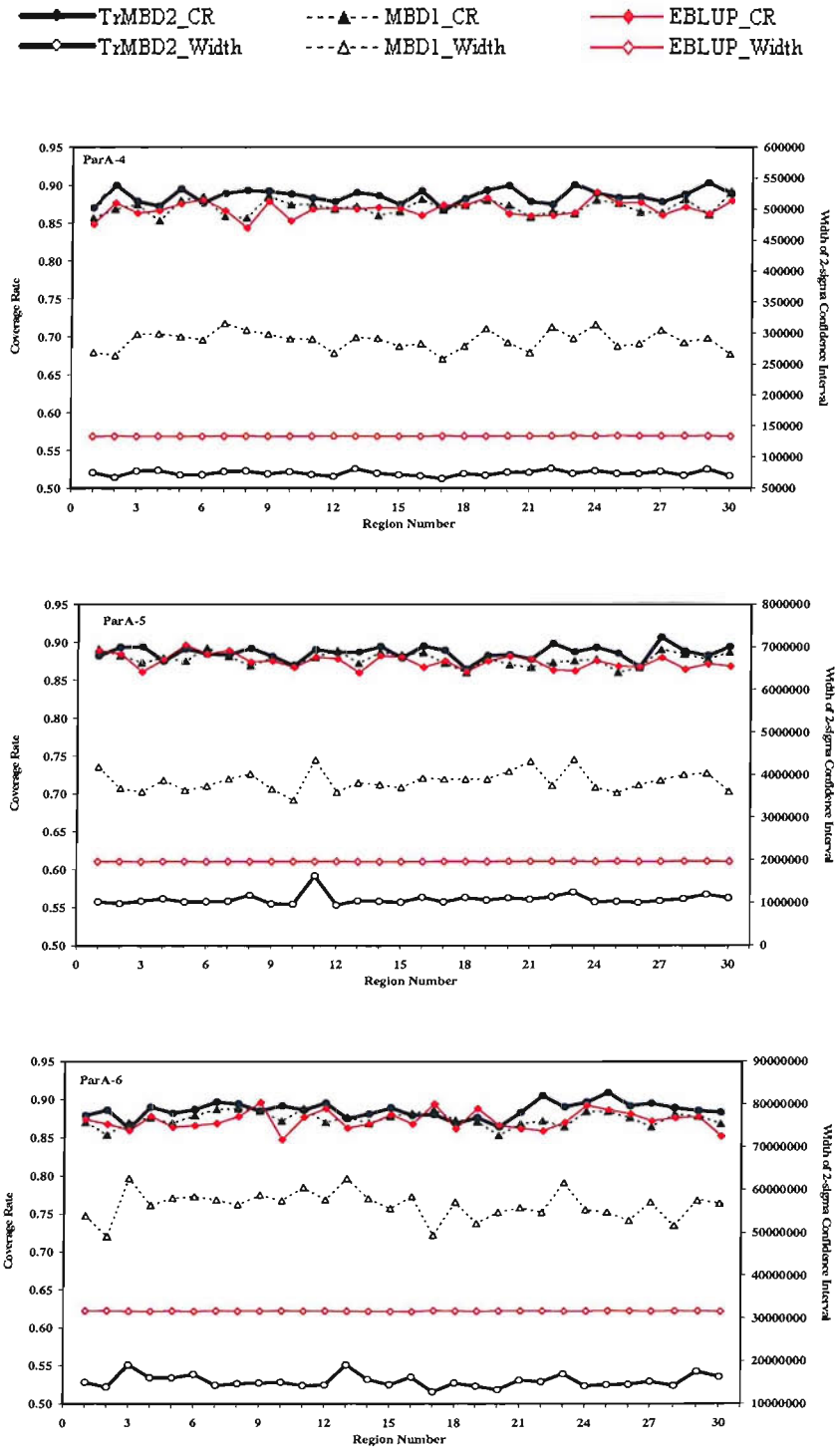


Figure 6.2 (Continued) Region-specific coverage rates and confidence interval widths for simulation set-A.



The MBD (MBD1 and MBD2) estimators and the EBLUP are based on raw-scale linear mixed model, while the MBD estimators (TrMBD1-TrMBD4) derived using model-calibrated EBLUP weights for skewed (5.7) is based on log-scale linear mixed model. These results show under linearity on log-scale, the proposed method for skewed data leads to efficient sets of small area estimate. In Set B of the model-based simulations we investigate the robustness of model-calibrated MBD estimation (TrMBD1-TrMBD4) to misspecification of the non-linear model. The results from Set-B correspond to population data that are non-linear both on the raw and log transform scale. Table 6.3 shows average relative biases (%), average relative root mean squared errors (%), average coverage rate and average 2-sigma confidence interval width for simulation Set-B.

The results in Table 6.3 show that in this case the biases generated by TrMBD2 increase as the actual non-linear model deviates more from the assumed non-linear model ($\gamma = 0.0$ in the Table). However, these biases are offset by small variability, so in terms of average RRMSE TrMBD2 still performs as well or better than EBLUP and continues to dominate MBD1. The biases generated by MBD1 and EBLUP are of the same order, while the average RRMSE of EBLUP dominates that of MBD1. Average coverage rates for EBLUP are marginally better than those of MBD1 and TrMBD2, but the average widths of the confidence intervals underpinning these rates tended to be smallest for TrMBD2, followed by EBLUP and then MBD1.

Figure 6.3 and 6.4 summarize the region-specific performance measures generated by three methods (TrMBD2, MBD1 and EBLUP) for Set-B. Figure 6.3 shows that relative

biases of TrMBD2 are larger than MBD1 and EBLUP for parameter set ParB1 and ParB5 (i.e. γ= -1 and +1 respectively). However, it is nearly same for all methods when values of γ (±0.5, i.e. near to zero) are small. The relative RMSEs of TrMBD2 are lower than both MBD1 and EBLUP in most of the areas for all parameter sets except ParB2 and ParB3, where EBLUP is marginally better. Figure 6.4 demonstrates that although coverage rates of TrMBD2 are marginally lower for ParB2-ParB5, widths of the confidence intervals are consistently smaller for all parameter choices (ParB1- ParB5). Our model-based simulation results for Set B indicate that although MBD-based SAE with model-calibrated weights is susceptible to model misspecification bias, the overall performance of this approach appears relatively unaffected by slight deviations from the assumed non-linear model.

As mentioned earlier the model-based simulation Set-C is similar to Set-A except the distribution of the random effects. In Set-A of the simulations, the random effects are generated from normal distribution while in Set-C these are generated from the gamma distribution. Table 6.4 reports the average relative biases (%), average relative root mean squared errors (%), average coverage rate and average interval width generated by different SAE methods for Set-C. The results generated by different methods of SAE in Set-C are identical to the results in the Set-A (Table 6.2 and 6.4). This indicates that the proposed method of SAE is robust with respect to distribution of these random effects. The region-specific performance measure generated by these methods (TrMBD2, MBD1 and EBLUP) for Set-C is presented in Appendix K (Figure K.1 and K.2).

Table 6.3 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%), average (ACR) coverage rate and average (AW) 2-sigma confidence interval width for simulation set-B.

Criterion	Estimator	ParB1	ParB2	ParB3	ParB4	ParB5
ARB (%)	TrMBD1	-57.67	-13.4	-3.6	-44.6	-83.46
	TrMBD2	3.46	0.37	0.14	-0.9	-7.54
	TrMBD3	126.41	1.45	0.26	-98.36	-72.4
	TrMBD4	4.92	0.66	0.15	-1.54	-8.74
	MBD1	-0.21	0.04	0.12	0.16	-0.85
	MBD2	-0.21	0.04	0.12	0.17	-0.84
	EBLUP	-0.19	0.04	0.13	0.17	-0.77
ARRMSE (%)	TrMBD1	0.69	0.34	0.32	0.56	0.99
	TrMBD2	0.35	0.33	0.33	0.34	0.39
	TrMBD3	71.16	0.39	0.34	49.47	7.06
	TrMBD4	0.39	0.35	0.34	0.37	0.42
	MBD1	0.56	0.36	0.34	0.53	1.2
	MBD2	0.56	0.36	0.34	0.53	1.2
	EBLUP	0.38	0.3	0.29	0.36	0.56
ACR	TrMBD1	0.96	0.91	0.91	0.93	0.92
	TrMBD2	0.93	0.92	0.92	0.91	0.86
	TrMBD3	0.95	0.92	0.92	0.92	0.92
	TrMBD4	0.94	0.92	0.92	0.91	0.86
	MBD1	0.91	0.92	0.92	0.92	0.9
	MBD2	0.91	0.92	0.92	0.92	0.9
	EBLUP	0.93	0.94	0.94	0.93	0.92
AW	TrMBD1	0.09	2.6	206	5×10^4	14×10^6
	TrMBD2	0.04	2.4	207	2×10^4	5×10^6
	TrMBD3	0.4	2.7	214	20×10^4	19×10^6
	TrMBD4	0.04	2.5	211	3×10^4	5×10^6
	MBD1	0.06	2.7	214	4×10^4	13×10^6
	MBD2	0.06	2.7	214	4×10^4	13×10^6
	EBLUP	0.05	2.6	214	3×10^4	10×10^6

Figure 6.3 Region-specific percentage relative biases and percentage relative RMSEs for simulation set-B.

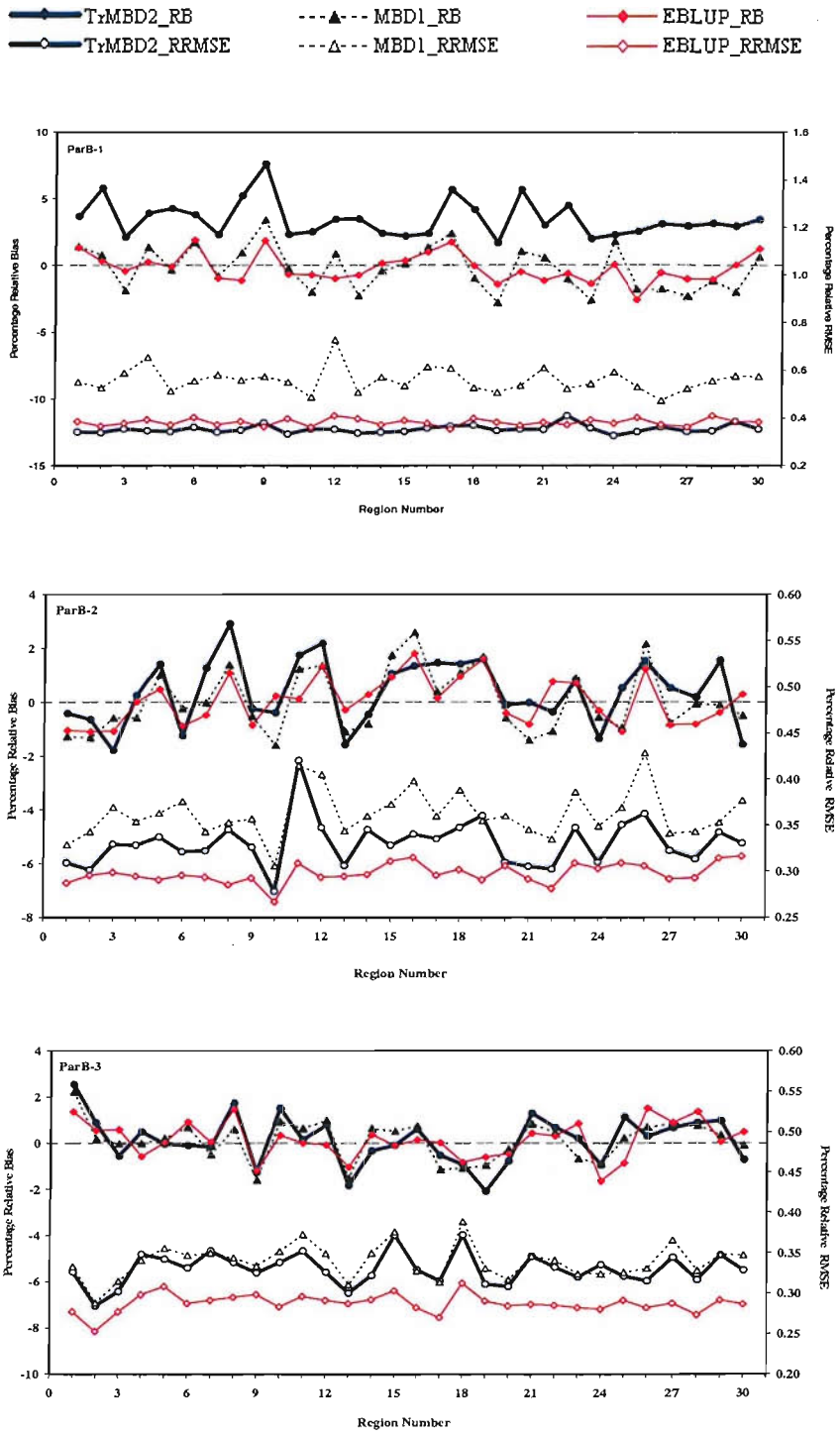


Figure 6.3 (Continued) Region-specific percentage relative biases and percentage relative RMSEs for simulation set-B.

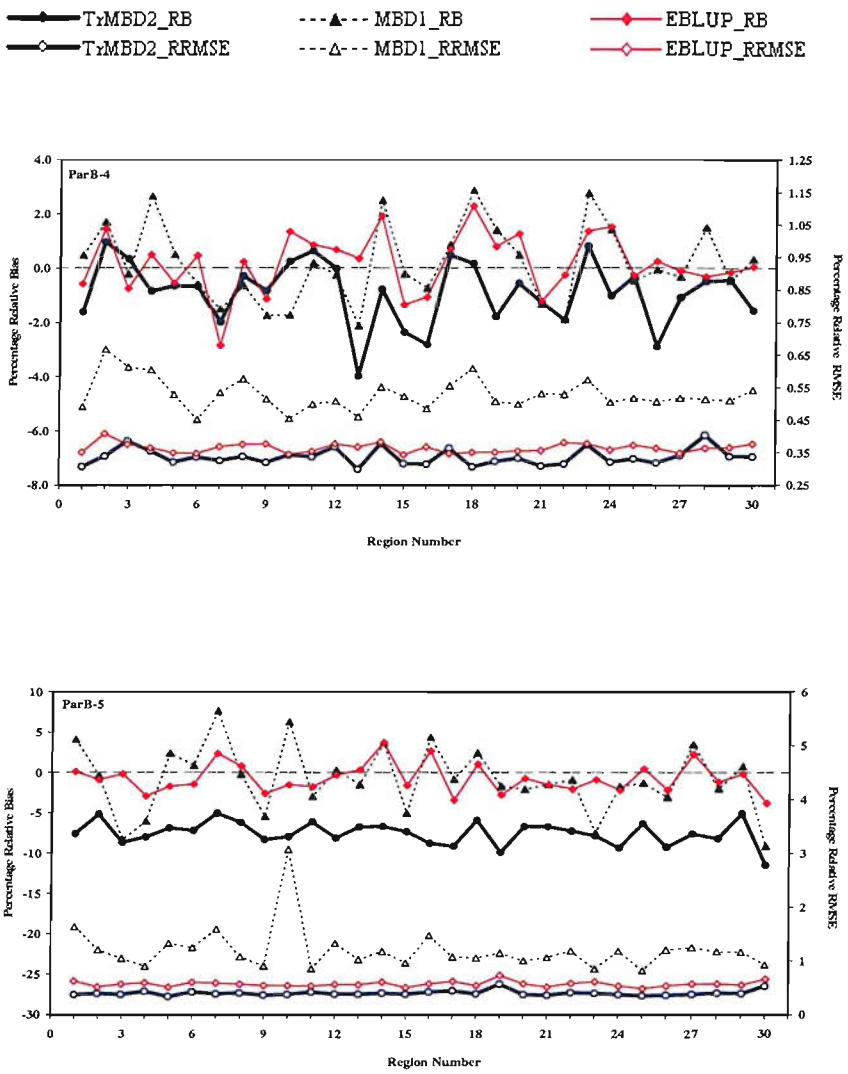


Figure 6.4 Region-specific coverage rates and confidence interval widths for simulation set-B.

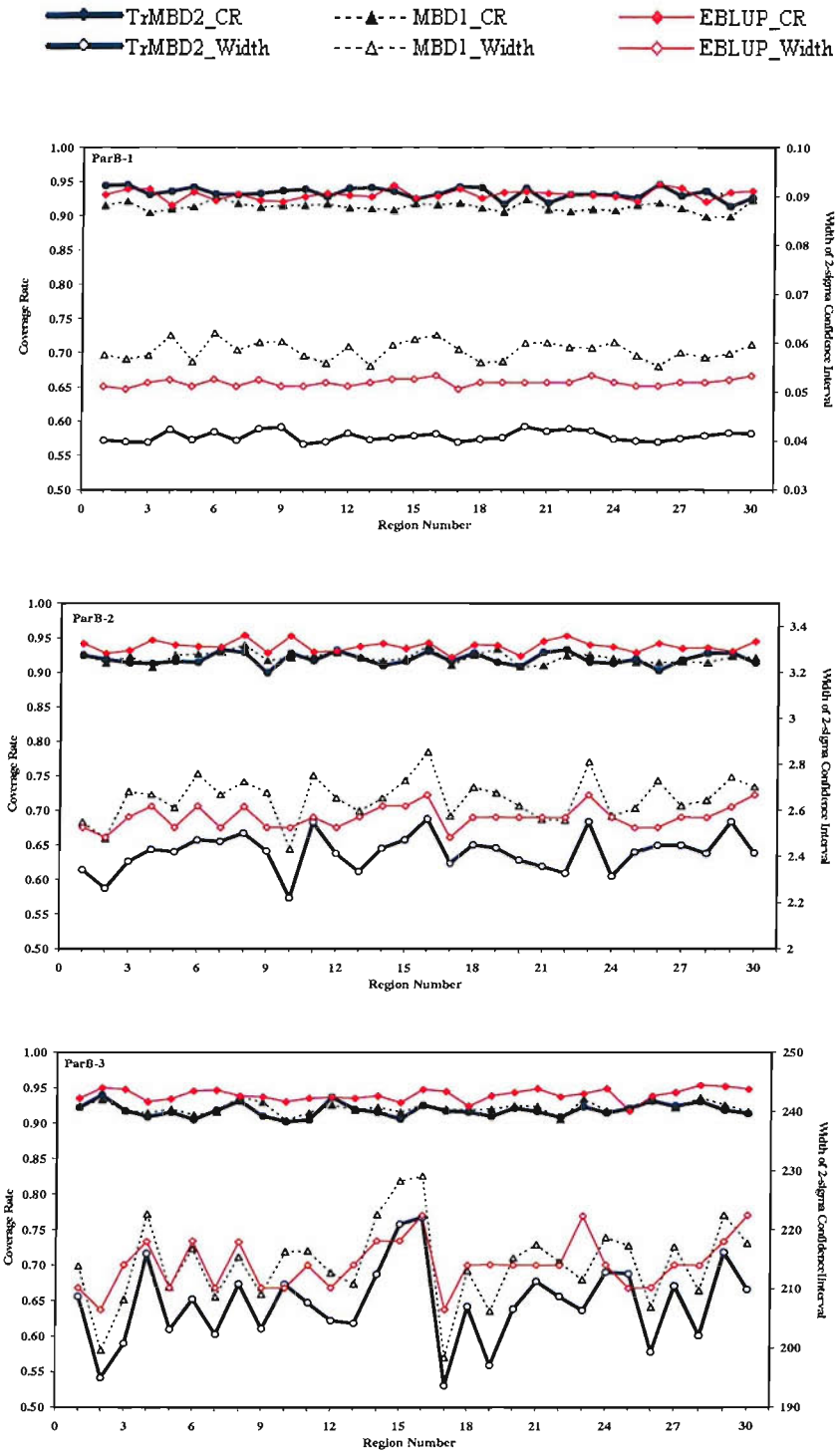


Figure 6.4 (Continued) Region-specific coverage rates and confidence interval widths for simulation set-B.

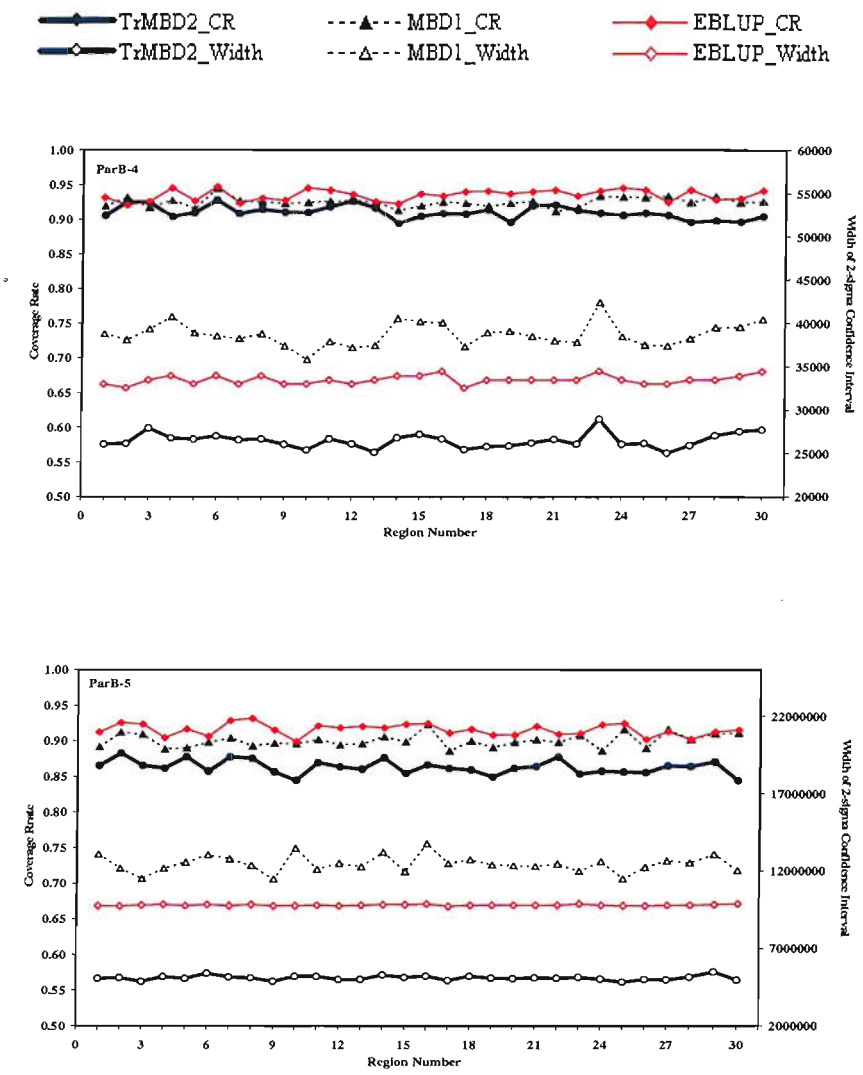


Table 6.4 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%), average (ACR) coverage rate and average (AW) 2-sigma confidence interval width for simulation set-C.

Criterion	Estimator	ParC1	ParC2	ParC3	ParC4	ParC5	ParC6
ARB (%)	TrMBD1	-85.96	-96.52	-98.46	-98.56	-98.46	-99.08
	TrMBD2	-0.08	-0.16	0.11	0.29	0.34	-0.31
	TrMBD3	-86.89	-95.79	-98.72	-98.28	-100.10	-98.92
	TrMBD4	0.02	-0.20	0.14	0.29	0.40	-0.24
	MBD1	11.30	5.27	-1.90	-3.65	-6.67	-7.36
	MBD2	13.55	6.38	-1.11	-3.38	-6.42	-7.24
	EBLUP	13.51	6.34	-0.96	-3.39	-6.64	-7.18
ARRMSE (%)	TrMBD1	0.96	1.18	1.32	1.36	1.59	1.78
	TrMBD2	0.42	0.39	0.55	0.65	0.91	1.16
	TrMBD3	3.09	1.52	1.49	1.45	2.08	1.79
	TrMBD4	0.44	0.40	0.56	0.65	0.92	1.17
	MBD1	1.48	1.68	1.90	2.70	2.61	4.03
	MBD2	1.75	1.76	1.95	2.69	2.64	4.05
	EBLUP	1.06	0.70	0.82	1.08	1.21	1.80
ACR	TrMBD1	0.97	0.96	0.94	0.93	0.91	0.90
	TrMBD2	0.85	0.89	0.88	0.87	0.88	0.87
	TrMBD3	0.97	0.96	0.94	0.93	0.91	0.90
	TrMBD4	0.85	0.90	0.88	0.87	0.88	0.87
	MBD1	0.85	0.84	0.84	0.86	0.87	0.87
	MBD2	0.84	0.84	0.84	0.86	0.87	0.87
	EBLUP	0.88	0.87	0.87	0.88	0.88	0.87
AW	TrMBD1	1881	30×10^3	21×10^4	78×10^5	53×10^6	6.4×10^7
	TrMBD2	493	7×10^3	6×10^4	26×10^5	20×10^6	27×10^7
	TrMBD3	2180	34×10^3	21×10^4	78×10^5	54×10^6	64×10^7
	TrMBD4	517	8×10^3	6×10^4	26×10^5	20×10^6	28×10^7
	MBD1	1797	32×10^3	22×10^4	97×10^5	68×10^6	98×10^7
	MBD2	1860	32×10^3	23×10^4	97×10^5	68×10^6	98×10^7
	EBLUP	784	14×10^3	10×10^4	50×10^5	38×10^6	59×10^7

6.5.2 Design Based Simulations

In section 6.5.1 we noticed that the estimator of choice for skewed data is TrMBD2. Therefore in the design-based simulations using real data from AAGIS survey we compare the performance of TrMBD2 with MBD1 and EBLUP. In previous section under model-based simulations we used a random intercept specification of model (5.9). In the design-based simulations we consider the random intercept and random slope specification of model (5.9). That is the model I and II respectively described in chapter 3. Both model I and II describes the AAGIS data, however model II (random slope model) fit is relatively better (see chapter 3). We notice that linear model fit is not very well for this data, although log-linear is slightly better, not very good (see Figure 3.2). It is interesting to see how log-log transformation based method work with this data. We used ZoneSize*FarmSize for fixed effects specification and random intercept and random intercept + random slopes for random effects specification for linear mixed model (5.9). Description on model fitting for AAGIS data is briefed in chapter 3.

Table 6.5 presents the percentage average relative biases, the percentage average relative root mean squared errors and the average coverage rates (averaged over 29 and 28 regions) generated by different estimators. Figure 6.5 and 6.6 displays the region-specific distribution of the relative biases, relative RMSEs and coverage rates under the random intercept and random slope model respectively. These results indicate relatively better performance under model II since this model is relatively better fit. Further these results show the average relative bias of TrMBD2 is smaller than EBLUP but larger than MBD1,

while the average RRMSE of TrMBD2 is marginally larger (with high average coverage rate) than the corresponding values for MBD1 and EBLUP. Inspection of Figure 6.5 and 6.6 shows that high relative bias and relative RMSE of TrMBD2 is essentially due to one region (21) in the original AAGIS sample that contained a massive outlier as noted in chapter 3. This leads to completely unrealistic estimates for region 21 being generated by the TrMBD2 and MBD1 methods. The right-hand column in Table 6.5 therefore shows the average performances of the different methods when this region is excluded. Here we see that now TrMBD2 and MBD1 are essentially on a par, with both dominating EBLUP. Region-specific results show the TrMBD2 dominates in some areas not in all (Figure 6.5 and 6.6). The fact that the TrMBD2 does not provide significant gains over the MBD1 in this case reflects the fact that the raw-scale and log-scale linear mixed models used in these estimators both provide relatively poor fits to the AAGIS data.

The TrMBD2 estimator provides significant gain under the linearity on transform model. However, gain may not be significant if linearity does not hold. At the same time, we noticed when transform model is approximately linear then it is safer to use TrMBD2 method. We recall that AAGIS data is extremely heteroskedastic and analysis of original sample data indicates a week linear relationship between Y (annul farm cost) and X (farm size) which improves when we fit a log-linear models (Figure 3.2). However, fitted model on log-transform is not exactly linear (although linear on log scale in few areas). Therefore, the TrMBD2 performs marginally better and provides a gain in those regions where linearity holds, not in all regions.

Table 6.5 Average (ARB) values of relative bias (%), average (ARRMSE) values of relative root mean squared error (%) and average (ACR) coverage rate for design based simulation using AAGIS data.

Model	Criterion	Estimator	Average of 29 regions	Average of 28 regions
I	ARB (%)	TrMBD2	3.00	2.54
		MBD1	-2.49	-2.58
		EBLUP	4.24	4.74
	ARRMSE (%)	TrMBD2	22.00	17.15
		MBD1	20.55	17.33
		EBLUP	19.92	19.40
	ACR	TrMBD2	0.99	0.99
		MBD1	0.92	0.93
		EBLUP	0.90	0.90
II	ARB (%)	TrMBD2	2.35	2.24
		MBD1	-2.13	-2.21
		EBLUP	2.98	3.36
	ARRMSE (%)	TrMBD2	21.31	17.13
		MBD1	20.15	16.91
		EBLUP	19.87	19.30
	ACR	TrMBD2	0.90	0.92
		MBD1	0.93	0.95
		EBLUP	0.85	0.85

Figure 6.5 Region-specific percentage relative biases and percentage relative RMSEs and coverage rates for AAGIS data under model-I.

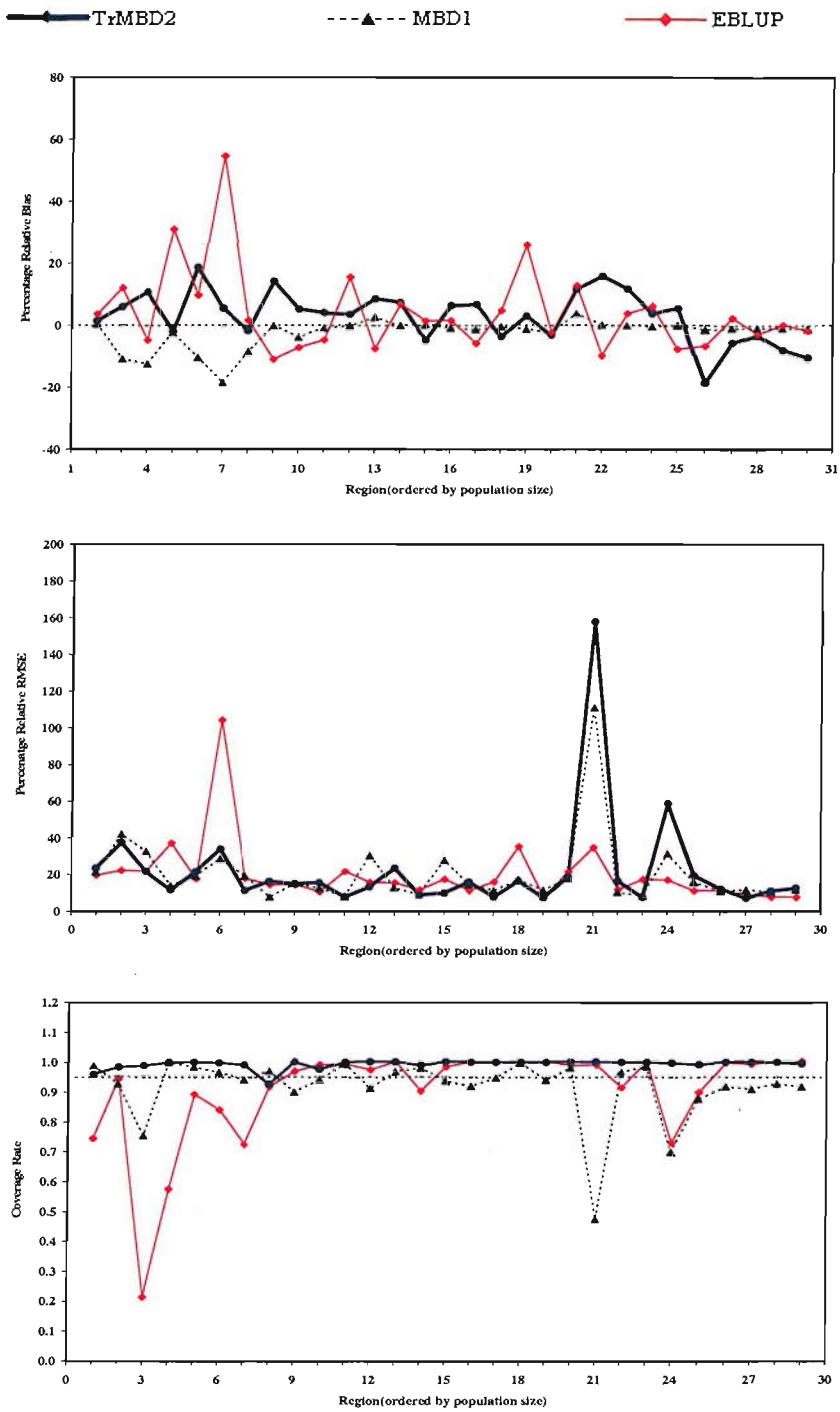
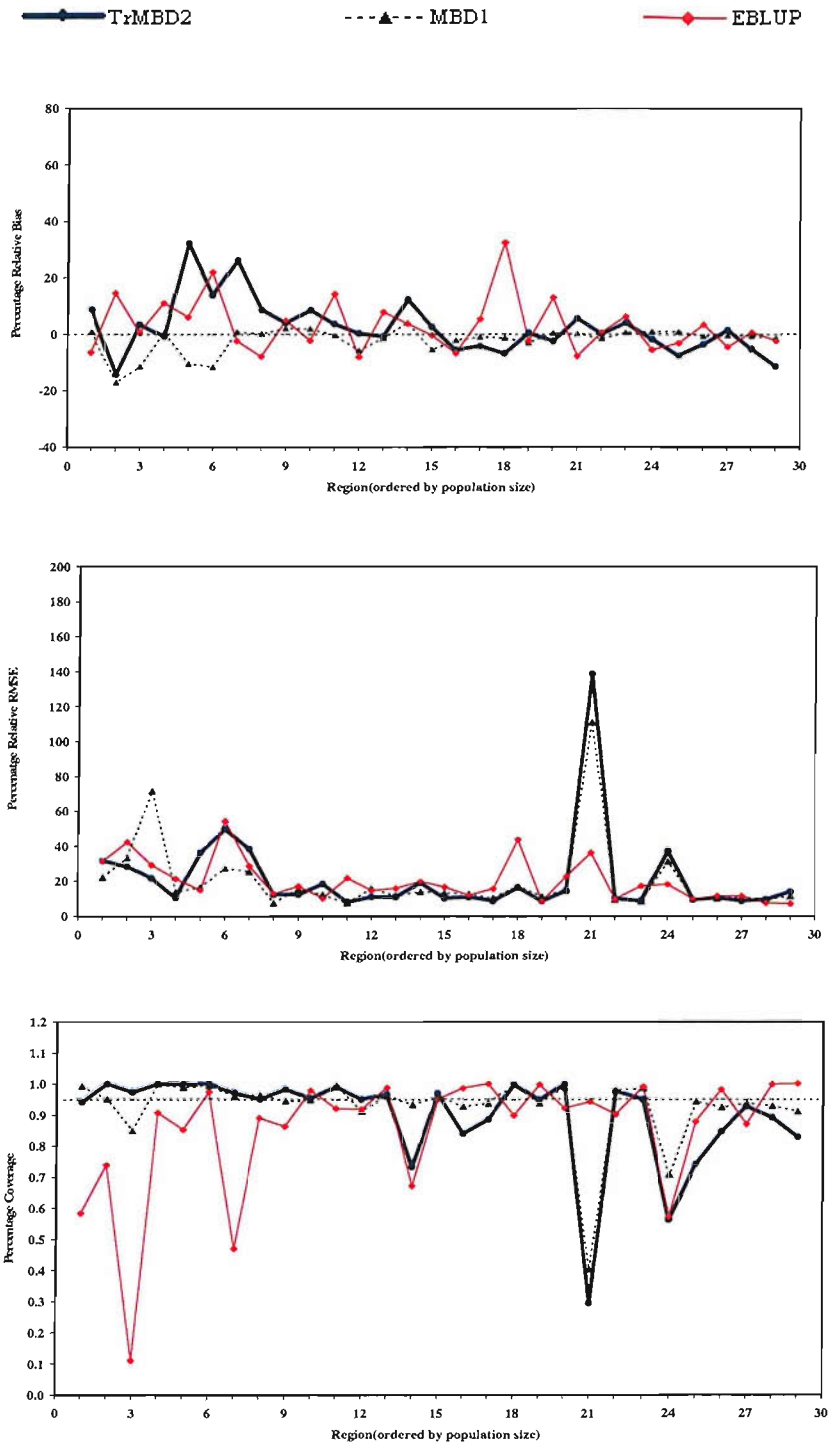


Figure 6.6 Region-specific percentage relative biases and percentage relative RMSEs and coverage rates for AAGIS data under model-II.



6.6 Conclusions

We now summarize the main points from the evaluation of methodology presented in chapter 5. The simulation results discussed in the previous sections show that combining model-calibrated weights with MBD estimation can bring significant gains in SAE efficiency if the population data are clearly non-linear. As one would expect, these gains are less when the assumed non-linear model is misspecified.

In chapter 5 we noticed that the proposed method of SAE is robust with respect to distribution of the random effects. We investigated the proposed method of SAE under normal and *gamma* distribution of random effects via simulation studies. Our conclusions are essentially unaffected when we carry out similar simulations using gamma distributed random effects. The application of the proposed SAE techniques to real data from AAGIS provides a satisfactory performance. The proposed method is advisable for skewed data but identification of appropriate transform model is crucial in application this method, otherwise results can be misleading. We also examine the performance of an empirical best predictor under a log-scale linear mixed model (TrEBP). The results generated by TrEBP are presented in Appendix L.

Our main caveat concerning the use of model-calibrated weights for SAE is their specificity. These weights do not appear to have the same 'multipurpose' characteristics as standard EBLUP weights based on the linear mixed models (see chapter 4). Further research is therefore required on how to build model-calibrated weights for SAE that are

less specific in the way they work. It is to be expected that such weights would not be as efficient as the variable specific weights (5.8), but hopefully this will be more than offset by their increased utility. A further issue that is extremely important in practice is that positively skewed survey variables can also take zero (or even negative) values. Consequently, the log-scale linear mixed model that underpins the model-calibration weighting considered in chapter 5 and 6 needs to be suitably generalised.

CHAPTER 7

SUMMARY AND FURTHER RESEARCH

7.1 Introduction

In this chapter we summarize the research work presented in different chapters of this thesis, highlighting the major points. We identify related topics and outstanding issues that require further attention. In section 7.2 we present the principal results and conclusions from different chapters of this dissertation. Finally, section 7.3 addresses the potential further research topics.

7.2 Summary

The purpose of the research presented in this thesis is to develop methodology for small area estimation (SAE) that is simple and also easy to implement. Further, using the real data set we investigated several existing methods for SAE and proposed a few new approaches to SAE that overcomes the problem identified in the existing techniques. We focused on weighted linear estimators for small areas and their mean squared error estimation. In particular, we used the calibrated weighting approach introduced in Chambers (2005). In this thesis we referred this approach as the model based direct (MBD) method for SAE. We compare the performance of the MBD method of SAE with the standard empirical best linear unbiased prediction (EBLUP) via empirical studies. Then we extended the MBD method of small area estimation for

multivariate and business surveys. In a broad sense, we can subdivide this thesis into three major topics:

(a) study the properties of the model-based direct estimators for small area quantities and compare with the EBLUP (in chapter 3) method, and some further application of MBD estimation, e.g. estimation of small areas for categorical survey variables.

(b) illustrate loss functions that can be used to compute optimal multipurpose weights suitable for use in small area estimation using MBD estimators for multivariate surveys (in chapter 4), and

(c) develop small area estimation methods for skewed data e.g. business surveys, where data are typically skewed and linear model assumptions are questionable (in chapter 5 and 6).

In this section we now summarize our basic findings from the different chapter and give some directions for future research in next section.

In chapter 2 we reviewed some of the important small area estimation methods existing in literature, identified some gaps existing in the present research and pointed out the problem to be addressed in this thesis. This chapter prepared a foundation for the rest of the thesis. Consequently, in chapter 3 we focus on small area estimators that are a weighted linear function of the area specific sample data that we referred to as MBD estimation. The EBLUP method is widely used approach for the estimation of small areas under unit level mixed effect models. However, this approach does not lead to small area estimators that are a weighted linear function of the sample data from these areas. As a result, several practical advantage of using such weighted estimators are lost, with probably the most important being the relative simplicity of

their mean squared error estimation. In this chapter we studied the properties of the MBD. This approach uses weights derived from a population level version of the random effects model to define weighted linear small area estimators and a simple expression for their MSE. The associated small area estimator appears to be a direct estimator based on the sample data from each area. However, it is not true in general. The sample weights are a function of the data from the entire sample. Note that unlike design based direct estimation, MBD weights borrow strength via random effects model that defines the weights.

In general, unbiased direct estimators for small area quantities are usually considered too variable to be of any practical use. In this chapter we observed that the MBD estimator for small area quantities appears to overcome this objection, in the sense that these estimators are comparable in efficiency to the indirect model-based small area estimators (e.g. EBLUPs) that are now widely used. There are many practical advantages associated with such MBD estimators, arising from the fact that they are computed as weighted linear combinations of the actual sample data from the small areas of interest. Note that in this case the weights 'borrow strength' via a model that explicitly allows for small area effects. One particular advantage that we explore in this chapter is that estimation of mean squared error is then straightforward, using well-known methods that are in common use for population level estimates. Empirical results reported in this chapter show that the MBD estimator represents a real alternative to the EBLUP, with the simple MSE estimator associated with the MBD estimator providing good coverage performance. We also report results that indicate that the MBD estimator may be more robust than the EBLUP when the small area model is incorrectly specified.

An application of MBD estimation to the categorical variable that takes 0 and 1 value shows a satisfactory performance of the methods. In particular, we observed no loss in efficiency by using a linearity assumption based MBD for the binary variable. Further, the method is comparable with usual indirect method of SAE based on a generalized linear mixed model. In contrast, on many occasions a standard EBLUP based on a linear mixed model generated estimates that are greater than 1. However, this is not the case with the MBD estimator.

Sample surveys are generally multivariate, in the sense that they measure more than one response variable. In theory, each variable can then be assigned an optimal weight for estimation purposes. However, it is often a distinct practical advantage to have a single weight that is used with all variables collected in the survey. In chapter 4 we consider SAE for a multivariate survey and introduce two loss functions that can be used to compute optimal multipurpose weights suitable for use in SAE using MBD estimators. We consider two case: (*a*) we ignore the correlations between the survey variables; and (*b*) we take these correlations between the survey variables into account. From the results generated under design-based simulations (using real population data) and model based simulations (using generated data under the model), we see that the performance of the corresponding multipurpose weighting based estimators under (a) and (b) are almost identical. That is, there are no real gains from taking account of the correlations between the survey variables when constructing the multipurpose weights. We discuss two methods of constructing multipurpose weights for use in MBD small area estimation based on: (i) weighted average of the variance components; and (ii) suitably averaging the variable specific EBLUP weights. Empirical results show that method (ii) is somewhat less efficient than the method (i).

Our results also show that these multipurpose weights remain efficient across a wide range of variables, even variables that have not been used in the definition of the multipurpose weights. This can be important in some situations (e.g. where variables have many zero values) where standard mixed models cannot be fitted and the usual EBLUP methods do not work. Further, in defining the multipurpose weights we can also assign importance factors based on the intrinsic variability among of the variables. In our empirical studies, we use two options for importance factors: $\phi_k = 1/\Sigma_{e,k}$ and $\phi_k = 1/V_k$, where $\Sigma_{e,k}$ and V_k are the individual and total variability of the k^{th} target variable. These results show that, for the population considered in the simulation study, there is little to choose between these different importance weighting factors.

The central theme of chapter 5 and 6 is SAE for skewed data. In business surveys, data typically are skewed and the standard approach for SAE based on linear mixed models lead to inefficient estimates. In chapter 5 we introduced SAE techniques for skewed data that are linear following a suitable transformation, focusing on the widely used log-log transformation. In particular, we extended the MBD approach described in chapter 3 and 4 to skewed data using a model with random area effects that is linear in the log scale and sample weights derived via model calibration. We presented the theoretical developments in this chapter. In chapter 6 we then provided illustrative empirical results that contrast the proposed MBD estimator for skewed data with the EBLUP and the MBD method under a linear mixed model.

The simulation results reported in chapter 6 show that combining model-calibrated weights with MBD estimation can bring significant gains in SAE efficiency if the

population data are clearly non-linear in the raw scale, but linear in the log scale. As one would expect, there are smaller gains when the assumed non-linear model is misspecified. Furthermore, our conclusions are essentially unaffected when our simulations use gamma, rather than Gaussian, distributed random effects. That is the proposed method is robust with respect to the usual normality assumption for the area effects. An application to real life business survey data (AAGIS data) provides a further demonstration of the satisfactory performance of the proposed MBD method. The proposed method is advisable for skewed data, however examination of appropriate model relationship is very crucial in application of this method, otherwise results can be misleading.

7.3 Further research

In chapter 3 we described the MBD estimation and simple mean squared error estimation for this estimator. This approach treats these estimators as simple weighted estimators of a domain mean. Under this approach the sample weights are considered fixed and the prediction variance is estimated using a standard heteroskedasticity robust variance estimator. A 'plug-in' estimate of the squared bias is then added to this estimated prediction variance to define a simple estimator of the mean squared error of these estimators. Chambers (2005) advocated the use of this MSE estimator with the justification that method is consistent with the way mean squared error is estimated at the population level. Empirical results reported in chapter 3 based on AAGIS data, show that the simple MSE estimator associated with the MBD estimator provides good coverage performance. Further, these results indicate that this estimator may be more robust than the EBLUP when the small area model is incorrectly

specified. Although this method of MSE estimation seems to be working reasonably well, however it remains to develop its theoretical proof and justifications. Further, generalization of this method of MSE estimation for any weighted direct or indirect (e.g., EBLUP or M-quantile methods, Chambers and Tzavidis, 2006) small area estimators is interesting and demanding as well.

Negative weights impact on the utility of the MBD method and this remains unresolved and needs further attention. For example, negative weights, which occurred in some regions in the simulation study reported in chapter 3, can lead to impossible (i.e. negative) estimates. Since such values are easily identified, they should not cause problems in real life. However, the problem remains of how to modify the weights to ensure they are strictly positive. A related issue that has already been noted is the impact of outlier Y-values on (3.15). Certainly this estimator, because it is a linear combination of just the small area data values, is more susceptible to outliers in specific areas than the EBLUP. Methods for dealing with negative weights under 'standard' regression models have been discussed in the literature (Huang and Fuller, 1978; Bardsley and Chambers, 1984; Deville and Sarndal, 1992; Chambers, 1996) but their application in the context of mixed models remains to be explored.

Throughout this thesis we assume that random area effects are independent between areas. However, we can extend the MBD approach under spatially correlated random area effect model (spatial-MBD) similar to the spatial-EBLUP (Singh, Shukla and Kundu, 2005, Petrucci and Salvati, 2004 and Pratesi and Salvati, 2005) and spatial M-quantile (Chambers, Pratesi, Salvati and Tzavidis, 2006) method of SAE. Further, it is

interesting to see the performance of Spatial MBD with spatial-EBLUP and spatial M-quantile methods of SAE for spatially correlated population data. Furthermore, we can extend nonparametric methods to MBD estimation, see for example Opsomer et al (2005). In chapter 4 we concluded that the MBD approach based on multipurpose weights can be important in some situations (e.g. where variables have many zero values) where standard mixed models cannot be fitted and the usual EBLUP methods do not work. In such cases, we can extend the EBLUP approach under the mixtures of linear mixed models.

In chapters 5 and 6 we proposed a method of SAE for skewed data based on the log-scale linear model where survey variables can have *only* strictly positive values. In practice positively skewed survey variables can also take zero (or even negative) values. Consequently, the log-scale linear mixed model that underpins the model-calibration weighting needs to be suitably generalised. Karlberg (2000a) and Fletcher *et at.* (2005) illustrate the application of a mixture model for skewed data with zeros. Further, one can use a generalized linear mixed model with Gamma or Poisson (for count data) or other class of distributions for skewed data with zeros. Joe, Chris, and Mark (2005) described the neglog transformation for skewed data with negative values. A further issue concerning the use of model-calibrated weights for SAE is their specificity. These weights do not appear to have the same 'multipurpose' characteristics as standard EBLUP weights based on linear mixed models. Further research is therefore required on how to build model-calibrated weights for SAE that are less specific in the way they work. It is to be expected that such weights would not be as efficient as the variable specific weights, but hopefully this will be more than offset by their increased utility.

APPENDIX A

COMPARING RANDOM EFFECTS SPECIFICATION FOR THE MIXED MODEL IN CHAPTER 3

In chapter 3 for judging the best-fitted model to the AAGIS data we use the Akaike Information Criterion (AIC) evaluated as $AIC = -2\log Lik + 2k$, where k is the number of parameters in the model and *LogLik* is log-likelihood of the model. Under this definition, smaller the value of AIC is the better. In addition, we use the likelihood ratio (LR) test as criteria to find the best model. The values of test criterions obtained from ANOVA function in **R** for the random intercept and random slope model (i.e. models I and II in chapter 3) using AAGIS data are set out in Table A.1.

Table A.1 Analysis of variance (ANOVA) results for comparing two models.

Model	degree of freedom	AIC	logLik	LR	p-value
I	14	49992	-24982		
II	16	49989	-24979	6.43	0.04

The small p-value for the test statistics indicates the model II is better than the model I. The AIC criterion is nearly same for both models (but marginally smaller for model II). Consequently we conclude that model II is relatively better than model I for this data.

APPENDIX B

REGION-SPECIFIC RESULT USING ML ESTIMATES OF VARIANACE COMPONENTS IN CHAPTER 3

Figure B.1 Region-specific percentage relative biases for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with ML estimates.

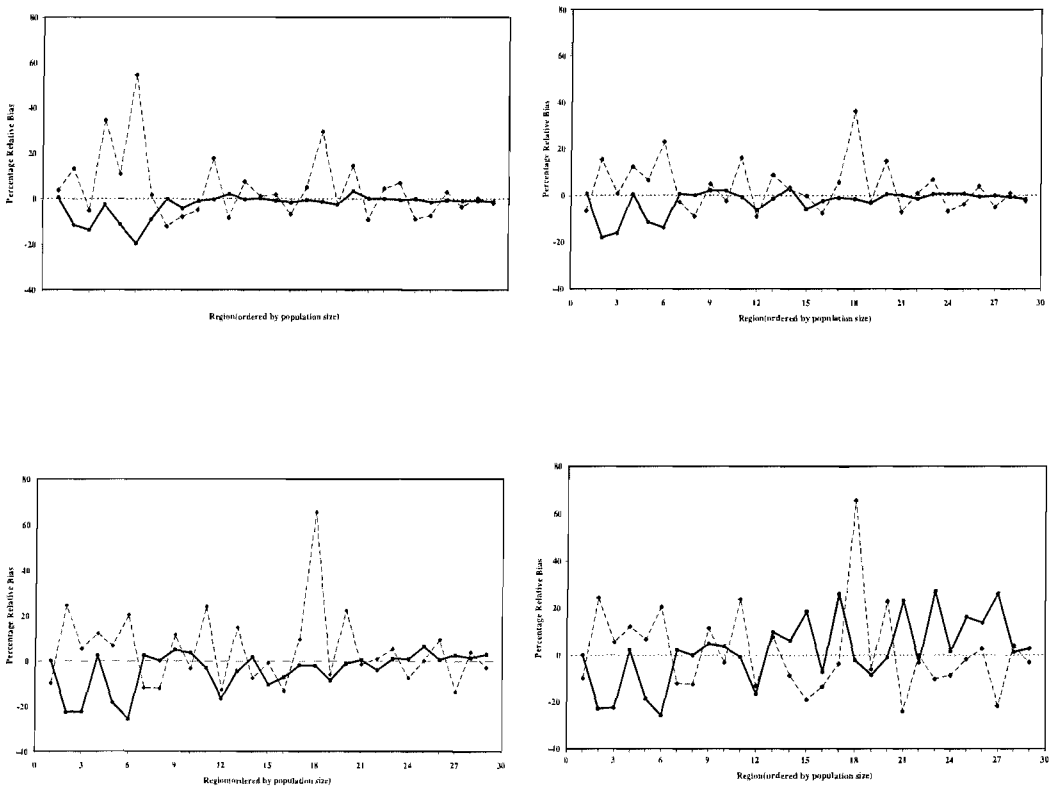


Figure B.2 Region-specific percentage relative RMSE for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with ML estimates.

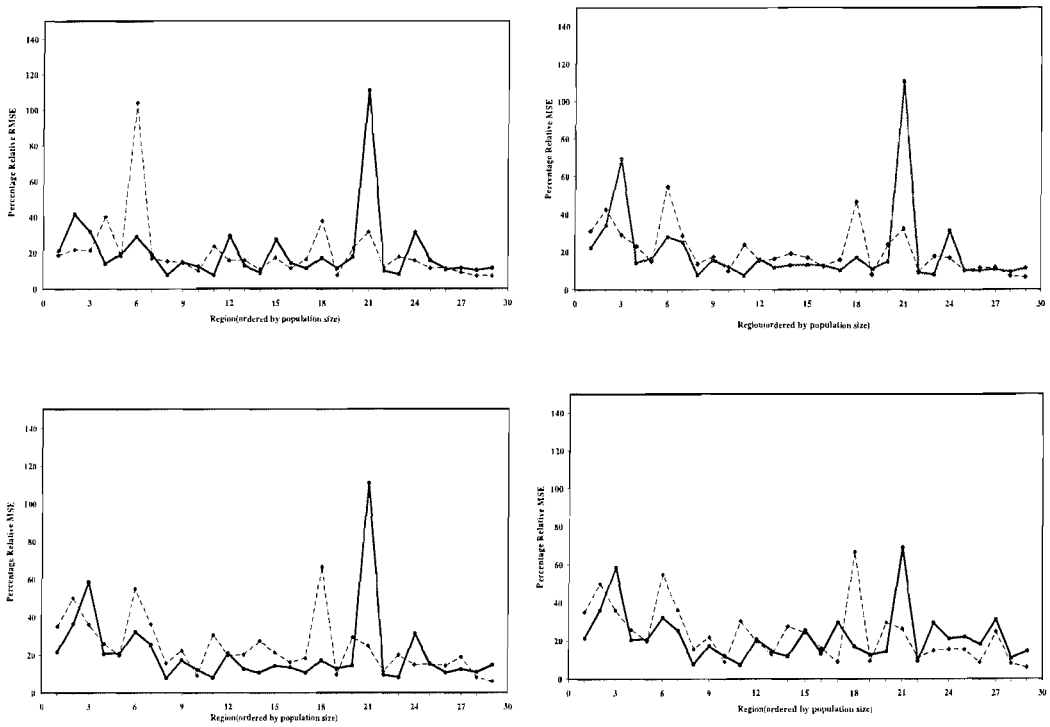
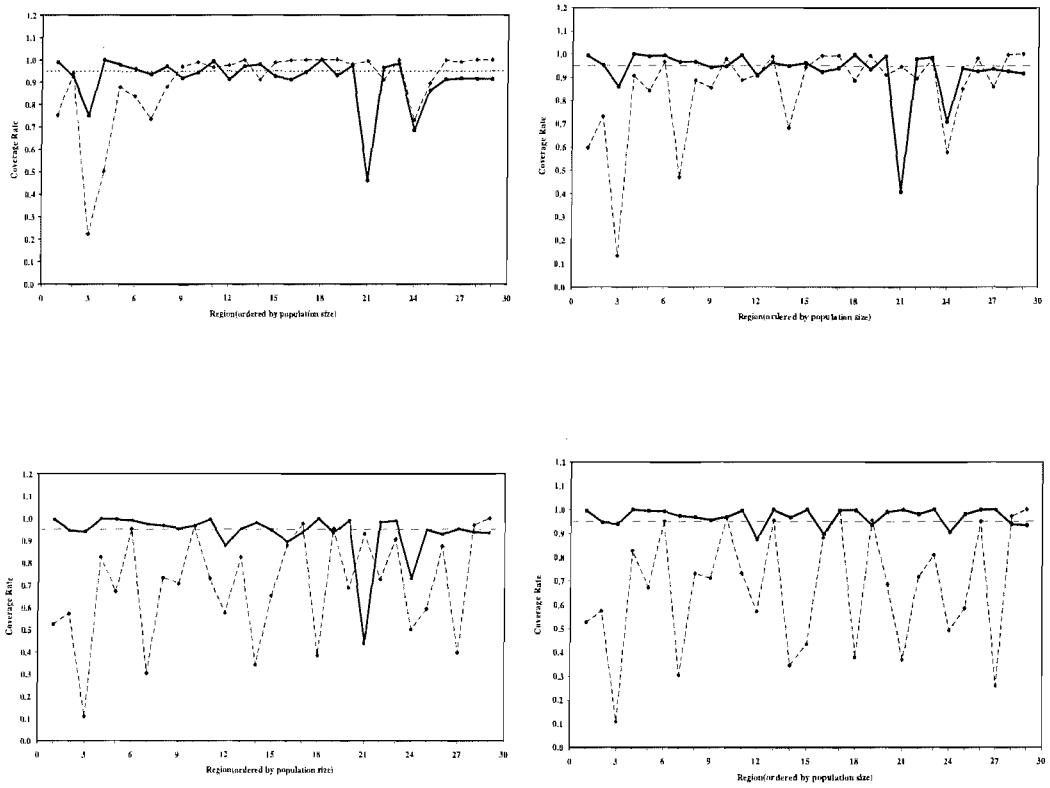


Figure B.3 Region-specific coverage rates for EBLUP (dashed line) and MBD (solid line) under model I (top left), model II (top right), model III (bottom left) and model IV (bottom right) with ML estimates.



APPENDIX C

BLUP, MBD AND DBD ESTIMATOR FOR SMALL AREAS

We present an analytical comparison of the BLUP, MBD and design based direct (DBD) estimators for the small area estimation (SAE). Let us consider the random intercept specification of linear mixed model (3.11) as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij} \qquad \text{(C.1)}$$

where y_{ij} and x_{ij} are the values of Y and X respectively for population unit $j(j = 1,....,N_i)$ in small area $i(i = 1,....m)$. Let $E(u_i) = E(e_{ij}) = 0$, $Var(u_i) = \sigma_u^2$, $Var(e_{ij}) = \sigma_e^2$, $Cov(y_{ij}, y_{ik}) = \sigma_u^2$, and $Var(y_{ij}) = \sigma_e^2 + \sigma_u^2$. We defined different terms defined similar to as below equation (3.11) in chapter 3 except that X is now scalar. Assuming model (C.1) holds with a special case of $\sigma_u^2 = 0$, the sample weights defining the BLUP of the population total of Y, $T_y = \sum_{j \in U} y_j$ are

$$\begin{aligned}
\hat{T}_y &= \sum_{j \in s} y_j + \sum_{j \in r} \hat{y}_j = \sum_{j \in s} y_j + \sum_{j \in r} (\hat{\beta}_0 + \hat{\beta}_1 x_j) \\
&= \sum_s y_j + \sum_r (\bar{y}_s - \hat{\beta}_1 \bar{x}_s + \hat{\beta}_1 x_j) \\
&= N\bar{y}_s + (N-n)\hat{\beta}_1 (\bar{x}_r - \bar{x}_s) \\
&= N\bar{y}_s + (N-n)(\bar{x}_r - \bar{x}_s) \frac{\sum_{j \in s} (x_j - \bar{x}_s) y_j}{\sum_{j \in s} (x_j - \bar{x}_s)^2} = \sum_{j \in s} w_j y_j
\end{aligned} \qquad \text{(C.2)}$$

where $w_j = \frac{N}{n} + \frac{(N-n)(\bar{x}_r - \bar{x}_s)(x_j - \bar{x}_s)}{(n-1)s_x^2}$

with $(n-1)s_x^2 = \sum_{j\in s}(x_j - \bar{x}_s)^2$

and $\bar{y}_s = n^{-1}\sum_{j\in s} y_j$, $\bar{x}_s = n^{-1}\sum_{j\in s} x_j$, $\bar{x}_r = (N-n)^{-1}\sum_{j\in r} x_j$.

If we use these BLUP weights (C.2) to define the MBD estimator at small area level for the population total of *Y* for small area *i* then

$$\begin{aligned} E(\hat{T}_{y_i} - T_{y_i}) &= \sum_{j\in s_i} w_j y_j - \sum_{j\in U_i} y_j \\ &= \sum_{j\in s_i} w_j(\beta_0 + \beta_1 x_j) - \sum_{j\in U_i}(\beta_0 + \beta_1 x_j) \\ &= \beta_0\left(\sum_{j\in s_i} w_j - N_i\right) + \beta_1\left(\sum_{j\in s_i} w_j x_j - \sum_{j\in U_i} x_j\right) \end{aligned} \qquad \text{(C.3)}$$

where s_i and $U_i (i = 1,....m)$ respectively denote the set of sample and population unit in the small area *i*. Let us write the weights in (C.2) as

$$w_j = \frac{N}{n}(1 + g_j), \qquad \text{(C.4)}$$

where $g_j = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right](x_j - \bar{x}_s)$ then

$$E(\hat{T}_{y_i} - T_{y_i}) = \beta_0\left[\sum_{j\in s_i}\frac{N}{n}(1+g_j) - N_i\right] + \beta_1\left[\sum_{j\in s_i}\frac{N}{n}(1+g_j)x_j - \sum_{j\in U_i} x_j\right]$$

$$= \beta_0\left(\frac{N}{n}n_i - N_i\right) + \beta_1\left(\frac{N}{n}n_i\bar{x}_{s_i} - N_i\bar{X}_i\right) + \frac{N}{n}\left(\sum_{j\in s_i} g_j(\beta_0 + \beta_1 x_j)\right) \qquad \text{(C.5)}$$

$$Var(\hat{T}_{y_i} - T_{y_i}) = Var\left(\sum_{j\in s_i} w_j y_j - \sum_{j\in U_i} y_j\right) = Var\left(\sum_{j\in s_i}(w_j - 1)y_j - \sum_{j\in r_i} y_j\right)$$

$$= \sum_{s_i}(w_j - 1)^2 Var(y_j) + \sum_{r_i} Var(y_j) - 2Cov\left(\sum_{s_i}(w_j - 1)y_j, \sum_{r_i} y_k\right)$$

$$= \left\{ \sum_{s_i} (w_j - 1)^2 + (N_i - n_i) \right\} \sigma_e^2 \text{ , since covariance term is zero}$$

$$= \left\{ \sum_{s_i} \left(\frac{N}{n}(1 + g_j) - 1 \right)^2 + (N_i - n_i) \right\} \sigma_e^2$$

$$= \left\{ \sum_{s_i} \left[\left(\frac{N}{n} - 1 \right)^2 + 2\frac{N}{n}\left(\frac{N}{n} - 1 \right) g_j + \left(\frac{N}{n} \right)^2 g_j^2 \right] + (N_i - n_i) \right\} \sigma_e^2$$

$$= \left\{ n_i \left(\frac{N}{n} - 1 \right)^2 + (N_i - n_i) + 2\frac{N}{n}\left(\frac{N}{n} - 1 \right) \left(\sum_{s_i} g_j \right) + \frac{N^2}{n^2} \left(\sum_{s_i} g_j^2 \right) \right\} \sigma_e^2 \quad \text{(C.6)}$$

C.1 Expansion Weights and Regression Weights for Small Area Estimation

Let $\hat{T}_{y_i}^{Reg}$ and $\hat{T}_{y_i}^{Exp}$ respectively denote the regression weighted and expansion weighted estimator for the population total of Y in small area i, defined as

$$\hat{T}_{y_i}^{Exp} = \sum_{j \in s_i} w_j y_j = \frac{N}{n} \left(\sum_{j \in s_i} y_j \right) \quad \text{(C.7)}$$

$$E(\hat{T}_{y_i}^{Exp} - T_{y_i}) = \frac{N}{n} \left(\sum_{j \in s_i} (\beta_0 + \beta_1 x_j) \right) - \left(\sum_{j \in U_i} (\beta_0 + \beta_1 x_j) \right)$$

$$= \beta_0 \left(\frac{N}{n} n_i - N_i \right) + \beta_1 \left(\frac{N}{n} n_i \bar{x}_{s_i} - N_i \bar{X}_i \right) \quad \text{(C.8)}$$

$$Var(\hat{T}_{y_i}^{Exp} - T_{y_i}) = Var\left(\frac{N}{n} \sum_{s_i} y_j - \sum_{U_i} y_j \right) = Var\left[\left(\frac{N}{n} - 1 \right) \sum_{s_i} y_j - \sum_{r_i} y_j \right]$$

$$= \left(\frac{N}{n} - 1 \right)^2 Var\left(\sum_{s_i} y_j \right) + Var\left(\sum_{r_i} y_j \right)$$

$$= \left\{ \left(\frac{N}{n} - 1 \right)^2 n_i + (N_i - n_i) \right\} \sigma_e^2 \quad \text{(C.9)}$$

$$E(\hat{T}_{y_i}^{Reg} - T_{y_i}) = E(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \frac{N}{n}\left(\sum_{j\in s_i} g_j(\beta_0 + \beta_1 x_j)\right) \quad \text{(C.10)}$$

$$Var(\hat{T}_{y_i}^{Reg} - T_{y_i}) = Var(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \left\{2\frac{N}{n}\left(\frac{N}{n} - 1\right)\left(\sum_{s_i} g_j\right) + \frac{N^2}{n^2}\left(\sum_{s_i} g_j^2\right)\right\}\sigma_e^2 \quad \text{(C.11)}$$

From (C.4) we can write

$$\sum_{j\in s_i} g_j = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right] n_i(\bar{x}_{s_i} - \bar{x}_s) \quad \text{(C.12)}$$

$$\sum_{s_i} g_j^2 = \left(1 - \frac{n}{N}\right)^2\left(\frac{n}{n-1}\right)^2\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]^2 \sum_{s_i}(x_j - \bar{x}_s)^2$$

$$= \left(1 - \frac{n}{N}\right)^2\left(\frac{n}{n-1}\right)^2\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]^2 \sum_{s_i}(x_j - \bar{x}_{s_i} + \bar{x}_{s_i} - \bar{x}_s)^2$$

$$= \left(1 - \frac{n}{N}\right)^2\left(\frac{n}{n-1}\right)^2\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]^2 \left[(n_i - 1)s_{ix}^2 + n_i(\bar{x}_{s_i} - \bar{x}_s)^2\right] \quad \text{(C.13)}$$

$$\sum_{s_i} g_j x_j = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right] \sum_{j\in s_i}(x_i - \bar{x}_s)x_j$$

$$= \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]\left\{\sum_{j\in s_i}(x_j - \bar{x}_{s_i})x_j + \sum_{j\in s_i}(\bar{x}_{s_i} - \bar{x}_s)x_j\right\}$$

$$= \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]\left\{(n_i - 1)s_{jx}^2 + n_i(\bar{x}_{s_i} - \bar{x}_s)\bar{x}_{s_i}\right\} \quad \text{(C.14)}$$

It follows

$$\sum_{s_i} g_j(\beta_0 + \beta_1 x_j) = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]\left\{n_i(\bar{x}_{s_i} - \bar{x}_s)(\beta_0 + \beta_1\bar{x}_{s_i}) + (n_i - 1)\beta_1 s_{ix}^2\right\}$$

(i) For $s_i = s$, i.e. $n_i = n$

$$E(\hat{T}_{y_i}^{Reg} - T_{y_i}) = E(\hat{T}_y^{Reg} - T_y) = E(\hat{T}_y^{Exp} - T_y) + \frac{N}{n}\sum_{s_i} g_j(\beta_0 + \beta_1 x_j)$$

$$= E(\hat{T}_y^{Exp} - T_y) + \frac{N}{n}\left(1-\frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]$$
$$\left\{n_i(\bar{x}_{s_i} - \bar{x}_s)(\beta_0 + \beta_1\bar{x}_{s_i}) + (n_i - 1)\beta_1 s_{ix}^2\right\}$$

$$= E(\hat{T}_y^{Exp} - T_y) + \frac{N}{n}\left(1-\frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{(\bar{x}_r - \bar{x}_s)}{s_x^2}\right]\left\{(n-1)\beta_1 s_x^2\right\}$$

$$= E(\hat{T}_y^{Exp} - T_y) + \beta_1(N-n)(\bar{x}_r - \bar{x}_s) \qquad \text{(C.15)}$$

(ii) For $\bar{x}_{s_i} \approx \bar{x}_s$

$$E(\hat{T}_{y_i}^{Reg} - T_{y_i}) = E(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \beta_1(N-n)(\bar{x}_r - \bar{x}_s)\frac{(n_i-1)s_{ix}^2}{(n-1)s_x^2}, \text{ and}$$

$$Var(\hat{T}_{y_i}^{Reg} - T_{y_i}) = Var(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \left\{\frac{(N-n)^2}{(n-1)^2}(n_i-1)\frac{(\bar{x}_r - \bar{x}_s)^2}{s_x^2}\frac{s_{ix}^2}{s_x^2}\right\}\sigma_e^2$$

If we also have $s_{ix}^2 \equiv s_x^2$

$$E(\hat{T}_{y_i}^{Reg} - T_{y_i}) = E(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \beta_1(N-n)(\bar{x}_r - \bar{x}_s)\frac{(n_i-1)}{(n-1)} \qquad \text{(C.16)}$$

$$Var(\hat{T}_{y_i}^{Reg} - T_{y_i}) = Var(\hat{T}_{y_i}^{Exp} - T_{y_i}) + \left\{\frac{(N-n)^2}{(n-1)^2}(n_i-1)\frac{(\bar{x}_r - \bar{x}_s)^2}{s_x^2}\right\}\sigma_e^2 \qquad \text{(C.17)}$$

C.2 BLUP for Small Area Population Total

We denote by $\tilde{T}_{y_i}$, the BLUP for the population total of Y for small area i as

$$\tilde{T}_{y_i} = \sum_{j\in s_i} y_j + \sum_{j\in r_i}(\hat{\beta}_0 + \hat{\beta}_1 x_j) = \sum_{j\in s_i} y_j + \sum_{j\in r_i}(\bar{y}_s - \hat{\beta}_1\bar{x}_s + \hat{\beta}_1 x_j)$$

$$= \sum_{j\in s_i} y_j + (N_i - n_i)\bar{y}_s + (N_i - n_i)(\bar{x}_{r_i} - \bar{x}_s)\hat{\beta}_1$$

$$= \sum_{j\in s}\delta_{ij} y_j + (N_i - n_i)\left[\frac{1}{n}\sum_{j\in s} y_j + \frac{\sum_{j\in s}(\bar{x}_{r_i} - \bar{x}_s)(x_j - \bar{x}_s)y_j}{(n-1)s_x^2}\right]$$

$$= \sum_{j \in s} \left\{ \delta_{ij} + (N_i - n_i) \left[\frac{1}{n} + \frac{(\bar{x}_{r_i} - \bar{x}_s)(x_j - \bar{x}_s)}{(n-1)s_x^2} \right] \right\} y_j \ ,$$

where $\delta_{ij} = \begin{cases} 1 & if \ \ j \in i \\ 0 & otherwise \end{cases}$. Then

$$w_{j,BLUP}^{(i)} = \delta_{ij} + \frac{(N_i - n_i)}{n} + (N_i - n_i) \frac{(\bar{x}_{r_i} - \bar{x}_s)}{(n-1)s_x^2} (x_j - \bar{x}_s)$$

$$= \left(\delta_{ij} - \frac{n_i}{n} \right) + \frac{N_i}{n} \left[1 + \left(1 - \frac{n_i}{N_i} \right) \left(\frac{n}{n-1} \right) \frac{(\bar{x}_{r_i} - \bar{x}_s)}{s_x^2} (x_j - \bar{x}_s) \right]$$

$$= \left(\delta_{ij} - \frac{n_i}{n} \right) + \frac{N_i}{n} (1 + g_{ij}) , \qquad \text{(C.18)}$$

where $g_{ij} = \left(1 - \frac{n_i}{N_i} \right) \left(\frac{n}{n-1} \right) \frac{(\bar{x}_{r_i} - \bar{x}_s)}{s_x^2} (x_j - \bar{x}_s)$.

We observe that $w_j \sim O\left(\frac{N}{n} \right)$ given in (C.4) while $w_{j,BLUP}^{(i)} \sim O\left(\frac{N_i}{n} \right)$ given in (C.18), so we expect the BLUP to be more efficient.

$$w_{j,BLUP}^{(i)} = \begin{cases} \left(1 - \frac{n_i}{n} \right) + \frac{N_i}{n} (1 + g_{ij}), & j \in i \\ -\frac{n_i}{n} + \frac{N_i}{n} (1 + g_{ij}), & j \notin i \end{cases}$$

Under what conditions are the MBD and BLUP 'close'?

$$\sum_{j \in s_i} w_j y_j - \sum_{j \in s} w_{j,BLUP}^{(i)} y_j$$

$$= \sum_{j \in s_j} \{ w_j - w_{j,BLUP}^{(i)} \} y_j - \sum_{s - s_i} w_{j,BLUP}^{(i)} y_j \qquad \text{(C.19)}$$

Suppose that $\bar{x}_{r_i} \equiv \bar{x}_r$. Then, to first order $g_{ij} = g_j$ and, to the same order of approximation, (C.19) equals

$$\sum_{s_i}\left\{\frac{N}{n}(1+g_j)-\frac{N_i}{n}(1+g_j)\right\}y_j-\sum_{s-s_i}\frac{N_i}{n}(1+g_j)y_j$$

$$=\frac{N-N_i}{n}\sum_{s_i}(1+g_j)y_j-\frac{N_i}{n}\sum_{s-s_i}(1+g_j)y_j$$

$$=\left(\frac{N-N_i}{n}\right)n_i\left[\sum_{s_i}\frac{1}{n_i}(1+g_j)y_j\right]-\frac{N_i}{n}(n-n_i)\left[\sum_{s-s_i}\frac{1}{n-n_i}(1+g_j)y_j\right] \qquad \text{(C.20)}$$

Suppose the two averages in square brackets in (C.20) are same, equal to A, say. Then (C.20) reduces to

$$\left\{\left(\frac{N-N_i}{n}\right)n_i-\frac{N_i}{n}(n-n_i)\right\}A=\left[\frac{N}{n}n_i-N_i\right]A=N\left[\frac{n_i}{n}-\frac{N_i}{N}\right]A.$$

Hence sufficient conditions for equivalence of the MBD and BLUP are

i. $\bar{x}_{r_i}=\bar{x}_r$

ii. $\frac{n_i}{n}=\frac{N_i}{N}$

iii. $\frac{1}{n_i}\sum_{s_i}(1+g_j)y_j=\frac{1}{n-n_i}\sum_{s-s_i}(1+g_j)y_j$.

APPENDIX D

EFFICIENCY OF BLUP AND DIRECT ESTIMATORS FOR SMALL AREA ESTIMATION

The efficiency of the BLUP and direct estimator for the population total of Y in small area i, T_{y_i} is studies via empirical example using AAGIS data described in chapter 3. Besides the MBD estimator, we also considered two design-based direct estimators for the population total of Y for small area i, defined as

$$\hat{T}_{y_i}^{DBD1} = \sum_{j \in s_i} w_j y_j = (N_i / n_i)\left(\sum_{j \in s_i} y_j\right), \tag{D.1}$$

$$\hat{T}_{y_i}^{DBD2} = \sum_{j \in s_i} w_j y_j = (N / n)\left(\sum_{j \in s_i} y_j\right). \tag{D.2}$$

These are expansion type estimators defined by area specific weights $w_{ij} = N_i / n_i$ and population level weights $w_{ij} = N / n$, denoted by DBD1 and DBD2 respectively, see Rao (2003), page 19.

In our empirical study under the random intercept model (model I in chapter 3), we fix the values for parameter β and σ_e^2 (obtained from original sample of AAGIS data), and then choose different values for the σ_u^2. Table D.1 shows the average of ratios of the mean squared error (MSE) between DBD and MBD and between MBD and BLUP. These results in Table D.1 and Figure D.1 indicate between the DBD and MBD methods,

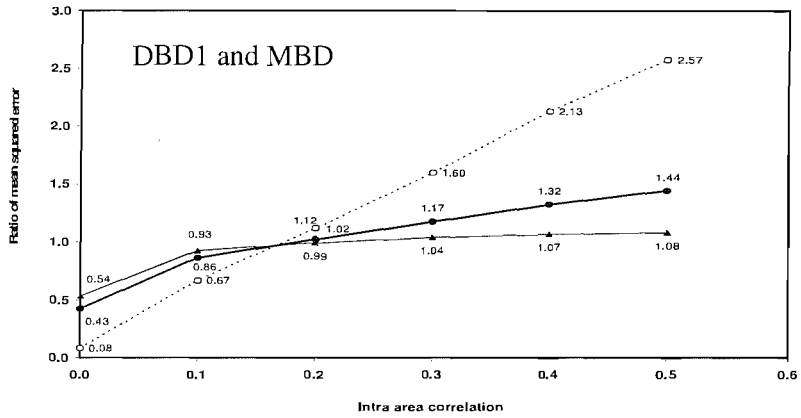
efficiency of the MBD increases with area effect. The MBD incorporates area effect while DBD does not. The gains due to the MBD are more significant when area specific sample sizes are smaller. At population level too, the MBD is consistently more efficient than the DBD (Figure D.2). In general, the BLUP is efficient than the MBD if model holds. The BLUP and MBD have equivalent performance if either area effect or small area sample sizes or both are large. The MBD provides an improvement over design-based methods and competes with BLUP.

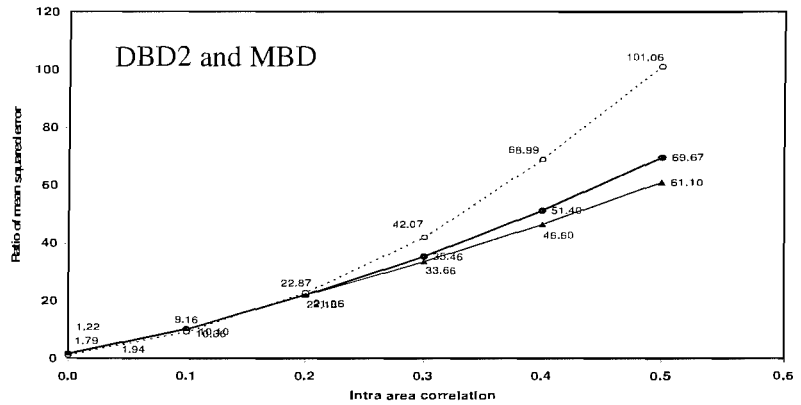
Table D.1 Ratio of mean squared errors (MSEs).

Ratio of MSE	Averaged over areas	Intra area effect					
		0.0	0.1	0.2	0.3	0.4	0.5
DBD1/MBD	29 areas	0.43	0.86	1.02	1.17	1.32	1.44
	7 areas ($n_i \leq 30$)	0.08	0.67	1.12	1.6	2.13	2.57
	22 areas ($n_i > 30$)	0.54	0.93	0.99	1.04	1.07	1.08
	Population	1.32	1.05	1.04	1.03	1.03	1.03
DBD2*/MBD	28 areas	1.79	10.10	22.15	35.46	51.40	69.67
	6 areas ($n_i \leq 30$)	1.22	9.16	22.87	42.07	68.99	101.06
	22 areas ($n_i > 30$)	1.94	10.36	21.96	33.66	46.60	61.10
	Population	17.48	15.78	17.76	20.33	24.02	29.06
MBD/BLUP	29 areas		5.14	2.44	1.71	1.41	1.29
	7 areas ($n_i \leq 30$)		14.76	5.74	3.29	2.31	1.92
	22 areas ($n_i > 30$)		2.08	1.39	1.2	1.13	1.09

* One area is dropped due to high value in of DBD2

Figure D.1 Average ratio of MSEs. Averaged over 29 regions (solid line), 7 regions with sample size less or equal to 30 (dashed line) and 22 regions with sample size greater than 30 (thin line). For DBD2 only 28 regions are taken.





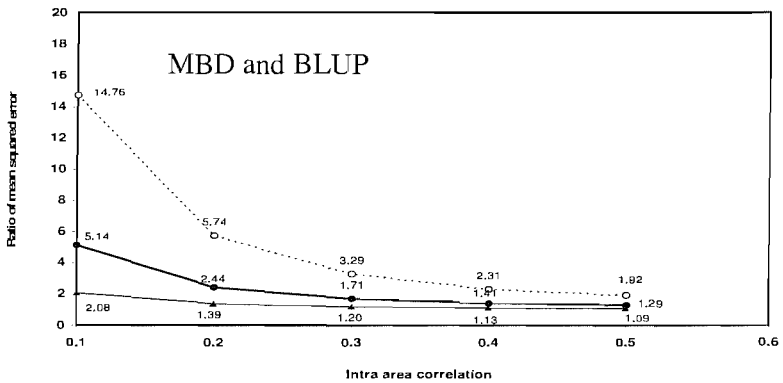
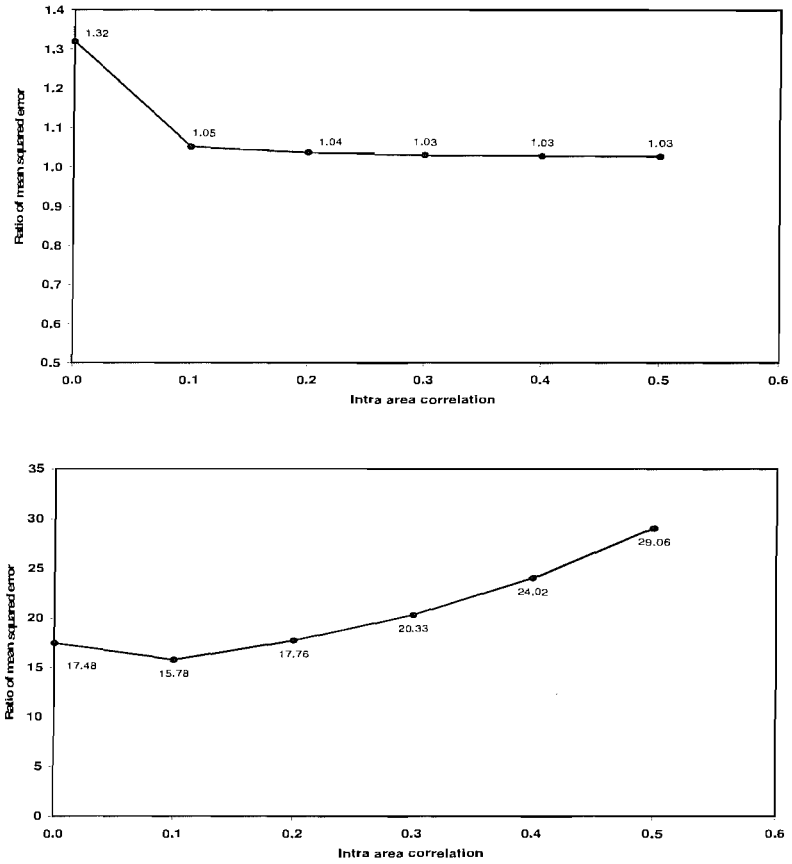


Figure D.2 Ratio of MSE of DBD1 and MBD (up) and MSE of DBD2 and MBD (down) at population level.



APPENDIX E

AN APPLICATION OF MBD METHOD OF SMALL AREA ESTIMATION TO THE BINARY VARIABLES

In chapter 3 and 4 we noticed that MBD estimators, whether they are based on variable-specific weights or multipurpose weights, are effectively linear estimators, and implicitly assume that variable of interest follows a linear mixed model. For categorical survey variables, it is well known that the indirect estimation methods based on a generalized linear mixed model (GLMM) can be used (Rao, 2003). Therefore, it is interesting to see how much efficiency is lost if the MBD under the linear assumption is used in this case. We examine this issue via empirical studies using the AAGIS data. In AAGIS data we created a binary (0-1) variable, Zero Debt, which takes value 1 if Debt (the response variable Farm Debt) is zero for the given farm and value 0 otherwise.

E.1 Small Area Estimation under Generalized Linear Mixed Models

Many often variables of interest in small area estimation (SAE) are not normally distributed, and therefore cannot be adequately modelled via linear mixed model. In such cases an appropriate model is GLMM. Under this type of model, distribution of the

values of the variable of interest Y is assumed to depend on η that is related to regression covariates and random component through the model of form

$$\eta_{ij} = g(\pi_{ij}) = x'_{ij}\beta + Z_{ij}u_i \quad (j = 1,....,n_{ij}; i = 1,....,m). \qquad \text{(E.1)}$$

Here notation used is similar to one described in chapter 3, see Saei and Chambers (2003). Under a random intercept specification of model (E.1), $\eta_{ij} = g(\pi_{ij}) = x'_{ij}\beta + u_i$. The linear predictor η_{ij} is connected to y_{ij} via a known function h (inverse of g) as $E(y_{ij} \mid u_i) = \pi_{ij} = h(\eta_{ij})$. This is the expectation of the conditional distribution of the outcome given the random effects. The predicted values of y_{ij} are given as $\hat{y}_{ij} = h(\hat{\eta}_{ij})$ with $\hat{\eta}_{ij} = x_{ij}\hat{\beta} + \hat{u}_i$. For a binary variable, the function $g(.)$ is logit or logistic function of the probability π_i that a population unit j in area i is a "success". In other words, $\pi_{ij} = \Pr(y_{ij} = 1)$, $j = 1,....., n_i; i = 1,..., m$. The empirical best predictor for population mean of Y for small area i (denoted by EBP) is

$$\hat{\bar{Y}}_i^{EBP} = N_i^{-1}(\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}) = N_i^{-1}\left(\sum_{s_i} y_{ij} + \sum_{r_i} \frac{\exp(x'_{ij}\hat{\beta} + \hat{u}_i)}{1 + \exp(x'_{ij}\hat{\beta} + \hat{u}_i)}\right). \qquad \text{(E.2)}$$

In empirical evaluation we consider four different types of estimators:

i) the EBLUP (3.20) under linear mixed model, denoted by EBLUP

ii) the empirical best predictor (E.2), denoted by EBP

iii) the MBD (3.15) based on variable specific weights for Zero Debt under the linear mixed model, denoted by MBD

iv) the MBD (3.15) based on multipurpose weights under the linear mixed model, denoted by MBD.MP.

The multipurpose weights used in MBD.MP estimator are based on five variables (TCC, TCR, FCI, Cattle and Sheep) other than Zero Debt from AAGIS data. See chapter 4. For EBLUP, MBD and MBD.MP, we followed the procedure described in chapter 3 and 4. However, EBP (E.2) we fit generalized linear mixed model via penalized quasi-likelihood (PQL) using 'glmmPQL' function in MASS library in R. See http://www.r-project.org. Results from the design-based simulation studies are reported in Table E.1.

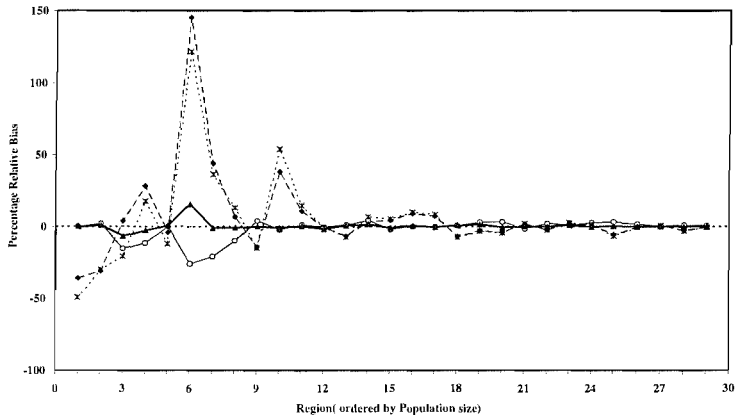
These results show the average relative biases of MBD.MP and average RRMSE of MBD are smaller overall. The MBD (MBD and MBD.MP) method is performing well. In this case EBLUP under the linear mixed model is ill-suited. Overall we do not observe any efficiency loss if the MBD based on the linear assumption is used. Figure E.1 shows the regional performances generated by these methods. We notice relatively better performance of MBD approach in regional estimation. In few regions, both EBLUP and EBP are very unstable. In particular, in two regions (1 and 6) both EBLUP and EBP produce unstable results, inspection of the population and sample data indicated that this is because of a few outlying estimates. In region 1 with sample size 6 there is one zero and rest (5 observations) 1's, and in population of 79 there are 15 zeros and rest (64 observations are) 1's. Further, out of 1000 samples there are 9 samples with no zeros and 16 sample only one zeros, this generated the number of outlying estimates. Similarly in region 6 of size 19, there are 13 zero and rest 1's in original sample. In population 465, there are 407 zero, which created lot of outlying estimates. However, MBD being direct estimator is still working well in such cases.

Table E.1 Average (ARB) relative biases (%) and average (ARRMSE) relative root mean squared errors (%) generated by different methods under the random intercept specification of mixed model for Zero Debt. The average is over 29 small areas.

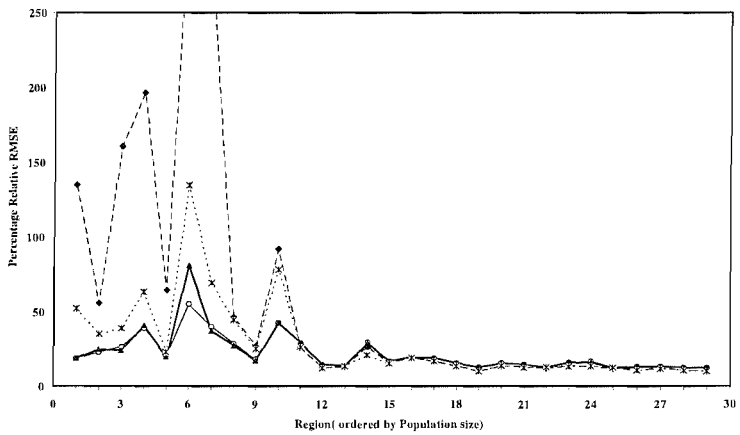
Criteria	EBLUP	EBP	MBD	MBD.MP
ARB	6.56	4.57	-1.92	0.29
ARRMSE	57.59	29.02	21.77	22.36

Figure E.1 Regional performances of EBLUP (dashed line), EBP (dotted line), MBD (thin line) and MBD.MP (thick line) for Zero Debt.

Relative Bias



Relative RMSE



APPENDIX F

ESTIMATES FOR β USED IN MSE ESTIMATION FOR THE MBD METHODS IN CHAPTER 4

The estimate of β used in the MSE estimate (4.18) under the multipurpose weighting methods is evaluated as below.

1. When multipurpose weights defined by the first approach, i.e. via (4.8) or (4.12), there can be three possible options for using $\hat{\beta}$:
   a) Use variable specific estimate, $\hat{\beta}_k$
   b) Use weighted average of variable-specific β estimates, $\hat{\beta} = \sum_k \phi_k \hat{\beta}_k$
   c) Use estimate of β evaluated from the weighted average of variances used in deriving the sample weights (4.8) or (4.12).

   Empirical results indicate use of option (a) or (c) does not make any substantial difference. However, option (b) seems to be less appropriate.
2. When the multipurpose weights defined by the second approach via (4.17), there can be two possible ways (a) and (b) to calculate the estimate of β. Our results show method (a) is more appropriate in this case.

In chapter 4, the design-based simulation studies use option (a). However, results with option (b) and (c) are illustrated below. These results (Table F.1 and F.2) hardly show any difference in the performance of the MBD method by using option (a), (b) or (c).

Table F.1 Average coverage rate (ACR) and average interval width (AW) for five variables best suited to the linear mixed modelling under MBD1-A and with equal relative weights ($\phi_k = 1/K$) to all variables.

Model	Criteria	Options	TCC	TCR	FCI	Cattle	Sheep
I	ACR	a	0.92	0.92	0.94	0.95	0.96
		b	0.91	0.9	0.97	1.00	1.00
		c	0.92	0.92	0.94	0.95	0.96
	AW	a	142754	186619	91641	1622	2705
		b	131848	160107	112934	85824	85664
		c	143250	186502	90742	1622	2689
II	ACR	a	0.93	0.93	0.94	0.95	0.96
		b	0.92	0.91	0.98	1.00	1.00
		c	0.93	0.93	0.95	0.95	0.96
	AW	a	173451	238140	111223	1873	4014
		b	136333	166732	138622	118127	117962
		c	174943	238456	113517	1877	3669

Table F.2 Average coverage rate (ACR) and average interval width for five variables best suited to the linear mixed modelling under MBD2 method and with equal relative weights ($\phi_k = 1/K$) to all variables.

Model	Criteria	Option	TCC	TCR	FCI	Cattle	Sheep
I	ACR	a	0.92	0.92	0.94	0.95	0.96
		b	0.91	0.91	0.97	1.00	1.00
	AW	a	142328	186212	91950	1617	2713
		b	132132	160943	113217	85868	85724
II	ACR	a	0.93	0.93	0.94	0.95	0.96
		b	0.92	0.92	0.98	1.00	1.00
	AW	a	184598	254392	116304	2010	4325
		b	142649	175450	145350	128111	128111

APPENDIX G

THE METHOD OF MOMENT ESTIMATION USED IN CHAPTER 4

We describe the method of moment estimation (also called Henderson's III method) for the variance components of the model (2.13). Here various terms used are defined similar to as below equation (2.13). A simple method of estimating σ_e^2 and σ_u^2 involves performing two ordinary least squares (OLS) regression and then using the method of moments to get unbiased estimators of σ_e^2 and σ_u^2. An unbiased estimator of σ_e^2 (using 'hat' to denote an estimate) is

$$\hat{\sigma}_e^2 = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\varepsilon}_{ij}^2 \qquad \text{(G.1)}$$

where $\{\hat{\varepsilon}_{ij}\}$ are the residuals from the OLS regression of $\tilde{y}_{ij} = (y_{ij} - \bar{y}_i)$ on $\tilde{x}_{ij} = (x_{ij} - \bar{x}_i)$, with $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{x}_i = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$ are the sample means in the small area *i*. That is $\hat{\varepsilon}_{ij} = (\tilde{y}_{ij} - \tilde{x}_{ij}'\hat{\beta}_{OLS}) = (y_{ij} - \bar{y}_i) - (x_{ij} - \bar{x}_i)'\hat{\beta}_{OLS}$. Here $\hat{\beta}_{OLS}$ is the OLS estimate of $\hat{\beta}$ and $D = (n - m - p + a)$ with $a = 0$ if the model (2.13) has no intercept term in fixed component of the model and $a = 1$ otherwise.

An unbiased estimator of σ_u^2 is

$$\hat{\sigma}_u^2 = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{e}_{ij}^2 - (n-p)\hat{\sigma}_e^2\right] = n_*^{-1}\left[\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{x}_i'\hat{\beta}_{OLS}) - (n-p)\hat{\sigma}_e^2\right] \qquad \text{(G.2)}$$

where $n_* = n - tr\left[(X'X)^{-1}\sum_{i=1}^{m} n_i^2 \bar{x}_i \bar{x}_i'\right] = \sum_i n_i \left\{1 - n_i \bar{x}_i \left(\sum_i \sum_j n_{ij} x_{ij} x_{ij}'\right)^{-1} \bar{x}_i'\right\}$ is function of x_{ij} with $X' = (x_1, .., x_m)$; $X'X = \sum_i x_i x_i' = \sum_i \sum_j x_{ij} x_{ij}'$ and $\{\hat{e}_{ij}\}$ are the residuals from the OLS regression of y_{ij} on x_{ij}, i.e. $\hat{e}_{ij} = y_{ij} - x_{ij}'\hat{\beta}_{OLS}$. Here $\hat{\sigma}_u^2$ can also take negative values so a truncated estimator of σ_u^2 is obtained as $\hat{\sigma}_u^2 = \max(0, \hat{\sigma}_u^2)$. Note that $\hat{\sigma}_u^2$ is no longer unbiased, but it is consistent as *m*, the number of small areas, increases. The estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ are equivalent to those found by using the well-known method of fitting of constants (F-C) due to Henderson (1953). The moment estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ are, therefore, also referred to as fitting-of-constants (MFC) estimators (Prasad and Rao, 1990).

Once σ_e^2 and σ_u^2 are estimated, then we can also get an improve estimates of β using an iterative generalized least squares (IGLS) method to estimate the fixed regression parameter β and the variance components (Goldstein, 1995). The IGLS method involves two applications of the generalized least square (GLS). The first step is to obtain the GLS estimate of β assuming σ_e^2 and σ_u^2 known. The second step is to use the GLS estimate of β to form the "raw" residuals. Then the estimation of σ_e^2 and σ_u^2 involves an application of GLS on the vector form of the cross-product matrix of the residuals, assuming normality. The IGLS method involves iterative updating between the GLS estimate of β and the GLS estimates of σ_e^2 and σ_u^2 until the procedure converges. We do not pursue this iterative procedure in this thesis.

When we have K variables, we introduce an extra subscript k ($k = 1,..,K$) to denote quantities with respect to the k^{th} variable. For any two variables, above method leads to following estimates:

$$\hat{\sigma}^2_{e1} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\varepsilon}^2_{1,ij} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\tilde{y}_{1,ij} - \tilde{x}'_{ij}\hat{\beta}_{1,OLS})^2$$

$$\hat{\sigma}^2_{e2} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\varepsilon}^2_{2,ij} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\tilde{y}_{2,ij} - \tilde{x}'_{ij}\hat{\beta}_{2,OLS})^2$$

$$\hat{\sigma}^2_{e12} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\varepsilon}_{1,ij}\hat{\varepsilon}_{2,ij} = D^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\tilde{y}_{1,ij} - \tilde{x}'_{ij}\hat{\beta}_{1,OLS})(\tilde{y}_{2,ij} - \tilde{x}'_{ij}\hat{\beta}_{2,OLS}) = \hat{\sigma}^2_{e21}$$

$$\hat{\sigma}^2_{u1} = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{e}^2_{1,ij} - (n-p)\hat{\sigma}^2_{e1}\right] = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{1,ij} - x'_{ij}\hat{\beta}_{1,OLS})^2 - (n-p)\hat{\sigma}^2_{e1}\right]$$

$$\hat{\sigma}^2_{u1} = \max(0, \hat{\sigma}^2_{u1})$$

$$\hat{\sigma}^2_{u2} = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{e}^2_{2,ij} - (n-p)\hat{\sigma}^2_{e2}\right] = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{2,ij} - x'_{ij}\hat{\beta}_{2,OLS})^2 - (n-p)\hat{\sigma}^2_{e2}\right]$$

$$\hat{\sigma}^2_{u2} = \max(0, \hat{\sigma}^2_{u2})$$

$$\hat{\sigma}_{u12} = n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{e}_{1,ij}\hat{e}_{2,ij} - (n-p)\hat{\sigma}^2_{e12}\right]$$

$$= n_*^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{1,ij} - x'_{ij}\hat{\beta}_{1,OLS})(y_{2,ij} - x'_{ij}\hat{\beta}_{2,OLS}) - (n-p)\hat{\sigma}^2_{e12}\right] = \hat{\sigma}_{u21}$$

$\hat{\sigma}_{u12} = 0$ if either $\hat{\sigma}^2_{u1} = 0 = \hat{\sigma}^2_{u2}$ or either of them is zero.

APPENDIX H

MODEL-BASED SIMULATIONS FOR MULTIPURPOSE SMALL AREA ESTIMATION IN CHAPTER 4

In the model-based simulations we choose a population of size $N = 15{,}000$ and then randomly generated small area population sizes N_i, $i = 1,...,m = 30$, so that $\sum_i N_i = N$. We consider $n =$ 600 and then generated small area sample sizes as $n_i = N_i(n/N)$ so that $\sum_i n_i = n$ and kept fixed throughout the simulations. We generated a multivariate normal (MVN) population for K =2 response variables. Two response variables y_1 and y_2 are generated under a multivariate linear mixed model of form

$$y_{1,ij} = \alpha_0 + \alpha_1 x_{ij} + u_{1,i} + e_{1,ij} \text{ and } y_{2,ij} = \beta_0 + \beta_1 x_{ij} + u_{2,i} + e_{2,ij}.$$

We fixed $\alpha_0 = 5, \alpha_1 = 1, \beta_0 = 5$ and $\beta_1 = 3$. The covariate values x_{ij} are generated from $\chi^2(50)$ distribution. The random area effects $u_{1,i}$ and $u_{2,i}$ are generated from a MVN with zero mean vector and covariance $\Sigma_u = \begin{pmatrix} \Sigma_{u,1} & \Sigma_{u,12} \\ \Sigma_{u,21} & \Sigma_{u,2} \end{pmatrix}$. That is $\begin{pmatrix} u_{1,i} \\ u_{2,i} \end{pmatrix} \sim MVN_2(0, \Sigma_u)$ with between area correlation $\rho_{u,12} = \Sigma_{u,12} / (\sqrt{\Sigma_{u,1}} \sqrt{\Sigma_{u,2}}) = \rho_{u,21}$. The individual random errors $e_{1,ij}$ and $e_{2,ij}$ generated from $\begin{pmatrix} e_{1,ij} \\ e_{2,ij} \end{pmatrix} \sim MVN_2(0, \Sigma_e)$, where $\Sigma_e = \begin{pmatrix} \sigma^2_{e,1} & \sigma_{e,12} \\ \sigma_{e,21} & \sigma^2_{e,2} \end{pmatrix}$, with $\rho_{e,12} = \sigma_{e,12} / (\sqrt{\sigma^2_{e,1}} \sqrt{\sigma^2_{e,2}}) = \rho_{e,21}$. We choose seven different values of $\rho_{u,12}$ and $\rho_{e,12}$

corresponding to set1-set7. The values of $\Sigma_{u,1}$, $\Sigma_{u,2}$ and $\Sigma_{u,12}$ as well as $\sigma^2_{e,1}, \sigma^2_{e,2}$ and $\sigma_{e,12}$ are fixed up so that intra-area correlations with respect to first and second variables are $\left(\Sigma_{u,1}/(\Sigma_{u,1}+\sigma^2_{e,1})\right) \approx 0.20$ and $\left(\Sigma_{u,2}/(\Sigma_{u,2}+\sigma^2_{e,2})\right) \approx 0.10$ respectively. Table H.1 sets out the results from this simulation study. These results indicate identical performance of MBD1-A and MBD1-B methods of SAE.

Table H.1 Performance measures generated by MBD0, MBD1-A and MBD1-B (see chapter 4) for two variables under model I. Method of Moment estimate for variance components are used. All averages are over the 30 small areas.

Variables	Criterions		Set1	Set2	Set3	Set4	Set5	Set6	Set7
		$\rho_{u,12}$	0.50	0.50	0.50	0.50	0.00	0.25	0.75
		$\rho_{e,12}$	0.00	0.25	0.50	0.75	0.50	0.50	0.50
1	MBD0	ARB	0.00	0.01	0.00	0.00	0.00	0.00	-0.01
		ARMSE	4.10	4.09	4.09	4.10	4.11	4.12	4.10
		ACR	0.96	0.97	0.96	0.97	0.97	0.96	0.96
	MBD1-A	ARB	0.00	0.01	0.00	0.00	0.00	0.00	-0.01
		ARMSE	4.10	4.09	4.09	4.10	4.11	4.12	4.10
		ACR	0.96	0.97	0.97	0.97	0.97	0.96	0.96
	MBD1-B	ARB	0.00	0.01	0.00	0.00	0.00	0.00	-0.01
		ARMSE	4.10	4.09	4.09	4.10	4.10	4.12	4.10
		ACR	0.96	0.97	0.97	0.97	0.97	0.96	0.96
2	MBD0	ARB	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		ARMSE	4.22	4.21	4.20	4.22	4.22	4.23	4.21
		ACR	0.75	0.76	0.76	0.75	0.75	0.75	0.75
	MBD1-A	ARB	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		ARMSE	4.22	4.21	4.20	4.22	4.22	4.23	4.21
		ACR	0.94	0.94	0.94	0.93	0.93	0.94	0.94
	MBD1-B	ARB	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		ARMSE	4.22	4.21	4.20	4.22	4.22	4.23	4.21
		ACR	0.94	0.94	0.94	0.93	0.94	0.94	0.94

APPENDIX I

COVARIANCE MATRIX UNDER RANDOM SLOPE SPECIFICATION OF MODEL (5.9) IN CHAPTER 5

Under model (5.9) the covariance matrix of $l_i = \log(Y_i)$ is

$$V_i = \sigma_e^2 I_{N_i} + G_i \Sigma(\theta) G_i' \text{ with } v_{ijk} = \sigma_e^2 I(j=k) + G_{ij}' \Sigma(\theta) G_{ij}.$$

The covariance matrix of Y_i given by (5.15) is $Var(Y_i) = \Omega_i = [\omega_{ijk}]$ with

$$\omega_{ijk} = \begin{cases} e^{(W_{ij}+W_{ik})'\beta}[e^{\frac{1}{2}(v_{ijj}+v_{ikk})}(e^{v_{ijk}}-1)] & if \quad j \neq k \\ e^{2W_{ij}'\beta}[e^{v_{ijj}}(e^{v_{ijj}}-1)] & if \quad j = k \end{cases}$$

$$= \begin{cases} e^{(W_{ij}+W_{ik})'\beta}[e^{\frac{1}{2}(\sigma_e^2+G_{ij}'\Sigma(\theta)G_{ij}+\sigma_e^2+G_{ik}'\Sigma(\theta)G_{ik})}(e^{G_{ij}'\Sigma(\theta)G_{ik}}-1)] & if \quad j \neq k \\ e^{2W_{ij}'\beta}[e^{\sigma_e^2+G_{ij}'\Sigma(\theta)G_{ij}}(e^{\sigma_e^2+G_{ij}'\Sigma(\theta)G_{ij}}-1)] & if \quad j = k. \end{cases}$$

We can rewrite

$\Omega_i = e^{\sigma_e^2}[E_i \Delta_i E_i']$ with $E_i = diag\{e^{W_{ij}'\beta}; 1 \le j \le N_i\}$ and

$$\Delta_i = \Delta_{i1} - \Delta_{i2} = \exp\left\{\sigma_e^2 I_{N_i} + \frac{1}{2}\Delta_{i1}^{(1)} + \frac{1}{2}\Delta_{i1}^{(2)} + \Delta_{i1}^{(3)}\right\} - \exp\left\{\frac{1}{2}\Delta_{i1}^{(1)} + \frac{1}{2}\Delta_{i1}^{(2)}\right\}, \text{ where}$$

$$\Delta_{i1} = \begin{bmatrix} e^{\sigma_e^2+2G_{i1}'\Sigma G_{i1}} & e^{\frac{1}{2}(G_{i2}'\Sigma G_{i2}+G_{i1}'\Sigma G_{i1}+2G_{i1}'\Sigma G_{i2})} & \cdots & e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1}+G_{iN_i}'\Sigma G_{iN_i}+2G_{i1}'\Sigma G_{iN_i}')} \\ e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1}+G_{i2}'\Sigma G_{i2}+2G_{i1}'\Sigma G_{i2})} & e^{\sigma_e^2+2G_{i2}'\Sigma G_{i2}} & & \\ \cdots & \cdots & & \cdots \\ e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1}+G_{iN_i}'\Sigma G_{iN_i}+2G_{i1}'\Sigma G_{iN_i})} & e^{\frac{1}{2}(G_{i2}'\Sigma G_{i2}+G_{iN_i}'\Sigma G_{iN_i}+2G_{i1}'\Sigma G_{iN_i})} & \cdots & e^{\sigma_e^2+2G_{iN_i}'\Sigma G_{iN_i}} \end{bmatrix}$$

$$= \exp\left\{\sigma_e^2 I_{N_i} + \frac{1}{2}\Delta_{i1}^{(1)} + \frac{1}{2}\Delta_{i1}^{(2)} + \Delta_{i1}^{(3)}\right\} \text{ with}$$

$$\Delta_{i1}^{(1)} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} & G_{i1}'\Sigma G_{i1} \ldots G_{i1}'\Sigma G_{i1} \\ G_{i2}'\Sigma G_{i2} & G_{i2}'\Sigma G_{i2} \ldots G_{i2}'\Sigma G_{i2} \\ \ldots\ldots \\ G_{iN_i}'\Sigma G_{iN_i} & G_{iN_i}'\Sigma G_{iN_i} \ldots G_{iN_i}'\Sigma G_{iN_i} \end{bmatrix} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} \\ G_{i2}'\Sigma G_{i2} \\ \ldots\ldots \\ G_{iN_i}'\Sigma G_{iN_i} \end{bmatrix} 1_{N_i}' \; ;$$

$$\Delta_{i1}^{(2)} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} & G_{i2}'\Sigma G_{i2} \ldots G_{iN_i}'\Sigma G_{iN_i} \\ G_{i1}'\Sigma G_{i1} & G_{i2}'\Sigma G_{i2} \ldots G_{iN_i}'\Sigma G_{iN_i} \\ \ldots\ldots \\ G_{i1}'\Sigma G_{i1} & G_{i2}'\Sigma G_{i2} \ldots G_{iN_i}'\Sigma G_{iN_i} \end{bmatrix} = 1_{N_i}\left[G_{i1}'\Sigma G_{i1} \;\; G_{i2}'\Sigma G_{i2} \ldots G_{iN_i}'\Sigma G_{iN_i}\right]'$$

$$\Delta_{i1}^{(3)} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} & G_{i1}'\Sigma G_{i2} \ldots G_{i1}'\Sigma G_{iN_i} \\ G_{i2}'\Sigma G_{i1} & G_{i2}'\Sigma G_{i2} \ldots G_{i2}'\Sigma G_{iN_i} \\ \ldots\ldots \\ G_{iN_i}'\Sigma G_{i1} & G_{iN_i}'\Sigma G_{i2} \ldots G_{iN_i}'\Sigma G_{iN_i} \end{bmatrix} = G_i \Sigma G_i'$$

and

$$\Delta_{i2} = \begin{bmatrix} e^{G_{i1}'\Sigma G_{i1}} & e^{\frac{1}{2}(G_{i2}'\Sigma G_{i2} + G_{i1}'\Sigma G_{i1})} & \ldots\; e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1} + G_{iN_i}'\Sigma G_{iN_i})} \\ e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1} + G_{i2}'\Sigma G_{i2})} & e^{G_{i2}'\Sigma G_{i2}} & \\ \ldots\ldots & & \\ e^{\frac{1}{2}(G_{i1}'\Sigma G_{i1} + G_{iN_i}'\Sigma G_{iN_i})} & e^{\frac{1}{2}(G_{i2}'\Sigma G_{i2} + G_{iN_i}'\Sigma G_{iN_i})} & \ldots\; e^{G_{iN_i}'\Sigma G_{iN_i}} \end{bmatrix} = \exp\left\{\frac{1}{2}\Delta_{i1}^{(1)} + \frac{1}{2}\Delta_{i1}^{(2)}\right\}.$$

We consider the sample and non-sample partition of covariance matrix as

$$\Omega_i = e^{\sigma_e^2}\left[E_i \Delta_i E_i'\right] = \begin{bmatrix} \Omega_{iss} & \Omega_{isr} \\ \Omega_{irs} & \Omega_{irr} \end{bmatrix} = e^{\sigma_e^2}\left\{\begin{bmatrix} E_{iss} & 0 \\ 0 & E_{irr} \end{bmatrix}\begin{bmatrix} \Delta_{iss} & \Delta_{isr} \\ \Delta_{irs} & \Delta_{irr} \end{bmatrix}\begin{bmatrix} E_{iss} & 0 \\ 0 & E_{irr} \end{bmatrix}'\right\}$$

$$= e^{\sigma_e^2}\begin{bmatrix} E_{iss}\Delta_{iss}E_{iss}' & E_{iss}\Delta_{isr}E_{irr}' \\ E_{irr}\Delta_{irs}E_{iss}' & E_{irr}\Delta_{irr}E_{irr}' \end{bmatrix} \text{ so that}$$

$$\Omega_{iss} = e^{\sigma_e^2}\left[E_{iss}\Delta_{iss}E_{iss}'\right] \text{ with } E_{iss} = diag\left\{e^{W_{ij}\beta}; 1 \le j \le n_i\right\} \text{and } \Delta_{iss} = (\Delta_{iss1} - \Delta_{iss2}).$$

Here $\Delta_{iss1} = \exp\left\{\sigma_e^2 I_{n_i} + \frac{1}{2}\Delta_{iss1}^{(1)} + \frac{1}{2}\Delta_{iss1}^{(2)} + \Delta_{iss1}^{(3)}\right\}$ and $\Delta_{iss2} = \exp\left\{\frac{1}{2}\Delta_{iss1}^{(1)} + \frac{1}{2}\Delta_{iss1}^{(2)}\right\}$ with

$$\Delta_{iss1}^{(1)} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} \\ G_{i2}'\Sigma G_{i2} \\ ..... \\ G_{in_i}'\Sigma G_{in_i} \end{bmatrix} 1_{n_i}' \ , \ \Delta_{iss1}^{(2)} = 1_{n_i}\left[G_{i1}'\Sigma G_{i1} \ \ G_{i2}'\Sigma G_{i2} \ ...G_{in_i}'\Sigma G_{in_i}\right]' \text{ and } \Delta_{iss1}^{(3)} = G_{is}\Sigma G_{is}'.$$

Similarly,

$\Omega_{isr} = e^{\sigma_e^2}\left[E_{iss}\Delta_{isr}E_{irr}'\right]$ with $E_{irr} = diag\left\{e^{W_{ij}\beta}; n_i + 1 \le j \le N_i\right\}$ and $\Delta_{isr} = \Delta_{isr1} - \Delta_{isr2}$, where

$\Delta_{isr1} = \exp\left\{\frac{1}{2}\Delta_{isr1}^{(1)} + \frac{1}{2}\Delta_{isr1}^{(2)} + \Delta_{isr1}^{(3)}\right\}$ and $\Delta_{isr2} = \exp\left\{\frac{1}{2}\Delta_{isr1}^{(1)} + \frac{1}{2}\Delta_{isr1}^{(2)}\right\}$ with

$$\Delta_{isr1}^{(1)} = \begin{bmatrix} G_{i1}'\Sigma G_{i1} \\ G_{i2}'\Sigma G_{i2} \\ .... \\ G_{in_i}'\Sigma G_{in_i} \end{bmatrix} 1_{N_i - n_i}' \ , \ \Delta_{isr1}^{(2)} = 1_{n_i}\left[G_{in_i+1}'\Sigma G_{in_i+1} \ \ G_{in_i+2}'\Sigma G_{in_i+2} \ ...G_{iN_i}'\Sigma G_{iN_i}\right]', \ \Delta_{isr1}^{(3)} = G_{is}\Sigma G_{ir}'.$$

We assume that $Var(u_i) = \Sigma = \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix}$, then

$$\begin{bmatrix} G_{i1}\Sigma G_{i1}' \\ G_{i2}\Sigma G_{i2}' \\ .... \\ G_{in_i}\Sigma G_{in_i}' \end{bmatrix} = \begin{bmatrix} 1 & G_{i1} \\ 1 & G_{i2} \\ & .... \\ 1 & G_{in_i} \end{bmatrix} \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} 1 & G_{i1} \\ 1 & G_{i2} \\ & .... \\ 1 & G_{in_i} \end{bmatrix}' = \begin{bmatrix} \sigma_{u1}^2 + 2G_{i1}\sigma_{u12} + G_{i1}^2\sigma_{u2}^2 \\ \sigma_{u1}^2 + 2G_{i2}\sigma_{u12} + G_{i2}^2\sigma_{u2}^2 \\ .... \\ \sigma_{u1}^2 + 2G_{in_i}\sigma_{u12} + G_{in_i}^2\sigma_{u2}^2 \end{bmatrix}$$

$$= \left[\sigma_{u1}^2 \ 1_{n_i} + 2\begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix}\sigma_{u12} + \sigma_{u2}^2 \begin{pmatrix} G_{i1}^2 \\ G_{i2}^2 \\ .. \\ G_{in_i}^2 \end{pmatrix} \right].$$

This indicates that

$$\Delta_{isr1}^{(1)} = \left[\sigma_{u1}^2 \; 1_{n_i} + 2 \begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix} \sigma_{u12} + \sigma_{u2}^2 \begin{pmatrix} G_{i1}^2 \\ G_{i2}^2 \\ .. \\ G_{in_i}^2 \end{pmatrix} \right] 1'_{N_i - n_i} = \left[\sigma_{u1}^2 \; 1_{n_i} 1'_{N_i - n_i} + 2\sigma_{u12} \begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix} 1'_{N_i - n_i} + \sigma_{u2}^2 \begin{pmatrix} G_{i1}^2 \\ G_{i2}^2 \\ .. \\ G_{in_i}^2 \end{pmatrix} 1'_{N_i - n_i} \right]$$

$$\Delta_{isr1}^{(2)} = 1_{n_i} \left[\sigma_{u1}^2 \; 1_{N_i - n_i} + 2 \begin{pmatrix} G\gamma_{in_i+1} \\ G_{in_i+2} \\ .. \\ G_{iN_i} \end{pmatrix} \sigma_{u12} + \sigma_{u2}^2 \begin{pmatrix} G_{in_i+1}^2 \\ G_{in_i+2}^2 \\ .. \\ G_{iN_i}^2 \end{pmatrix}' \right] = \left[\sigma_{u1}^2 \; 1_{n_i} 1'_{N_i - n_i} + 2\sigma_{u12} 1_{n_i} \begin{pmatrix} G_{in_i+1} \\ G_{in_i+2} \\ .. \\ G_{iN_i} \end{pmatrix}' + \sigma_{u2}^2 1_{n_i} \begin{pmatrix} G_{in_i+1}^2 \\ G_{in_i+2}^2 \\ .. \\ G_{iN_i}^2 \end{pmatrix}' \right]$$

$$\Delta_{isr1}^{(3)} = G_{is} \Sigma G_{ir} = \begin{bmatrix} 1 & G_{i1} \\ 1 & G_{i2} \\ .... & \\ 1 & G_{in_i} \end{bmatrix} \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u12} \\ \sigma_{u21} & \sigma_{u2}^2 \end{bmatrix} \begin{bmatrix} 1 & G_{in_i+1} \\ 1 & G_{in_i+2} \\ & .... \\ 1 & G_{iN_i} \end{bmatrix}' = \left[1_{n_i} (\sigma_{u1}^2 \sigma_{u12})' + \begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix} (\sigma_{u21} \sigma_{u2}^2)' \right] \begin{bmatrix} 1 & G_{in_i+1} \\ 1 & G_{in_i+2} \\ & .... \\ 1 & G_{iN_i} \end{bmatrix}'$$

$$= 1_{n_i} \left\{ \sigma_{u1}^2 1'_{N_i - n_i} + \sigma_{u12} \begin{bmatrix} G_{in_i+1} \\ G_{in_i+2} \\ .... \\ G_{iN_i} \end{bmatrix}' \right\} + \begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix} \left\{ \sigma_{u21} 1'_{N_i - n_i} + \sigma_{u2}^2 \begin{bmatrix} G_{in_i+1} \\ G_{in_i+2} \\ .... \\ G_{iN_i} \end{bmatrix}' \right\}.$$

The last term in 'fitted value' model (5.6) based model calibration weights (5.7) is

$\Omega_{iss}^{-1} \Omega_{isr} 1_{N_i - n_i} = e^{\sigma_e^2} \Omega_{iss}^{-1} (E_{iss} \Delta_{isr} E'_{irr}) 1_{N_i - n_i} = e^{\sigma_e^2} \Omega_{iss}^{-1} E_{iss} \Delta_{isr} E_{ir}$. Here

$A_{ir} = E'_{irr} 1_{N_i - n_i} = (e^{W_{in_i+1}\beta} e^{W_{in_i+2}\beta} ... e^{W_{iN_i}\beta})'$ is a vector of order $(N_i - n_i)$. Consequently

$\Delta_{isr} E_{ir} = (\Delta_{isr1} - \Delta_{isr2}) E_{ir} = \Delta_{isr1} E_{ir} - \Delta_{isr2} E_{ir} = D_1 - D_2$. Note that

$$D_1 = \Delta_{isr1} E_{ir} = \left[\sigma_{u1}^2 \; 1_{n_i} 1'_{N_i - n_i} + 2\sigma_{u12} \begin{pmatrix} G_{i1} \\ G_{i2} \\ .. \\ G_{in_i} \end{pmatrix} 1'_{N_i - n_i} + \sigma_{u2}^2 \begin{pmatrix} G_{i1}^2 \\ G_{i2}^2 \\ .. \\ G_{in_i}^2 \end{pmatrix} 1'_{N_i - n_i} \right] E_{ir}$$

and

$$D_2 = \Delta_{isr2} E_{ir} = \left[\sigma_{u1}^2 \; 1_{n_i} 1'_{N_i - n_i} + 2\sigma_{u12} 1_{n_i} \begin{pmatrix} G_{in_i+1} \\ G_{in_i+2} \\ .. \\ G_{iN_i} \end{pmatrix}' + \sigma_{u2}^2 1_{n_i} \begin{pmatrix} G_{in_i+1}^2 \\ G_{in_i+2}^2 \\ .. \\ G_{iN_i}^2 \end{pmatrix}' \right] E_{ir}$$

are the vectors of order n_i and thus can be easily evaluated.

APPENDIX J

COVARIANCE MATRIX OF THE ESTIMATED VARIANCE COMPONENTS IN CHAPTER 5

We shall illustrate the expressions to obtain covariance matrix of the estimated variance components under random the slope specification of model (5.9). From section 5.3, we write

$$G_{ij}\Sigma(\hat{\theta})G_{ij}' = (1 \;\; G_{ij})\begin{pmatrix} \hat{\sigma}_{u1}^2 & \hat{\sigma}_{u12} \\ \hat{\sigma}_{u21} & \hat{\sigma}_{u2}^2 \end{pmatrix}(1 \;\; G_{ij})' = \hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2$$

$$Var(\hat{v}_{ijj}) = Var(\hat{\sigma}_e^2 + G_{ij}\Sigma(\hat{\theta})G_{ij}') = Var(\hat{\sigma}_e^2 + \hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2)$$

$$= Var(\hat{\sigma}_e^2) + Var(\hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2) + 2Cov(\hat{\sigma}_e^2, \hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2)$$

where

$$Var(\hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2) = Var(\hat{\sigma}_{u1}^2) + 4G_{ij}^2 Var(\hat{\sigma}_{u12}) + G_{ij}^4 Var(\hat{\sigma}_{u2}^2) +$$
$$+ 4G_{ij}Cov(\hat{\sigma}_{u1}^2, \hat{\sigma}_{u12}) + 2G_{ij}^2 Cov(\hat{\sigma}_{u1}^2, \hat{\sigma}_{u2}^2) + 4G_{ij}^3 Cov(\hat{\sigma}_{u12}, \hat{\sigma}_{u2}^2)$$

and $Cov(\hat{\sigma}_e^2, \hat{\sigma}_{u1}^2 + 2\hat{\sigma}_{u12}G_{ij} + \hat{\sigma}_{u2}^2 G_{ij}^2) = Cov(\hat{\sigma}_e^2, \hat{\sigma}_{u1}^2) + 2G_{ij}Cov(\hat{\sigma}_e^2, \hat{\sigma}_{u12}) + G_{ij}^2 Cov(\hat{\sigma}_e^2, \hat{\sigma}_{u2}^2)$,

Let $\Sigma(\hat{\theta}) = \begin{bmatrix} \hat{\sigma}_{u1}^2 & \hat{\sigma}_{u12} \\ \hat{\sigma}_{u21} & \hat{\sigma}_{u2}^2 \end{bmatrix}$ and denote the vector of estimated variance components as

$\hat{\delta} = \left[\hat{\sigma}_{u1}^2,\ \hat{\sigma}_{u2}^2,\ \hat{\sigma}_{u12},\ \hat{\sigma}_e^2\right]'$. Then covariance matrix of estimated variance components is

$$Var(\hat{\delta}) = \begin{bmatrix} Var(\hat{\sigma}^2_{u1}) & Cov(\hat{\sigma}^2_{u1},\hat{\sigma}^2_{u2}) & Cov(\hat{\sigma}^2_{u1},\hat{\sigma}_{u12}) & Cov(\hat{\sigma}^2_{u1},\hat{\sigma}^2_{e}) \\ Cov(\hat{\sigma}^2_{u2}\hat{\sigma}^2_{u1}) & Var(\hat{\sigma}^2_{u2}) & Cov(\hat{\sigma}^2_{u2}\hat{\sigma}^2_{u12}) & Cov(\hat{\sigma}^2_{u2}\hat{\sigma}^2_{e}) \\ Cov(\hat{\sigma}_{u12},\hat{\sigma}^2_{u1}) & Cov(\hat{\sigma}_{u12},\hat{\sigma}^2_{u2}) & Var(\hat{\sigma}_{u12}) & Cov(\hat{\sigma}_{u12},\hat{\sigma}^2_{e}) \\ Cov(\hat{\sigma}^2_{e},\hat{\sigma}^2_{u1}) & Cov(\hat{\sigma}^2_{e},\hat{\sigma}^2_{u2}) & Cov(\hat{\sigma}^2_{e},\hat{\sigma}^2_{u12}) & Var(\hat{\sigma}^2_{e}) \end{bmatrix} = I^{-1}(\hat{\delta}).$$

To evaluate covariance matrix $Var(\hat{\delta})$, we need to solve inverse of Fisher information matrix $I(\hat{\delta})$, is defined as $S = \{S_{ij}\}$ with $S_{ij} = \frac{1}{2} tr(P_{ss} V_i P_{ss} V_j)$. Here,

$$P_{ss} = V_{ss}^{-1} - V_{ss}^{-1} X_s (X_s' V_{ss}^{-1} X_s)^{-1} X_s' V_{ss}^{-1}, \text{ with}$$

$$V_{ss} = \sigma_e^2 I_n + G_s \Sigma(\theta) G_s' = \sigma_e^2 I_n + G_s \begin{bmatrix} \sigma^2_{u1} & \sigma_{u12} \\ \sigma_{u21} & \sigma^2_{u2} \end{bmatrix} G_s', \text{ and}$$

$$V_i = \frac{\partial V_{ss}}{\partial \sigma_i^2} = \begin{cases} I_n & when \quad \sigma_i^2 = \sigma_e^2 \\ G_s \left(\frac{\partial \Sigma}{\partial \sigma_i^2} \right) G_s' & otherwise \end{cases}, i = 1,...,4.$$

This leads to

$$V_1 = \frac{\partial V_{ss}}{\partial \sigma^2_{u1}} = G_s \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} G_s', \qquad V_2 = \frac{\partial V_{ss}}{\partial \sigma^2_{u2}} = G_s \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} G_s', \qquad V_3 = \frac{\partial V_{ss}}{\partial \sigma_{u12}} = G_s \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} G_s',$$

$$V_4 = \frac{\partial V_{ss}}{\partial \sigma_e^2} = I_n,$$

$$S_{11} = \frac{1}{2} tr \left\{ P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma^2_{u1}} \right) P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma^2_{u1}} \right) \right\} = \frac{1}{2} tr \left\{ \left[P_{ss} \left(G_s \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} G_s' \right) \right]^2 \right\},$$

$$S_{22} = \frac{1}{2} tr \left\{ P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma^2_{u2}} \right) P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma^2_{u2}} \right) \right\} = \frac{1}{2} tr \left\{ \left[P_{ss} \left(G_s \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} G_s' \right) \right]^2 \right\},$$

$$S_{33} = \frac{1}{2} tr \left\{ P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma_{u12}} \right) P_{ss} \left(\frac{\partial V_{ss}}{\partial \sigma_{u12}} \right) \right\} = \frac{1}{2} tr \left\{ \left[P_{ss} \left(G_s \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} G_s' \right) \right]^2 \right\},$$

$$S_{44} = \frac{1}{2} tr\left\{ P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_e^2}\right) P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_e^2}\right)\right\} = \frac{1}{2} tr\left\{\left[P_{ss} I_n\right]^2\right\},$$

$$S_{12} = S_{21} = \frac{1}{2} tr\left\{ P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u1}^2}\right) P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u2}^2}\right)\right\} = \frac{1}{2} tr\left\{ P_{ss}\left(G_s \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} G_s' \right) P_{ss}\left(G_s \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} G_s' \right)\right\},$$

$$S_{13} = S_{21} = \frac{1}{2} tr\left\{ P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u1}^2}\right) P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u12}}\right)\right\} = \frac{1}{2} tr\left\{ P_{ss}\left(G_s \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} G_s' \right) P_{ss}\left(G_s \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} G_s' \right)\right\},$$

$$S_{23} = S_{32} = \frac{1}{2} tr\left\{ P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u2}^2}\right) P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u12}}\right)\right\} = \frac{1}{2} tr\left\{ P_{ss}\left(G_s \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} G_s' \right) P_{ss}\left(G_s \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} G_s' \right)\right\},$$

$$S_{34} = S_{43} = \frac{1}{2} tr\left\{ P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_{u12}}\right) P_{ss}\left(\frac{\partial V_{ss}}{\partial \sigma_e^2}\right)\right\} = \frac{1}{2} tr\left\{ P_{ss}\left(G_s \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} G_s' \right) P_{ss}\left(I_n\right)\right\}.$$

Collecting the terms we get $Var(\hat{\delta}) = I^{-1}(\hat{\delta}) = S^{-1}$.

APPENDIX K

REGION-SPECIFIC PERFORMANCE MEASURES FOR SIMULATION SET-C IN CHAPTER 6

Figure K.1 Region-specific percentage relative biases and percentage relative RMSE for simulation set-C

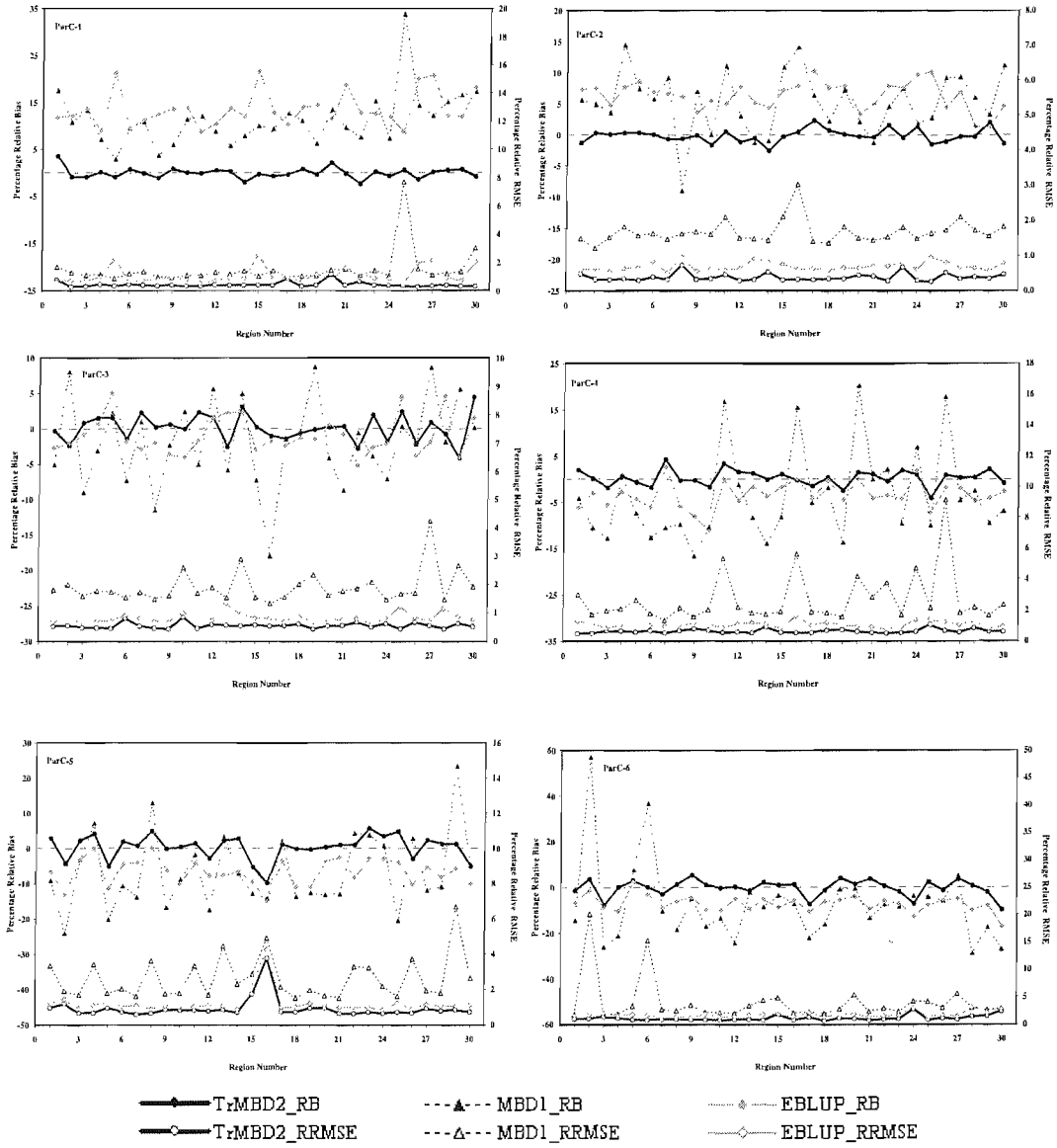
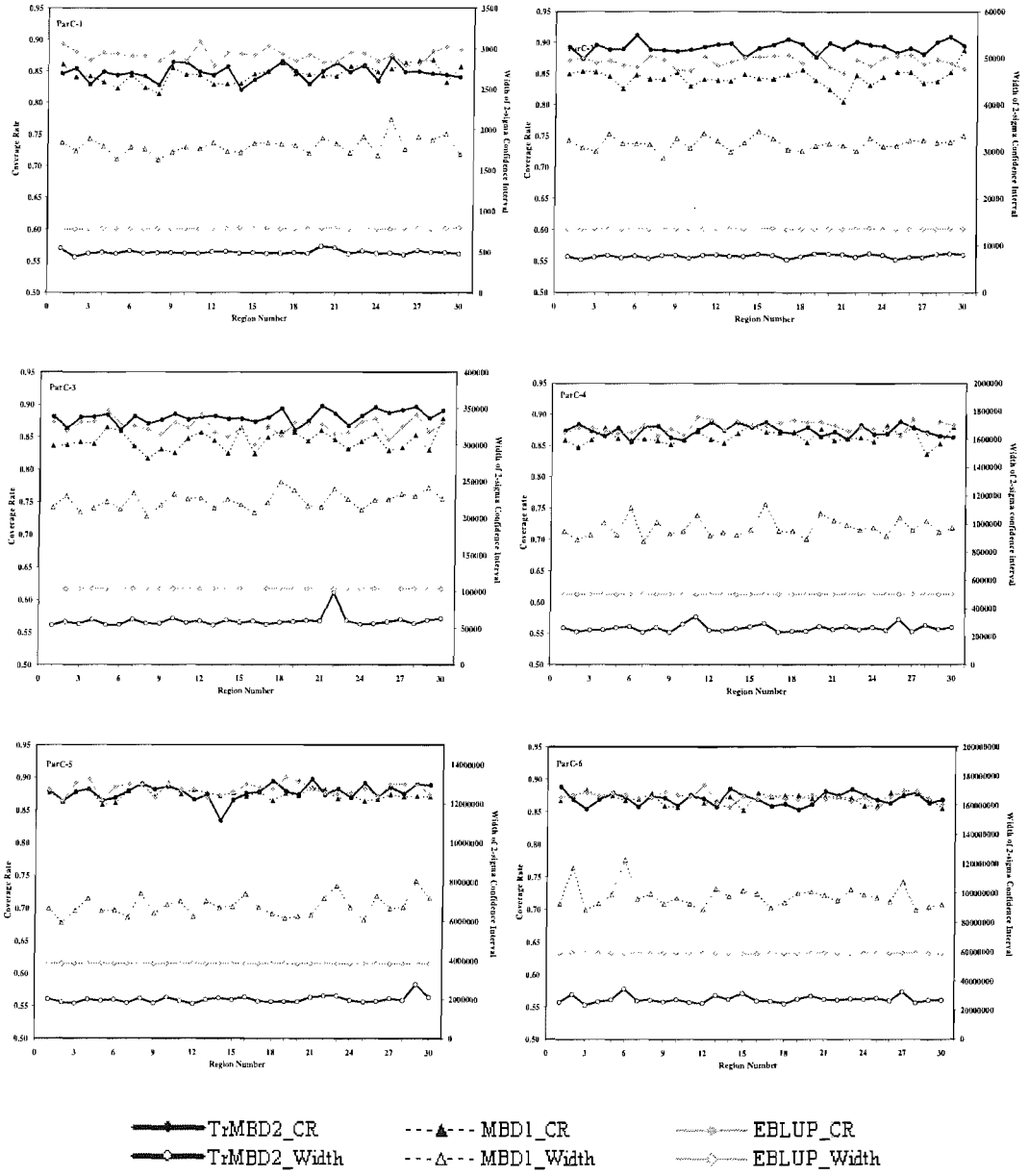


Figure K.2 Region-specific coverage rates and confidence interval widths for simulation set-C



APPENDIX L

EMPRICAL BEST PREDICTOR OF SMALL AREA MEANS

We use the model based simulations to compare the performance of Empirical Best Predictor (5.20, denoted by TrEBP) based on a log-scale linear mixed model (5.9) to the TrMBD2, MBD1 and EBLUP methods of small area estimation. See chapter 5 and 6. The set up of simulation experiment is similar to the simulations Set-A in chapter 6. In particular, we choose N=15000 (and n=600) and randomly generated the small area population (and sample) sizes N_i $\left(\text{and } n_i = N_i(n/N),\ i=1,...,m=30\right),)$ so that $\sum_i N_i = N$ (and $\sum_i n_i = n$) and kept fix throughout the simulations. Similar to Set A, population values of y_{ij} are generated for 30 areas from a multiplicative model $y_{ij} = 5.0 x_{ij}^{\beta} u_i e_{ij}$ and then draw random samples of sizes n_i from these areas. We choose six values of β (0.5, 0.8, 1.0, 1.3, 1.5 and 2.0). The random errors e_{ij} are independently generated from a LN $(0, \sigma_e)$. The random area effects u_i are generated from LN $(0, \sigma_u)$. The covariate values x_{ij} are generated from LN $(6, \sigma_x)$. Remaining part of the simulation is identical to simulation Set A. The results generated from these simulations (average relative bias and average relative RMSE) are reported in Table L.1.

Table L.1 Simulation results for Empirical Best Predictor (5.20).

	β	TrEBP	TrMBD2	MBD1	EBLUP
Average Relative Bias	0.5	-0.031	-0.026	11.897	13.762
	0.8	0.156	0.100	5.503	6.767
	1.0	0.132	-0.062	0.829	1.990
	1.3	0.650	0.321	-4.611	-4.227
	1.5	0.714	0.563	-6.457	-6.263
	2.0	1.492	1.523	-6.595	-6.466
Average Relative RMSEs	0.5	0.307	0.146	1.066	0.873
	0.8	0.417	0.263	1.474	0.611
	1.0	0.579	0.449	2.211	0.825
	1.3	0.686	0.513	2.120	0.859
	1.5	0.765	0.618	2.121	0.891
	2.0	1.033	0.853	3.260	1.400

These results show the TrEBP for skewed data under the log-transform model dominates the MBD1 and EBLUP. However, the TrMBD2 method is superior overall. The average relative biases of TrEBP are nearly same as the TrMBD2, however, average relative RMSEs of TrMBD2 are consistently smaller than the TrEBP. Although the TrEBP seems an alternative method to TrMBD2 but the mean squared error estimation is not straightforward like TrMBD2. In terms of efficiency, the TrMBD2 is more efficient. We do not carry forward this approach in our thesis.

REFERENCES

Bardsley, P. and Chambers, R.L. (1984). Multipurpose Estimation from Unbalanced Samples. *Applied Statistics*, **33**, 290–299.

Bates, D.M. and Pinheiro, J.C. (1998). Computational Methods for Multilevel Models. PostScript or PDF formats at http://franz.stat.wisc.edu/pub/NLME/.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error Components Models for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**, 28-36.

Brackstone, G.J. (1987). Small Area Data: Policy Issues and Technical Challenges. *Small Area Statistics Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.* Wiley, New York, 3-20.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* New York: Chapman and Hall.

Casella, G. and Berger, R.L. (1990). *Statistical Inference.* Duxbury Press, Belmont, California.

Chambers, R.L. (1996). Robust Case-Weighting for Multipurpose Establishment Surveys. *Journal of Official Statistics*, **12**, 3–32.

Chambers, R.L. (1997). Weighting and Calibration in Sample Survey Estimation. *Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth (Editors C. Malaguerra, S. Morgenthaler and E. Ronchetti).* Basel: Birkhäuser Verlag.

Chambers, R. L. (2005). Calibrated Weighting for Small Area Estimation. *Southampton Statistical Sciences Research Institute*, University of Southampton, U.K., WP/M05/04.

Chambers, R.L. and Dorfman, A.H. (2003). Transformed Variables in Survey Sampling. *Southampton Statistical Sciences Research Institute*, University of Southampton, U.K., WP/M03/21.

Chambers, R. L. and Chandra H. (2006). Improved Direct Estimators for Small Areas. *Southampton Statistical Sciences Research Institute*, Methodology Working Papers, M06/07, University of Southampton, U.K. http://eprints.soton.ac.uk/38465.

Chambers, R., Chandra H. and Tzavidis, N. (2007). On Robust Mean Squared Error Estimation for Linear Predictors for Domains. *In preparation.*

Chambers, R., Pratesi, M., Salvati, N. and Tzavidis, N. (2006). Spatial M-quantile Models for Small Area Estimation. Working paper No 279. *Dipartimento di Statisticae Matematica Applicata all'Economia, Università di Pisa.*

Chambers, R.L. and Skinner, C.J. (1999). Intelligent Calibration? *Proceedings of the Meeting of the International Association of Survey Statisticians*, Helsinki, 221–231.

Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika*, **93(2)**, 225-268.

Chandra, H. (2006). Small Area Estimation for Business Surveys. *Proceedings of the American Statistical Association*, Survey Research Method Section, page 2803-2809.

Chandra, H. and Chambers, R. (2006a). Small Area Estimation with Skewed Data. *Southampton Statistical Sciences Research Institute*, Methodology Working Papers, M06/05, University of Southampton, http://eprints.soton.ac.uk/38417 (submitted to the *Survey Methodology*).

Chandra, H. and Chambers, R. (2006b). Multipurpose Small Area Estimation. *Southampton Statistical Sciences Research Institute*, Methodology Working Papers, M06/06, University of Southampton, http://eprints.soton.ac.uk/38464 (submitted to the *Journal of Official Statistics*).

Chandra, H. and Chambers, R. L. (2006c). An Empirical Comparison of EBLUP Estimation and Model Based Direct Estimation for Small Areas. *Proceeding of the Q2006-European Conference on Quality in Survey Statistics,* Cardiff, United Kingdom. http://www.statistics.gov.uk/events/q2006/

Chandra, H. and Chambers, R. L. (2006d). Comparison of EBLUP Estimation and Model Based Direct Estimation for Small Areas. *Proceeding of the Stochastik-Tage-2006, German Open Conference on Probability and Statistics*, Frankfurt, Germany.

Chandra, H. and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation, *Statistics in Transition*, **7**, 637-648.

Chen, G. and Chen, J. (1996). A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood. *Survey Methodology*, **22**, 139-146.

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Cressie, N. (1993). *Statistics for Spatial Data* (Revised Edition). Wiley: New York.

Datta, G.S., Day, B. and Basawa, I. (1999). Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation. *Journal of Statistical Planning and Inference*, **75**, 269-279.

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

Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes Estimation of Median Income of Four-person Families by State Using Time Series and Cross-sectional Data. *Journal of Statistical Planning and Inference*, **102,** 83-97.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S. *Journal of the American Statistical Association*, **94**, 1074-1082.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, **76**, 341-353.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Drew, D., Singh, M. P. and Choudry, G. H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. *Survey Methodology*, **8**, 17-47.

Ericksen, E.P.(1974). A Regression Method for Estimating Population Changes of Local Areas. *Journal of the American Statistical Association*, **69**, 867 - 875.

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**, 269-277.

Fletcher, D., MacKenzie, D. and Villouta, E. (2005). Modelling Skewed Data with Many Zeros: A Simple Approach Combining Ordinary and Logistic Regression. *Journal of Environment and Ecological Statistics*, **12** (1), 45-54.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B. (1998). Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association*, **93**, 273-282.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Sciences*, **9**, 55-93.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold: New York, Wiley.

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimators. *Proceedings of the Social Statistics Section*, American Statistical Association, 33-36.

Gonzalez, M.E. and Hoza, C. (1978). Small Area Estimation with Applications to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, **73**, 7-15.

Gonzalez, M.E, and Waksberg, J.E. (1973). Estimation of the Error of Synthetic Estimates. *Proceedings of the first meeting of the International Association of Survey Statisticians*, Vienna, Austria.

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320–338.

Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics*, **31**, 423-447.

Hidiroglou, M.A., and Smith, P.A. (2005). Developing Small Area Estimates for Business Surveys at the ONS. *Statistics in Transition*, **7**, 527-539.

Holt, D., Smith, T.M.F. and Tomberlin, T.J. (1979). A Model-Based Approach to Estimation for Small Subgroups of a Population. *Journal of the American Statistical Association*, **74**, 405-410.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative Regression Estimation for Survey Data. *Proceedings of the American Statistical Association*, 300-305.

Jiang, J. and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, **15(1)**, 1-96.

Joe, W., Chris, W., and Mark, S. (2005). The Neglog Transformation and Quantile Regression for the Analysis of a Large Credit Scoring Database. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), **54(5)**, 863-878.

Kackar, R. N. and Harville, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Association*, **79**, 853-862.

Karlberg, F. (2000). Population Total Prediction Under a Lognormal Superpopulation Model. *Metron*, 53-80.

Karlberg, F. (2000a). Survey Estimation for Highly Skewed Populations in the Presence of Zeros. *Journal of Official Statistics*, **16**, 229-241.

Kleffe, J. and Rao, J.N.K. (1992). Estimation of Mean Square Error of Empirical Best linear Unbiased Predictors Under a Random Error Variance Linear Model. *Journal of Multivariate Analysis*, **43**, 1-15.

Kott, P. (1989). Robust Small Domain Estimation Using Random Effects Modelling. *Survey Methodology,* **15**, 1-12.

Kott, P.S. (2003). On Calibration Weighting. Available in PDF format at www.nass.usda.gov/research/reports/2003-jsm-**kott**.pdf.

Laake, P. (1979). A Predictive Approach to Subdomain Estimation in Finite Populations. *Journal of the American Statistical Association*, **74**, 355-358.

Lahiri, P., and Rao, J. N. K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **90**, 758-766.

Levy, P.S. and French, D.K. (1977). Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. *Vital and Health Statistics: Series 2, No. 75, DHEW Publication (PHS) 78 - 1349*. Washington: U.S. Government Printing Office.

Lewis, P.A.W. and Orlav, E.J. (1989). *Simulation Methodology for Statisticians, Operations analysts, and Engineers*. Volume 1, Belmont, Wadsworth.

Lui, K.J. and Cumberland, W. G. (1989). A Bayesian Approach to Small Domain Estimation. *Journal of Official Statistics*, **5**, 143-156.

Lui, K.J. and Cumberland, W. G. (1991). A Model-Based Approach: Composite estimators for Small Area Estimation. *Journal of Official Statistics*, **7**, 69-76.

Marker, D.A. (1999). Organization of Small Area Estimators Using a Generalized Linear Regression Framework. *Journal of Official Statistics* **15**, 1-24.

Marshall, R.J. (1991).Mapping Disease and Mortality Rates Using Empirical Bayes Estimators. *Applied Statistics*, **40**, 283-294.

McCulloch, C.E., and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

Militino, A.F., Ugarte, M. D. and Goicoa, T (2007). Combining Sampling and Model Weights in Agriculture Small Area Estimation. *Environmetrics*, **18**, 87–99.

Morgan, B.J.T (1984). *Elements of Simulations*. London, Chapman and Hall.

Moura, F.A.S. and Holt, D. (1999). Small Area Estimation Using Multilevel Models. *Survey Methodology*, **25**, 73-80.

National Health Interview Survey (1968). Synthetic State Estimates of Disability. *PHS Publication No. 1759. Public Health Service*, Washington: U.S. Government Printing Office.

Opsomer, J.D., G. Claeskens, M.G. Ranalli, G. Kauermann and F.J. Breidt (2005). Nonparametric Small Area Estimation Using Penalized Spline Regression. Preprint

Series 05-01, *Department of Statistics, Iowa State Univ*ersity, *USA*, http://www.public.iastate.edu/~jopsomer/papers/Pspline_SME.pdf

Park, M. and Fuller, W.A. (2005). Towards Nonnegative Regression Weights for Survey Samples. *Survey Methodology*, **31**, 85-93.

Peixoto, J.L. and Harville, D.A. (1986). Comparisons of Alternative Predictors Under the Balanced One-Way Random Model. *Journal of the American Statistical Association*, **81**, 431-436.

Petrucci, A., Pratesi, M. and Salvati, N. (2005). Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects. *Statistics in Transition*, **7(3)**, 609-623.

Petrucci, A. and Salvati, N. (2004). Small Area Estimation Considering Spatially Correlated Errors: the Unit Level Random Effects Model. Working Paper 2004/10. *Dipartimento di Statistica "G. Parenti",Firenze.*

Pfeffermann, D. (2002). Small Area Estimation: New Developments and Directions. *International Statistical Review*, **70**, 125-143.

Pfeffermann, D. and Burck, L. (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. *Survey Methodology,***16**, 217-237.

Pfeffermann, D., Feder, M., and Signorelli, D. (1998). Estimation of Auto-correlations of Survey Errors with Application to Trend Estimation in Small Areas. *Journal of Business and Economic Statistics,* **16**, 339-348.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Prasad, N.G.N. and Rao, J.N.K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, **25**, 67-72.

Pratesi, M. and Salvati, N. (2005). Small Area Estimation: the EBLUP Estimator with Autoregressive Random Area Effects, *Dipartimento di Statistica e Matematica Applicata all'Economia, Pisa,* Report No 26.

Pratesi, M., Salvati, N., Ranalli, M.G. and Chambers, R. (2006). Nonparametric M-Quantile Small-Area Estimation via Penalized Splines. *Proceedings of the Section on Survey Research Method, American Statistical Association*, 2006.

Purcell, N. J. and Kish, L. (1979). Estimation for Small Domain. *Biometrics,* **35**, 365-384.

Purcell, N. and Linacre, S. (1976). Techniques for the Estimation of Small Area Characteristics. *Proceedings of the third Australian Statistical Conference*, 18-20.

Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology*, **25**, 175-186.

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

Rao, J.N.K. and Choudry, G. H. (1995). Small Area Estimation: Overview and Empirical Study, in B.G. Cox et al (editors). *Business Survey Methods*, John Wiley & Sons, New York, 527-542.

Rao, J.N.K. and Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data. *Canadian Journal of Statistics*, **22**, 511-528.

Royall, R.M. (1976). The Linear Least-Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association,* **71**, 657-664.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population sampling. *Journal of the American Statistical Association,* **73**, 351-358.

Saei, A. and Chambers, R. (2003). Small Area Estimation: A Review of Methods Based on the Application of Mixed Models. *Southampton Statistical Sciences Research Institute*, University of Southampton, U.K., WP-M03/16.

Särndal, C.E. and Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association*, **84**, 266-275.88.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer- Verlag, New York.

Schaible, W.L. (1978). Choosing Weights for Composite Estimators for Small Area Statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 741-746.

Schaible, W.L., Brock, D.B. and Schnack, G.A. (1977). An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics. *Proceedings of the Social Statistics Section*, American Statistical Association, 1017-1021.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatial-Temporal Models in Small Area Estimation. *Survey Methodology*, **31**, 183-195.

Skinner, C.J. (1993). The Use of Synthetic Estimation Techniques to Produce Small Area Estimates. *New Methodology Series*, *NM18*Office of Population, Censuses and Surveys.

Stukel, D.M. and Rao, J.N.K. (1999). On Small Area Estimation Under Two-fold Nested Error Regression Models. *Journal of Statistical Planning and Inference*, **78**, 131-147.

Tiller, R.R. (1992). A Time Series Approach to Small Area Estimation. *Proceedings of the Social Statistics Section*, American Statistical Association, page 10-19. http://www.amstat.org/Sections/Srms/Proceedings/papers/1992_002.pdf

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference- A Prediction Approach*. John Wiley & Sons.

Wu, C. and Sitter, R.R. (2001). A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data. *Journal of the American Statistical* Association, **96**, 185 -193.

You, Y. and Rao, J.N.K. (2002). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *Canadian Journal of Statistics*, **30**, 431-439.

You, Y., Rao, J.N.K. and Kovacevic, M. (2003). Estimating Fixed Effects and Variance Components in a Random Intercept Model Using Survey Data. *Proceedings of Statistics Canada Symposium*, Challenges in Survey Taking for the Next Decade.