SOME NEW MODELS FOR SMALL AREA ESTIMATION

By

Hao Ren

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics

2011

ABSTRACT

SOME NEW MODELS FOR SMALL AREA ESTIMATION

By

Hao Ren

This dissertation includes some new models for small area estimation. There are four parts in total. The first part studied the selection of fixed effects covariates in linear mixed models. A modified bootstrap selection procedure for linear model from literature was extended to linear mixed effects models. Both theoretical work and simulations showed the effectiveness of this procedure for linear mixed effects models.

In the second part, a new approach by shrinking both means and variances of small areas was introducted. This method modeled the small area means and variances in a unified framework. The smoothed variance estimators used information of direct point estimators and their sampling variances, and consequently, for the smoothed small area estimators. Conditional mean squared error of prediction was also studied in this part to evaluate the performance of predictors.

The third part studied the confidence intervals of small area estimators introduced in the second part. The literature of small area estimation is dominated by point estimation and their standard errors. The standard normal or student-t confidence intervals do not produce accurate intervals. The confidence intervals produced in this part are from a decision theory perspective.

The fourth part estimated the small areas means with clustering of the small areas. In the realistic application, the estimation may not be appropriate to "borrow strength" from all other small areas universally, if cluster effects exist between clusters of small areas. A model

based on clustering was studied in this part, which included an additional cluster effect to the basic area level model. Since the partition of clusters was not known, a stochastic search procedure from literature was adapted first to find the clustering partition.

ACKNOWLEDGMENT

I would like to express my sincere appreciation to my major dissertation advisor, Professor Tapabrata Maiti; for his invaluable guidance, support, and encouragement throughout my dissertation study. This dissertation could not have been completed without his guidance and support.

I would also like to thank Professors Sarat Dass, Lifeng Wang, and William Schmidt for serving on my guidance committee. My special thanks go to Professor Sarat Dass for his kindly help and great advice for my dissertation. I also want to thank Professor Lifeng Wang and William Schmidt for their valuable help and support to my dissertation.

I also want to thank Adele Brandstrom and Jessalyn Smith for their kind help on style editing.

I would like to thank my wife Qi Diao for her support, encouragement and love. Finally, I would like to thank my beloved parents and sister for their consistent support and love. Thank you!

TABLE OF CONTENTS

List of Tables . . . . . **vii**

List of Figures . . . . . **ix**

1 Introduction **1**
1.1 Small Area Estimation 1
1.2 Classical Approach for Sample Survey 3
1.3 Model Based Estimation 5
1.3.1 Area Level Model 6
1.3.2 Unit Level Model 7
1.4 Mixed Model 8
1.5 Study Topics 10

2 Bootstrap Model Selection for Linear Mixed Effects Models: Application to Small Area Estimation **17**
2.1 Introduction 17
2.2 Linear Model 18
2.3 Linear Mixed Model 21
2.3.1 Fay-Herriot Model 21
2.3.2 Nested-Error Regression Model 24
2.4 The Theoretical Framework 27
2.5 Simulation Study 34
2.5.1 Algorithms and Settings for Different Models 34
2.5.2 Results 37

3 Prediction Error of Small Area Predictors Shrinking both Mean and Variances **53**
3.1 Introduction 53
3.2 Model and Estimation Method 54
3.2.1 Model with Assumption 54
3.2.2 Estimation of the Small Area Parameters 56
3.2.3 Estimation of the Stuctural Parameters 59
3.3 Prediction Error Calculation 62
3.3.1 Mean Squared Error of Prediction 62
3.3.2 Bias Correction for $\nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2)$ 63
3.3.3 Approximation of $c_i(\boldsymbol{B}; X_i, S_i^2)$ 66
3.4 Simulation Study 67

3.5 Appendix: Matrix Calculation Results 71
3.5.1 Computation of $\hat{\boldsymbol{B}} - \boldsymbol{B}$ 71
3.5.2 Second Order Correction of ν_i 75

4 Confidence Interval Estimation of Small Area Parameters Shrinking both Mean and Variances 79
4.1 Introduction 79
4.2 Proposed Model 80
4.3 Confidence Interval 81
4.3.1 Definition 81
4.3.2 Choice of the Tuning Parameter 83
4.4 Theoretical Justification of Tuning Parameter 84
4.5 Alternative Model for Confidence Interval 87
4.6 Simulation Study 90
4.7 A Real Data Analysis 94

5 Clustering Based Small Area Estimation: An Application to MEAP Data 99
5.1 Introduction 99
5.2 Data Set 104
5.3 Proposed Model 106
5.3.1 Prior Information 106
5.3.2 Model-Based Objective Functions 107
5.3.3 Estimation of Mean Scores 109
5.3.4 MSE of Estimator 110
5.4 Stochastic Search 111
5.4.1 Biased Random Walk 111
5.4.2 Split-Merge Moves 112
5.5 Data Analysis 113

6 Discussion 119

Bibliography 123

LIST OF TABLES

2.1 Selection probabilities for linear model 40

2.2 The coverage probabilities and lengths of confidence intervals for linear model 41

2.3 Selection probabilities for Fay-Herriot model, loss function without random effect 42

2.4 The coverage probabilities and lengths of confidence intervals for the Fay-Herriot model, loss function without random effect 43

2.5 Selection probabilities for Fay-Herriot model, loss function with random effect 44

2.6 The coverage probabilities and lengths of confidence intervals for the Fay-Herriot model, loss function with random effect 45

2.7 Selection probabilities for Nested-Error regression model with group level . . 46

2.8 The coverage probabilities and lengths of confidence intervals for for Nested-Error regression model with group level 47

2.9 Selection probabilities for Nested-Error regression model with element level . 48

2.10 The coverage probabilities and lengths of confidence intervals for Nested-Error regression model with element level 49

2.11 Compare of loss functions without and with random effect for Nested-Error model, data size $n = 160$. 50

2.12 The coverage probabilities and lengths of confidence intervals for Nested-Error regression model with different loss functions, data size $n = 160$. 51

3.1 Results of the simulation study, and here we present estimate (Est.), empirical standard deviation (SD) for β and τ^2. We set $\beta = 10$. 69

3.2 Results of MSE estimator, $n = 36$, $m = 9$ 70

3.3 Results of MSE estimator, $n = 180$, $m = 18$ 70

4.1 Estimation of model parameter. The left panel is for β and the right panel is for τ^2 91

4.2 Simulation results for prediction when $\tau^2 = 0.25$ 92

4.3 Simulation results for prediction when $\tau^2 = 1$ 92

4.4 Simulation results for prediction when $\tau^2 = 4$ 93

4.5 Crop data from You and Chapman(2006) 95

4.6 Estimation results of corn 95

5.1 Number of public school 4th graders in some school districts in Michigan state 100

LIST OF FIGURES

4.1 Corn hectares estimation. The vertical line for each county displays the confidence interval of $\hat{\theta}_i$, with $\hat{\theta}_i$ marked by the circle, for (I) Proposed method, (II)Wang and Fuller (2003), (III)Hwang *et al.* (2009) and (IV)Qiu and Hwang (2007). 96

4.2 Boxplot of estimates of corn hectares for each county. (I) to (IV) are the 4 methods corresponding to Figure 4.1. 97

5.1 Standard deviation of public school students' assessment scores in available school districts in Michigan state. 100

5.2 Poverty level of available school districts in Michigan state. 102

5.3 Math scores of 4th graders in public schools in available school districts in Michigan state. 104

5.4 Trace plot of different m values. 114

5.5 Map of clusters. 115

5.6 Mean scores vs. poverty levels for school districts in the 7 clusters and all the districts. 116

5.7 Coefficient of variation for the model-based estimates and observed mean scores. 117

5.8 MSE of estimators for models with clustering (5.2) or no clustering (5.1). . 117

Chapter 1

Introduction

1.1 Small Area Estimation

The term small area is commonly used to denote a small geographical area, such as a district. It can also be used to denote a small demographic group, such as a small group with certain social economic status or a sex/race/ethnicity group. Usually, a small area is defined when the domain specific sample is not large enough to support direct estimates with an adequate level of statistical precision.

The history of small area statistics is very long and can be traced back at least to eleventh century England and seventeenth century Canada. However, those are based on either census or administrative records targeting a complete enumeration (Brackstone, 1987) and sampling is usually not involved in those studies.

In recent years, sample surveys have become more and more popular because they are cost-effective and can help solve the issue that occurs when the population is dynamic and the individuals making up the population are moving constantly. Some basic sample survey

methods can be simple random sampling, systematic or stratified sampling. From the collected sample, the direct measures, such as the mean and standard error, can be calculated as the estimates for the population or each domain. However, if a small area exists, that is, if the sample for some domain is not large enough, it may result in unacceptably large standard errors if it is only based on direct survey estimators and only from the sample area data. Thus, it is important to research small area statistics to develop more reliable measures.

Small area estimation has gained more attention in recent years because of greatly increasing demand from both public and private sectors. For public sectors, from countries like the U.S. and Canada, they have a growing demand for small area statistics for the purpose of formulating policies and programs, in cases like government funding allocation and local regional planning. In order to identify areas that are in need of funding, such as certain school districts or some subpopulation, the reliable estimates from the small area are required. The increasing demand for small area statistics also comes from private sectors, like small businesses. Often, the business decisions rely on the local social-economic environment and other regional conditions. Therefore, estimates with adequate precision are needed from small areas. Since direct estimates from sample surveys sometimes cannot meet such requirements for small areas, the research on small area statistics has become more and more important.

What is more, when central and eastern European countries and the former Soviet Union countries moved away from a centralized decision making system, they also demand survey results not only for large areas but also small areas. Again, this created an increasing demand for the small area statistics.

With high power computers, the processing of large and complex data become feasible and it helps with the development of the small area statistics. Powerful statistical methods with a sound theoretical foundation have been developed for the analysis of small area data. Such methods "borrow strength" from related or similar small areas through implicit or explicit models which provide a link between related small areas. These types of "borrow strength" models will be the main focus of this dissertation. For review on small area estimation, papers include Ghosh and Rao (1994), Rao(1999), Marker(1999), Rao (2001) and Pfeffermann (2002), etc.

1.2 Classical Approach for Sample Survey

Direct Estimators

The variables of interest in a sample survey are usually the total measures or mean of the area or domain. Direct estimators are commonly used to provide estimate of such variables in a domain and use the sample data only in that domain. The typical direct estimators are design-based. A more extensive reference of direct estimation in sampling theory can be found in Lohr (1999). Direct estimators will sometimes suffice, such as domains with sufficiently large sample size and particularly after addressing survey design issues. But, it is well known that direct estimators for small areas are usually unreliable for the unacceptable large standard errors due to unduly sample sizes of small areas.

Model-based methods for direct estimators are also developed. They provide valid conditional inferences about the particular sample drawn, regardless of the sampling design. However, model-based methods depend heavily on the correct specification of models. The methods can perform poorly under misspecification even if the sample size is large.

Demographic Methods

The most powerful demographic method is census. Censuses are usually conducted at 10-year or 5-year intervals to provide population counts for specific geographical areas or subpopulations defined by age, sex, marital status and other demographic variables. But the information from a census becomes outdated due to changes in the size and composition of the resident population over time. Therefore, various demographic methods, other than census, are developed to provide population estimation in the noncensal years.

The changes of demographic variables are strongly related to changes of local population. Administrative registers contain current data of local population on various demographic variables. Such variables are called symptomatic indicators, such as number of births and deaths and net emigration during the period since the last census. Traditional demographic methods employ indirect estimators based on implicit linking models, which related the population estimates and symptomatic variables. These methods may be categorized as either symptomatic accounting techniques or regression symptomatic procedures. Symptomatic accounting techniques provide indirect estimators under some implicit linking models with the symptomatic variables. Regression symptomatic procedures use multiple linear regression to estimate local area populations. The symptomatic variables are used as independent variables. For detailed description of these methods, one can see Rao (2003).

Typically, demographic indirect estimators use only administrative and census data and sampling is not involved in these methods.

Indirect Estimators

As introduced previously, the unacceptably large standard errors of direct estimators are due

to unduly sample size of the small areas. Therefore, it is necessary to find indirect estimators that increase the effective sample size and thus decrease the standard error.

Traditional indirect estimation methods are based on implicit models that provide a link to related small areas through supplementary data. Such estimators include synthetic estimators, composite estimators, and James-Stein (or shrinkage) estimators. If a large area covers several small areas and the small areas are assumed to have the same characteristics as the large area, a reliable direct estimator for the large area can be used to derive an indirect estimator for a small area. Such an estimator is called synthetic estimator (Gonzalez 1973). The global measures (averaged over small areas) are often used with synthetic estimates. If an estimator is a weighted average of a synthetic estimator and a direct estimator, it is called a composite estimator. Actually, any estimator that has the composite form can be called a composite estimator, both design-based and model-based. For a composite estimator, a suitable weight needs to be chosen. The common weight approach uses a common weight for the composite estimators for all small areas, then the total MSE is minimized with respect to the common weight. The James-Stein estimator (James and Stein, 1961) is similar to the common weight estimator and attracted a lot of attention in mainstream statistics literature. It achieves large gains in efficiency in the traditional design-based framework without assuming a model on the small area parameters. A detailed introduction of these estimators can be found in Rao (2003).

1.3 Model Based Estimation

The traditional indirect estimators are briefly introduced in the previous section. Reduction in MSE is the main reason for using indirect estimators. Indirect estimators are largely

based on sample survey data in conjunction with auxiliary population data. However, the traditional indirect estimators only provide an implicit link between the small areas. The model based estimators which provide an explicit link between the small areas are introduced in this section.

Model based estimators are indirect estimators based on small area models. Small area models take the random area-specific effects into account and include additional auxiliary variables in the model to explain the effects, which make specific allowance for between area variation. Such models define the way how the related data are incorporated in the estimation process. The use of explicit models makes model diagnostics possible which can be used to find suitable models that fit the data well. For example, a selection of auxiliary variables is conducted later in this dissertation. Area-specific measures of precision also can be associated with each small area estimate. For complex data structures, more complex model structures can be adopted, such as mixed models and generalized linear models. And the existing methodologies for these models can be utilized directly to achieve accurate small area estimation.

In general, small area models can be classified into two broad types: (i) area level model: modeling the small area direct estimators with area-specific covariates. When unit level data are not available, such area level models are necessary. (ii) unit level model: modeling the unit direct estimators with unit-specific covariates.

1.3.1 Area Level Model

Let $\bar{Y}_i$ be the small area means and $g(\cdot)$ be a specified function that link the model based estimator θ_i and $\bar{Y}_i$, $\theta_i = g(\bar{Y}_i)$. The use of $g(\cdot)$ makes the model more robust. But in our

study, we choose $g(\cdot)$ as the identical function, i.e. $\bar{Y}_i = \theta_i$. The area level model relates the θ_i and the area-specific auxiliary data $z_i = (z_{1i}, \ldots, z_{pi})^T$ in the following way,

$$\theta_i \quad = \quad z_i^T \beta + b_i v_i, \quad i = 1, \ldots, m \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is the $p \times 1$ vector of regression parameters, b_i's are known positive constants, and v_i's are area-specific random effects that are iid with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$. In this dissertation, a normal distribution is always adopted for v_i. However, it is possible to choose other distributions which makes the model robust. m is the total number of areas.

For making inferences about the θ_i's under model (1.1), we assume that direct estimators, $\hat{\theta}_i$, are available and

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, \ldots, m \tag{1.2}$$

where the e_i's are sampling errors, $E(e_i) = 0$ and $Var(e_i) = \sigma_{e_i}^2$. That is, the estimators $\hat{\theta}_i$ are design-unbiased. And we always assume that the sampling variances, $\sigma_{e_i}^2$, are known.

The area level models have various extensions. A famous example is the Fay-Herriot model. The model was developed by Fay and Herriot in 1979 to estimate per-capita income for a small area (population less than 1,000). The Fay-Herriot model is adapted in this dissertation and the details will be introducted later.

1.3.2 Unit Level Model

When unit-specific auxiliary data for each population element in each small area is available, the unit level model is adapted. Let y_{ij} be the variable of interest and $X_{ij} =$

$(x_{ij1}, \ldots, x_{ijp})^T$ be the available element-specific auxiliary data. Then a one-fold nested error linear regression model is given

$$y_{ij} = X_{ij}^T \beta + \nu_i + e_{ij} \tag{1.3}$$
$$j = 1, \ldots, n_i; \quad i = 1, \ldots, m.$$

Here ν_i are the area-specific effects and are assumed to be iid random variables with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$. $e_{ij} = k_{ij}\tilde{e}_{ij}$ and the $\tilde{e}_{ij}$'s are iid random variables, independent of the ν_i's, with $E(\tilde{e}_{ij}) = 0$ and $Var(\tilde{e}_{ij}) = \sigma_e^2$, the k_{ij}'s are known constants, and n_i are the number of elements in the ith area. In addition, normality of the ν_i's and $\tilde{e}_{ij}$'s is often assumed. The parameters of inferential interest are the small area totals $Y_i = \sum_{i=1}^{n_i} y_{ij}$ or the means $\bar{Y}_i = Y_i/n_i$.

The unit level model does not include sample selection bias in the model; that is, the sample values are assumed to obey the model. Simple random sampling is satisfied for this condition. But for more complex sampling designs, it may not be appropriate. For example, in stratified multistage sampling, the design features are not incorporated in the model. However, there are various extensions to account for such features.

1.4 Mixed Model

The small area models introducted in the previous section may be regarded as special cases of the mixed models. The research of Mixed Models has gained much attention in recent years. It has many names in in a wide variety of disciplines in the physical, biological and social sciences. The mixed model is particularly useful in settings where repeated measurements

are made on the same statistical units (longitudinal data), or where measurements are made on clusters of related statistical units (clustered or panel data). Therefore, it can be called the model for repeated measurements, or a hierarchical model.

The most important difference of the mixed model from classical statistics is that observations are not necessary from the same population, where independent and identically distributed is a typical assumption. Mixed model data may have a more complex, multilevel, hierarchical structure. The general assumption of mixed model data is that observations from one cluster can be correlated, but independent between different clusters; and the subpopulations of clusters can be different. Therefore, a random effect at each cluster level is introduced into the model. The forms of a mixed model can be linear, generalized linear (such Logistic and Poisson), and nonlinear. In this dissertation, we focus on the linear mixed models only.

A general linear mixed model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e} \tag{1.4}$$

where $\boldsymbol{y}$ is the vector of sample observations, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known matrices, and $\boldsymbol{v}$ and $\boldsymbol{e}$ are distributed independently with means 0 and covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$, respectively, depending on some variance components parameters.

One approach to this model can be obtained by using the general theory of Henderson (1975) for a mixed linear model. If the parameters of variance components are known, the best linear unbiased estimator of $\boldsymbol{\theta} = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{v}$ is given by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{1.5}$$

where $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T$ is the variance-covariance matrix of $\boldsymbol{y}$ and

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y})$$

is the generalized least squares estimator of $\boldsymbol{\beta}$.

There are a lot of other approaches available, such as empirical best linear unbiased prediction, empirical Bayes method, and hierarchical Bayes method. Some of these methods will be involved in our studies later. The details of these methods can be found in literatures, such as Ghosh and Rao (1994) and Rao (2003).

1.5 Study Topics

As I introduced in previous sections, small area estimation and the statistical techniques therein have become a topic of growing importance in recent years and this is the reason why I chose small area estimation as the topic of my dissertation. To be more specific, the works in this dissertation includes the following aspects.

Model Selection for Linear Mixed Effects Models

The need for reliable small area estimates is felt by many agencies, both public and private, for making useful policy decisions. For example, the small area statistics being used to monitor socio-economic and health conditions for different stratum defined by age, sex, racial groups over small geographical areas.

It is now widely recognized that the direct survey estimates for small areas are usually unreliable, being accompanied with large standard errors and coefficients of variation. This

makes it necessary to use models, either explicit or implicit, to connect the small areas and obtain estimates of improved precision by borrowing strength across areas.

In the previous sections, some basic models for small area estimation is presented. Before we start working on more complex approach methods, it is important to note the influence on the choice of models, particularly on the choice of auxiliary variables. The success of any model-based method depends on the availability of good auxiliary data. Therefore, more attention should be given to the selection of auxiliary variables that are good predictors of the study variables.

The model selection study here arises from potential choices of the fixed effects covariates in linear mixed effects models. A bootstrap model selection method is adapted in this dissertation. The procedure based on the bootstrap has some important features: first, the bootstrap method can provide more accurate inference (Adkins and Hill, 1990; Hall,1989) at the same time when the model selection is undergoing. In other words, the bootstrap based inference on regression parameters takes into account the model selection procedure. Second and most importantly, the bootstrap selection procedure does not depend on a specific probability distribution.

However, a straightforward application of the bootstrap does not yield a consistent model selection procedure (Shao, 1996). A simple modification was applied to the straightforward bootstrap procedure: generate m (instead of n, and $m < n$, n is the size of all available data) iid bootstrap observations from the empirical distribution which puts equal mass on each pair of available data. After the modification, it will result with consistency if and only if $m/n \to 0$ and $m \to \infty$.

However, only the linear model is studied in the literature. In this dissertation, the modified bootstrap selection procedure is extended to linear mixed effects models, including the Fay-Herriot Model and the Nested-Error Regression Model, that are commonly used in small area estimation.

Small Area Predictors by Shrinking both Mean and Variances

The basic area level model is introducted in section 1.3.1. The survey based direct small area estimates and their variance estimates are the main ingredient to build area level small area models. Typical modeling strategies assume that the sampling variances are known while a suitable linear regression model is assumed for the means. For detailed developments, one can see Ghosh and Rao (1994), Pfeffermann (2002) and Rao (2003). The typical area-level models are subject to two main criticisms: (i) in practice, the sampling variances are estimated quantities and these are subject to substantial errors due to the fact that they are often based on equivalent sample sizes as the direct estimates are being calculated and (ii) assumption of known and fixed sampling variances does not take into account the uncertainty of estimation into the overall small area estimation strategy.

Previous attempts have been made to model only the sampling variances by Maples et al. (2009), Gershunskaya and Lahiri (2005), Huff et al. (2002), Cho et al. (2002), Valliant (1987), and Otto and Bell (1995). Wang and Fuller (2003) and Rivest and Vandal (2003) extended the asymptotic mean squared error (MSE) estimation of Prasad and Rao (1990) when the sampling variances are modeled with few additional parameters. You and Chapman (2006) considered sampling variance modeling. However, they adopted full Bayesian estimation techniques.

The issues have been recently raised by the practitioners. For example, Herrador et al. (2008) investigated estimates of design variances of model based and model-assisted small area estimators. The latest developments are nicely summarized in a recent article by William Bell of the United States Census Bureau (Bell, 2008). He carefully examined the effect of the above two issues in the context of MSE estimation for model based small area estimators. He also provided numerical evidence of the effect of assuming known sampling variances in the estimation of MSE in the context of the Fay-Herriot model. The developments so far made to this issue can be "loosely" viewed as (i) smoothing the direct sampling error variances to obtain stable variance estimates with low bias and (ii) (partial) account of sampling variance uncertainty by extending the Fay-Herriot model.

Much less or no attention has been given to accounting the sampling variances effectively while modeling the mean compared to the volume of research that has been devoted for modeling the means and their inferential issues. Thus, there is a lack of systematic development in small area estimation which includes shrinking both means and variances. In other words, we would like to exploit the technique of "borrowing strength" from other small areas to "improve" the variance estimates as we do to improve the small area mean estimates. The methodology introduced here develops a dual "shrinkage" estimation for both the small area means and variances in a unified framework. In this process, the smoothed variance estimators use information of direct point estimators and their sampling variances, and consequently, for the smoothed small area estimators.

The modeling perspective is closely related to Wang and Fuller (2003), Rivest and Vandal (2003), You and Chapman (2006) and Hwang et al. (2009). An EM-based estimation approach is developed. Numerical evidences are provided to show effectiveness of the dual

shrinkage estimation.

One statistic reported is conditional mean squared error of prediction (CMSEP) which is more akin to Booth and Hobert (1998). Booth and Hobert argued strongly for CMSEP as opposed to unconditional mean squared error of prediction (UMSEP). Recently, this technique has again been emphasized by Lohr and Rao (2009) in the context of nonlinear mixed effect models. These authors favor CMSEP particularly for non-normal models when the posterior variance of small area parameters depends on the area specific responses. Although they were interested only in generalized linear mixed models where the posterior variance depends on area specific responses, this property of posterior variance is perhaps true for situations with posterior non-linearity.

Another inference adopted is confidence intervals of small area means. The small area estimation literature is dominated by point estimation and their standard errors. It is well known that the standard practice of (pt. est. $\pm\, q$s.e.), q is a Z (standard normal) or a t cut-off point, does not produce accurate intervals. See, Hall and Maiti (2006) and Chatterjee et al. (2008) for more details. The previous works are based on the bootstrap procedure and has limited use due to repeated estimation of model parameters. The confidence intervals produced in this dissertation are from a decision theory perspective.

Clustering Based Small Area Estimation

For the introduced small area models, both the area level model and the unit level model assumed iid random area effects. That is, when we make inference of each small area, we "borrow strength" from all other small areas universally. But the realistic geographical (spatial) and socio-economic status of small areas in a large region may be quite different from each other. For example, the demographic composition of a small area might be close

to the adjacent small areas, but might not be all the adjacent areas; a similar thing exists for house-hold incomes, poverty levels and many other types of information. When we "borrow strength" from other small areas under such conditions, the inference may even be misleading. Therefore, it is more appropriate to divide all small areas into groups (clusters) if diversities exist between groups of small areas.

In this dissertation, a data set from the Michigan Educational Assessment Program (MEAP) is analyzed. The numbers of students for each school district is diverse. The standard deviations of students' math scores for each school district are also quite different from each other. A detailed description of the data set is given the later chapters. These information show that small area models are appropriate to be adopted to analyze the data set; and the comparisons of mean scores based on model-based estimate results are more meaningful than the direct comparison of mean score. In addition, the poverty levels of school districts suggest that small area models with clustering the school districts is more appropriate.

There are a lot of clustering methods available. However, it usually happens that we do not have partition information for clustering, or even the number of clusters. A stochastic search procedure from Booth et al. (2008) is adopted to solve the problem. The partition of clustering is involved in the model as a parameter. A cluster-specific random effect is also included in the model, which allows the cluster means departure from the assumed base model. The objective function is constructed based on the posterior distribution of the undergoing partition. The partition maximizes the objective function is chosen as the "optimal" clustering partition.

In Booth et al. (2008), they assumed the variances of objects were unknown but iid within a cluster. The main difference between our model and the model in Booth et al. (2008) is that different variances are assigned to each school district and assumed as known since they are reported in the MEAP data set. The posterior distribution based on different variances for each school district does not have a close form and only can be calculated by numerical method.

Chapter 2

Bootstrap Model Selection for Linear Mixed Effects Models: Application to Small Area Estimation

2.1 Introduction

Let $\{(x_i, y_i), i = 1, \ldots, n\}$ be the available data set. Some models always need to be fitted if the objective is to discover the relationship between x and y. Often, all the components of x may not be related or important to y. An optimal model contains the necessary components of x. The procedure of selecting the variables is a model selection problem in which each model corresponds to a particular set of the components of x.

Among the many existing variable/model selection procedures, such as AIC, BIC, Mellow's C_p, R^2 etc., the procedure based on the bootstrap has some important features: first, the bootstrap method can provide more accurate inference (Adkins and Hill, 1990; Hall,

1989) at the same time when the model selection is undergoing. In other words, the bootstrap based inference on regression parameters takes into account the model selection procedure. Second and most importantly, the bootstrap selection procedure does not depend on a specific probability distribution.

A very important theoretical study of a bootstrap selection procedure is its consistency. Shao (1996) discovered that a straightforward application of the bootstrap does not yield a consistent model selection procedure. A simple modification was applied to the straightforward bootstrap procedure: generate m (instead of n, and $m < n$) independent and identical (iid) bootstrap observations from the empirical distribution $\hat{F}$, which puts mass n^{-1} on each pair (x_i, y_i), $i = 1, \ldots, n$. After the modification, it will result in consistency if and only if $m/n \to 0$ and $m \to \infty$.

However, only the linear model is studied in the literature. In this chapter, the modified bootstrap selection procedure is extended to different aspects of the mixed model, including the Fay-Herriot Model and Nested-Error Regression Model, that are commonly used in small area estimation. The model selection study here arises from potential choices of the fixed effects covariates. In the following sections, we first start with a replication of Shao's algorithm in the linear model. Then the modified algorithms of model selection procedures for linear mixed models is provided. After that, the simulation studies were carried out with respect to the previous algorithms. The results are listed in the final section.

2.2 Linear Model

The model is given by

$$y_i = x_i^T \beta + e_i, i = 1, \ldots, n \tag{2.1}$$

where x_i is the ith value of a $p \times 1$ vector of explanatory variables and y_i is the response at x_i, β is a $p \times 1$ vector of unknown parameters. In our study, p is fixed, and does not increase as n increases. We assume that e_i, $i = 1, \ldots, n$, are iid Normal$(0, \sigma_e^2)$ and

$$\mu_i = E(y_i|x_i) = x_i^T \beta \quad var(y_i|x_i) = \sigma_e^2 \quad i = 1, \ldots, n$$

The unknown parameters are estimated by the least squares estimator (LSE)

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.2}$$

For the variable/model selection procedures, let α be a subset of $1, \ldots, p$ of size p_α and let $x_{i\alpha}$ (or β_α) be the subvector of x_i (or β) containing the components of x_i (or β) indexed by the integers in α. Then a candidate model, say model α, is

$$y_i = x_{i\alpha}^T \beta_\alpha + e_i, \ i = 1, \ldots, n$$

$$\mu_{i\alpha} = E(y_i|x_i) = x_{i\alpha}^T \beta_\alpha \quad var(y_i|x_i) = \sigma_e^2$$

and the parameter is estimated using equation (2.2)

$$\hat{\beta}_\alpha = (X_\alpha^T X_\alpha)^{-1} X_\alpha^T Y$$

Then an average conditional expected loss is defined to measure the efficiency of model α

$$\Gamma_n(\alpha) = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i\alpha}^T \hat{\beta}_\alpha)^2 | Y, X\right] \tag{2.3}$$

The model $\alpha \in \mathcal{A}$ which minimizes $\Gamma_n(\alpha)$ will be chosen as the estimate of the optimal model, where $\mathcal{A}$ is the collection of some subsets of $1, \ldots, p$. The largest possible $\mathcal{A}$ is the one containing all nonempty subsets of $1, \ldots, p$. For practice, we will consider a smaller collection of subsets. The optimal model, called α_0, in this sense is the correct model with the smallest size and the smallest $\Gamma_n(\alpha)$. As $\Gamma_n(\alpha)$ involves the unknown parameter β, α_0 must be estimated by $\hat{\alpha}$.

The model selection procedure is consistent if

$$\lim_{n\to\infty} P(\hat{\alpha} = \alpha_0) = 1$$

With the modified bootstrapping pairs method, generate m ($< n$) iid pairs (x_i^*, y_i^*) from the empirical distribution $\hat{F}$ putting mass n^{-1} on (x_i, y_i), $i = 1, \ldots, n$.

Then a bootstrap estimator of $E[\Gamma_m(\alpha)]$, for $m < n$, is

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \frac{\|Y - X_\alpha \tilde{\beta}^*_{\alpha,m}\|^2}{n} \tag{2.4}$$

where $\|a\| = \sqrt{a^T a}$ for any vector a, $\tilde{\beta}^*_{\alpha,m}$ is the bootstrap analog of $\hat{\beta}_\alpha$ based on the generated m iid bootstrapping observations,

$$\tilde{\beta}^*_{\alpha,m} = \left(\sum_{i=1}^{m} x^*_{i\alpha} x^{*\prime}_{i\alpha} \right)^{-1} \sum_{i=1}^{m} x^*_{i\alpha} y^*_i$$

and E_* represents expectation approximated by Monte Carlo with size B,

$$\hat{\Gamma}^{(B)}_{n,m}(\alpha) = \frac{1}{B} \sum_{b=1}^{B} \frac{\|Y - X_\alpha \tilde{\beta}^{*b}_{\alpha,m}\|^2}{n}$$

The objective of the selection procedure is to select a model $\hat{\alpha}_{n,m} \in \mathcal{A}$ that minimizes $\hat{\Gamma}_{n,m}(\alpha)$. From the literature (Shao, 1996), this bootstrap selection procedure is consistent; that is

$$\lim_{n\to\infty} P(\hat{\alpha}_{n,m} = \alpha_0) = 1$$

provided that m satisfies $m/n \to 0$ and $m \to \infty$.

2.3 Linear Mixed Model

In this section, the modified bootstrapping selection procedure will be applied to the Linear Mixed Model (LMM). The general mixed linear model is

$$y = X\beta + Z\nu + e$$

Where y is the vector of sample observations, X and Z are known matrices, β is the unknown parameters, and ν and e are distributed independently with mean 0 and covariance matrices G and R. Here two special cases of LMM, the Fay-Herriot model and Nested-Error regression model, will be used to illustrate the modified bootstrap selection procedure.

2.3.1 Fay-Herriot Model

If the character of interest is the sample mean $\bar{y}$ of each group, $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_n)$, Fay and Herriot (1979) developed this model to estimate per-capita income for small areas (population

less than 1,000). The model can be stated as

$$\begin{aligned} \bar{y}_i &= \mu_i + e_i \\ \mu_i &= x_i^T \beta + \nu_i \quad i = 1, \ldots, n \end{aligned} \tag{2.5}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^T$ is available for each area i, β is the unknown parameters, and $e = (e_1, \ldots, e_n)^T$ and $\nu = (\nu_1, \ldots, \nu_n)^T$ are distributed independently as Normal$(0, \mathrm{D})$ and Normal$(0, \mathrm{AI})$ respectively, where $D = diag(D_1, \ldots, D_n)$ and D_i is known. Therefore, μ_i is Normal$(\mathrm{x_i^T}\beta, \mathrm{A})$ and $\bar{y}$ is Normal(μ, D) if μ is given.

For the Fay-Herriot model, the best linear unbiased estimator of μ_i is

$$\begin{aligned} \hat{\mu}_i &= x_i^T \hat{\beta} + \frac{A}{A + D_i}(\bar{y}_i - x_i^T \hat{\beta}), \ i = 1, \ldots, n \\ \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} \bar{y} \end{aligned} \tag{2.6}$$

where $V = diag(A + D_1, \ldots, A + D_n)$ and $X = col_{1 \le i \le n}(x_i^T)$.

The model selection procedure in the Fay-Herriot model is similar to that in the linear model. Consider bootstrapping pairs, let $(x_i^*, \bar{y}_i^*)$, $i = 1, \ldots, m$, be iid sample from the empirical distribution putting mass n^{-1} to each $(x_i, \bar{y}_i)$, $i = 1, \ldots, n$. And D_i^*, $i = 1, \ldots, m$, are the corresponding known variance components. Then the bootstrap analogs of β and $\hat{\mu}_i$ under model α with bootstrap sample size m are

$$\begin{aligned} \tilde{\beta}_{\alpha,m}^* &= \left(\sum_{i=1}^{m} \frac{x_{i\alpha}^* x_{i\alpha}^{*\prime}}{\hat{A}_m + D_i^*} \right)^{-1} \sum_{i=1}^{m} \frac{x_{i\alpha}^* \bar{y}_i^*}{\hat{A}_m + D_i^*} x_{i\alpha}^* \bar{y}_i^* \\ \tilde{\mu}_{i\alpha} &= x_{i\alpha}^T \tilde{\beta}_{\alpha,m}^* + \frac{\hat{A}_m}{\hat{A}_m + D_i^*}(\bar{y}_i - x_{i\alpha}^T \tilde{\beta}_{\alpha,m}^*) \end{aligned}$$

where $\hat{A}_m$ is the analog of $\hat{A}$ with bootstrap size m and $\hat{A}$ is an unbiased quadratic estimator of A, where

$$\tilde{A} = \frac{1}{n-p}\left[\sum_{i=1}^{m}\hat{\mu}_i^2 - \sum_{i=1}^{m} D_i\left(1 - x_i^T(X^TX)^{-1}x_i\right)\right]$$
$$\hat{A} = max(\tilde{A}, 1/n) \tag{2.7}$$

where $\hat{\mu}_i = \bar{y}_i - x_i^T\hat{\beta}$ and $\hat{\beta} = (X^TX)^{-1}X^T\bar{y}$.

Define

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \frac{1}{n}\sum_{i=1}^{m}(\bar{y}_i - x_{i\alpha}^T\tilde{\beta}^*_{\alpha,m})^2 \tag{2.8}$$

Or define the loss function with the random effect estimated by the bootstrap sample data

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \left\{ \frac{1}{m}\sum_{i\in I_B}\left(\bar{y}_i - x_{i\alpha}^T\tilde{\beta}^*_{\alpha,m} - \frac{\hat{A}_m}{\hat{A}_m + D_i}(\bar{y}_i - x_{i\alpha}^T\tilde{\beta}^*_{\alpha,m})\right)^2 + \right.$$
$$\left. \frac{1}{n-m}\sum_{i\notin I_B}(\bar{y}_i - x_{i\alpha}^T\tilde{\beta}^*_{\alpha,m})^2 \right\} \tag{2.9}$$

where I_B is the set of indexes of which groups are chosen as bootstrap sample. The model selected by this bootstrap procedure is $\hat{\alpha}_{n,m} \in \mathcal{A}$ that minimizes $\hat{\Gamma}_{n,m}(\alpha)$.

2.3.2 Nested-Error Regression Model

This model was proposed by Battese et al. (1988) to estimate mean acreage under a crop for counties (small area) in Iowa. Their model is given by

$$y_{ij} = x_{ij}^T\beta + \nu_i + e_{ij} \quad i = 1, \ldots, t,\ j = 1, \ldots, n_i,\ \sum_{i=1}^{t} n_i = n \tag{2.10}$$
$$\nu_i \sim N(0, \sigma_\nu^2),\ e_{ij} \sim N(0, \sigma_e^2)$$

where y_{ij} is the character of interest for the jth sampled unit in the ith sample area, $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$, β is the unknown parameters, and n_i is the number of sampled units in the ith small area. The random errors ν_i and e_{ij} are independent of each other. The variance-covariance matrix of y is $V = diag(V_1, \ldots, V_t)$ with $V_i = \sigma_e^2 I_{n_i} + \sigma_\nu^2 1_{n_i} 1_{n_i}^T$.

The generalized least squares estimator of β is the same as (2.6)

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{2.11}$$

The corresponding estimator of ν_i is

$$\hat{\nu}_i = \gamma_i(\bar{y}_i - \bar{x}_i^T \hat{\beta})$$

where $\gamma_i = \sigma_\nu^2(\sigma_\nu^2 + \sigma_e^2 n_i^{-1})^{-1}$, and $\bar{y}_i$ and $\bar{x}_i$ are the sample mean of y_{ij} and x_{ij} in the ith group.

If variance components are unknown in LMM, variance components are estimated by

Henderson's method 3 (for nested error model):

$$\begin{aligned}\hat{\sigma}_e^2 &= (n-t-k+\lambda)^{-1}\sum\sum \hat{e}_{ij}^2 \\ \tilde{\sigma}_\nu^2 &= n_*^{-1}\left[\sum\sum \hat{u}_{ij}^2 - (n-k)\hat{\sigma}_e^2\right]\end{aligned} \tag{2.12}$$

where $n_* = n - tr\left[(\bar{X}^T\bar{X})^{-1}\sum_{j=1}^t n_j^2 \bar{x}_j \bar{x}_j^T\right]$. $\lambda = 0$ if the model has no intercept term and $\lambda = 1$ otherwise. $\{\hat{e}_{ij}^2\}$ are the residuals from the ordinary least squares regression of $y_{ij} - \bar{y}_i$ on $\{x_{ij1} - \bar{x}_{i\cdot 1}, \cdots, x_{ijk} - \bar{x}_{i\cdot k}\}$ and $\{\hat{u}_{ij}\}$ are the residuals from the ordinary least squares regression of y_{ij} on $\{x_{ij1}, \cdots, x_{ijk}\}$. To avoid negative $\tilde{\sigma}_\nu^2$, we define $\hat{\sigma}_\nu^2 = max(\tilde{\sigma}_\nu^2, 1/n)$.

The model selection procedure in the nested-error regression model is a little different from previous models. First, the bootstrap will be done with the group level, instead of elements in each group. Let (x_i^*, y_i^*), $i = 1, \ldots, m$, be iid sample groups from the empirical distribution putting mass t^{-1} to each group (x_i, y_i), $i = 1, \ldots, t$, where $x_i = (x_{i1}^T, \ldots, x_{in}^T)^T$ and $y_i = (y_{i1}, \ldots, y_{in})^T$. Then the bootstrap analogs of $\hat{\beta}$ under model α with bootstrap sample size m is

$$\tilde{\beta}_{\alpha,m}^* = (\hat{X}_\alpha^{*T} V^{*-1} \hat{X}_\alpha^*)^{-1} \hat{X}_\alpha^{*T} V^{*-1} \hat{Y}^* \tag{2.13}$$

where $\hat{X}_\alpha^*$ and $\hat{Y}^*$ are the bootstrap sample data with the corresponding estimated variance components.

Define

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \frac{1}{n}\sum_{i=1}^n (y_{ij} - x_{ij}\tilde{\beta}_{\alpha,m}^*)^2 \tag{2.14}$$

Or define the loss function by considering the random effect estimated by the bootstrap

sample data:

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \left\{ \frac{1}{m} \sum_{i \in I_B} \left(y_{ij} - x_{ij,\alpha}^T \tilde{\beta}^*_{\alpha,m} - \tilde{\gamma}_i (\bar{y}_i - \bar{x}_{i\alpha}^T \tilde{\beta}^*_{\alpha,m}) \right)^2 + \frac{1}{n-m} \sum_{i \notin I_B} (y_{ij} - x_{ij,\alpha}^T \tilde{\beta}^*_{\alpha,m})^2 \right\} \tag{2.15}$$

where $\tilde{\gamma}_i = \tilde{\sigma}_\nu^2 (\tilde{\sigma}_\nu^2 + \tilde{\sigma}_e^2 n_i^{-1})^{-1}$, and $\tilde{\sigma}_\nu^2$ and $\tilde{\sigma}_e^2$ are estimated variance components under model α with bootstrap sample size m, I_B is the set of indexes of which groups are chosen as the bootstrap sample.

The alternative algorithm is to bootstrap with the element level. Since the data are correlated within a group, the transformation is needed first to reduce the correlation:

$$\begin{aligned} \widehat{y}_{ij} &= y_{ij} - \alpha_i \bar{y}_{i\bullet} \quad \widehat{x}_{ij} = x_{ij} - \alpha_i \bar{x}_{i\bullet} \\ \widehat{\nu}_i &= \nu_i - \alpha_i \nu_i \quad \widehat{e}_{ij} = e_{ij} - \alpha_i \bar{e}_{i\bullet} \end{aligned} \tag{2.16}$$

where

$$\alpha_i = 1 - \left[\sigma_e^2 / (\sigma_e^2 + n_i \sigma_\nu^2) \right]^{1/2}$$

After the transformation, the new model is a "Linear Model":

$$\widehat{y}_{ij} = \widehat{x}_{ij}^T \beta + u_{ij} \quad where\ u_{ij} \sim \text{Normal}(0, \sigma_{\text{e}}^2) \tag{2.17}$$

Then β is estimated by

$$\hat{\beta} = (\widehat{X}^T \widehat{X})^{-1} \widehat{X}^T \widehat{Y}$$

where $\widehat{X} = col_{1 \le i \le t} col_{1 \le j \le n_i} (\widehat{x}_{ij}^T)$ and $\widehat{Y} = col_{1 \le i \le t} col_{1 \le j \le n_i} (\widehat{y}_{ij})$. After the trans-

formation of the data, the bootstrap procedure of the linear model can be used to the nested-error regression model.

2.4 The Theoretical Framework

Recently, Das et al. (2004) presented a general linear mixed model

$$\mathbf{Y} = \mathbf{X}_0\beta_0 + \mathbf{Z}\mathbf{v} + \mathbf{e}_n \tag{2.18}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of observed responses, $\mathbf{X}_0, n \times p_0$, $\mathbf{Z}_{n\times q}$ are known matrices and $\mathbf{v}$ and $\mathbf{e}_n$ are independent random variables with dispersion matrices $D(\psi)$ and $R_n(\psi)$ respectively. Here $\beta_0 \in \mathbb{R}^{p}0$ and $\psi \in \mathbb{R}^k$ are fixed parameters. The mixed ANOVA model, longitudinal models including the Fay-Herriot model and the nested error regression model are special cases of the above framework.

The model selection issue here arises from potential choices of the fixed effects covariates, which are reported in the columns of the matrix $\mathbf{X}_0$. In particular, the true value of p_0 is not known. We have a collection of vectors $X_i \in \mathbb{R}^n,\ i = 1, \ldots, p$, and any candidate $\mathbf{X}$ has a sub-collection of these vectors as columns.

Let $\mathcal{S} = \{1, \ldots, p\}$ denote the set of indices of the columns of covariates, and a typical subset of $\mathcal{S}$ is given by $\boldsymbol{\alpha} = \{j_1, \ldots, j_{p_\alpha}\} \subseteq \mathcal{S}$, and has $p_\alpha \in \{0, 1, \ldots, p\}$ elements. $\mathbf{X}_{\boldsymbol{\alpha}} = [X_{j_1} : \ldots : X_{j_{p_\alpha}}] \in \mathbb{R}^n \times \mathbb{R}^{p_\alpha}$, that is, the columns of $\mathbf{X}_{\boldsymbol{\alpha}}$ are those vectors whose indices match with those of $\boldsymbol{\alpha}$. In a slight abuse of notation, we write that the model

$\boldsymbol{\alpha}$ is given by

$$\mathbf{Y} = \mathbf{X}_{\boldsymbol{\alpha}}\beta_{\boldsymbol{\alpha}} + \mathbf{Z}\mathbf{v} + \mathbf{e}_n. \tag{2.19}$$

The number and choice of candidate models may be restricted by imposing conditions on the subset $\boldsymbol{\alpha} = \{j_1, \ldots, jp_\alpha\} \subseteq \mathcal{S}$. For example, nested models may be considered as a restrictive choice of candidate models $\boldsymbol{\alpha}$. It can be seen that the true data generating process, given by (2.18), is one of the candidate models. The goal of model selection is to consistently identify the true model.

We start with a random vector $\mathbf{w} = (w_1, \ldots, w_n)$ which is a realization from a $\mathbb{R}^n$ dimensional random variable, sometimes referred to as bootstrap weights. We impose the restriction that $w_i \geq 0$, which also reinforces the notion that these are weights. For convenience, we define the diagonal matrix $\mathcal{W}$, whose diagonal entires are given by $\mathbf{w}$, that is, the i^{th} entry is w_i. When $\mathbf{w}$ has the Multinomial $(m; 1/n, \ldots, 1/n)$ distribution for some $m \leq n$, this method is identical to the m-out-of-n bootstrap. In particular, when $m = n$, we get the classical botstrap of Efron (1979). Other weights may be used as well, for example to obtain the Bayesian bootstrap.

We use these bootstrap weights in a most interesting way. We define the following inner product between two vectors x and y in $\mathbb{R}^n$:

$$< x, y >_{\mathcal{W}} = x^T \mathcal{W} y = \sum_{i=1}^{n} w_i x_i y_i.$$

Thus, this inner product is a weighted Euclidean inner product, where the weights are the bootstrap weights. Notice that we recover the usual Euclidean inner product when all the

weights are equal to 1. In general, however, $< \cdot, \cdot >_{\mathcal{W}}$ is a semi-inner product, as some of the weights can be zero. Suppose $\mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}}$ is the projection matrix for a projection on the column space of $\mathbf{X}_{\boldsymbol{\alpha}}$ in terms of this randomly weighted semi-inner product $< \cdot, \cdot >_{\mathcal{W}}$. Thus, for any vector $v \in \mathbb{R}^n$, its projection on the column space of $\mathbf{X}_{\boldsymbol{\alpha}}$ is given by $v_{\mathcal{W},\boldsymbol{\alpha}} = \mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}} v$.

Let $\mathbf{I}_n$ denote the identity matrix in $\mathbb{R}^n$. For each model $\boldsymbol{\alpha}$, we obtain the following number:

$$T_{\boldsymbol{\alpha}} = \mathrm{E}_{\mathrm{B}} ||(\mathbf{I}_n - \mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}})\mathbf{Y}||^2,$$

where E_{B} stands for the bootstrap expectation, that is, expectation conditional on the observed data. Thus, in terms of our framework, E_{B} stands for integration with respect to the random matrix $\mathcal{W}$ conditional on all other random terms.

We now develop some conditions for the bootstrap weights $\{w_1, \ldots, w_n\}$. First, we assume that all the weights are semi-exchangeable up to order 4, that is, all marginal distributions involving subsets of size 4 or below from $\{w_1, \ldots, w_n\}$ have an exchangeability property. In particular, we assume that the univariate marginals of w_i for various $i = 1, \ldots, n$ are all identical, and the two-dimensional marginal of any pair (w_i, w_j) for $i \neq j = 1 \ldots, n$ also have the same distribution.

We assume several lower order moments of $\mathbf{w}$. In particular, we reserve the notation $\mathbb{E}w_1 = \mu_w$ and $\mathbb{V}w_1 = \sigma_w^2$, and define the centered and scaled weights $W_i = \sigma_w^{-1}(w_i - \mu_w)$ for $i = 1, \ldots, n$. Our technical assumptions will involve conditions on various cross-moments, which we state later.

In terms of the centered and scaled bootstrap weights $\mathbf{W} = (W_1, \ldots, W_n)$, we now write the random matrix of bootstrap weights as $\mathcal{W} = \mu_w \mathbf{I}_n + \sigma_w W$, where $W = \mathrm{diag}(W_1, \ldots, W_n)$,

the diagonal matrix with centered and scaled weights. For use in our technical proofs, we define the matrix $U = \sigma_w \mu_w^{-1} W$. In the m-out-of-n bootstrap (moon-bootstrap) case, we have $\sigma_w^2 = n^{-1}m(1-n^{-1})$, $\mu_w = n^{-1}m$ and $c_{w11} = -n^{-2}m$.

We also define the notation $n^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T\mathbf{X}_{\boldsymbol{\alpha}} = D_{n,\boldsymbol{\alpha}}$. The inverse of $D_{n,\boldsymbol{\alpha}}$ exists if and only if $\mathbf{X}_{\boldsymbol{\alpha}}$ has full column rank, and we assume that this is the case. Our technical methodology applies to cases where $D_{n,\boldsymbol{\alpha}}$ does not have an inverse. However, this scenario results in a problem of identifiability of models, and we assume that $\mathbf{X}_{\boldsymbol{\alpha}}$ has full column rank to avoid any issue of unique identification of models.

We also define

$$\begin{aligned}
\mathbf{A}_{\boldsymbol{\alpha}} &= n^{-1}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2}\mathbf{X}_{\boldsymbol{\alpha}}^T U \mathbf{X}_{\boldsymbol{\alpha}}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \\
\mathbf{P}_{\boldsymbol{\alpha}} &= \mathbf{X}_{\boldsymbol{\alpha}}\left(\mathbf{X}_{\boldsymbol{\alpha}}^T\mathbf{X}_{\boldsymbol{\alpha}}\right)^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2}\mathbf{X}_{\boldsymbol{\alpha}}^T \\
\mathbf{B}_{\boldsymbol{\alpha}} &= U\mathbf{X}_{\boldsymbol{\alpha}}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2}\left(\mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}}\right)^{-1}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2}\mathbf{X}_{\boldsymbol{\alpha}}^T U \\
\tilde{\mathbf{P}}_{\boldsymbol{\alpha}} &= \mathbf{P}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n - U + n^{-1}\mathbf{B}_{\boldsymbol{\alpha}}\right]\mathbf{P}_{\boldsymbol{\alpha}} \\
\mathbf{E}_{\boldsymbol{\alpha}} &= \mathbf{P}_{\boldsymbol{\alpha}} - \tilde{\mathbf{P}}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n + U\right]
\end{aligned}$$

In terms of these, we have

$$\begin{aligned}
\mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}} &= \mathbf{X}_{\boldsymbol{\alpha}}\left(\mathbf{X}_{\boldsymbol{\alpha}}^T\mathcal{W}\mathbf{X}_{\boldsymbol{\alpha}}\right)^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T\mathcal{W} \\
&= \mathbf{X}_{\boldsymbol{\alpha}}\left(\mathbf{X}_{\boldsymbol{\alpha}}^T\left[\mu_w\mathbf{I}_n + \sigma_w W\right]\mathbf{X}_{\boldsymbol{\alpha}}\right)^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T\left[\mu_w\mathbf{I}_n + \sigma_w W\right]
\end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\mu_w \mathbf{X}_{\boldsymbol{\alpha}} \left(n^{-1}\mu_w \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \mathbf{X}_{\boldsymbol{\alpha}} \right)^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \left(n^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \mathbf{X}_{\boldsymbol{\alpha}} \right)^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \left(\mathbf{D}_{\boldsymbol{\alpha}} + n^{-1}\mathbf{X}_{\boldsymbol{\alpha}}^T U \mathbf{X}_{\boldsymbol{\alpha}} \right)^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \left[\mathbf{D}_{\boldsymbol{\alpha}}^{1/2} \left\{ \mathbf{I}_n + n^{-1}\mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{X}_{\boldsymbol{\alpha}}^T U \mathbf{X}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \right\} \mathbf{D}_{\boldsymbol{\alpha}}^{1/2} \right]^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \left[\mathbf{D}_{\boldsymbol{\alpha}}^{1/2} \left\{ \mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}} \right\} \mathbf{D}_{\boldsymbol{\alpha}}^{1/2} \right]^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \left[\mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}}\right]^{-1} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \left(\mathbf{I}_n - \mathbf{A}_{\boldsymbol{\alpha}} + \mathbf{A}_{\boldsymbol{\alpha}} \left[\mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}}\right]^{-1} \mathbf{A}_{\boldsymbol{\alpha}} \right) \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= n^{-1}\mathbf{X}_{\boldsymbol{\alpha}} \Big[\mathbf{D}_{\boldsymbol{\alpha}}^{-1} - \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{A}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \\
&\quad + \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{A}_{\boldsymbol{\alpha}} \left(\mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}}\right)^{-1} \mathbf{A}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \Big] \mathbf{X}_{\boldsymbol{\alpha}}^T \left[\mathbf{I}_n + U\right] \\
&= \Big[\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}} U \mathbf{P}_{\boldsymbol{\alpha}} + n^{-1}\mathbf{P}_{\boldsymbol{\alpha}} U \mathbf{X}_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \left(\mathbf{I}_n + \mathbf{A}_{\boldsymbol{\alpha}}\right)^{-1} \\
&\quad \mathbf{D}_{\boldsymbol{\alpha}}^{-1/2} \mathbf{X}_{\boldsymbol{\alpha}}^T U \mathbf{P}_{\boldsymbol{\alpha}} \Big] \left[\mathbf{I}_n + U\right] \\
&= \mathbf{P}_{\boldsymbol{\alpha}} \left[\mathbf{I}_n - U + n^{-1}\mathbf{B}_{\boldsymbol{\alpha}} \right] \mathbf{P}_{\boldsymbol{\alpha}} \left[\mathbf{I}_n + U\right] \\
&= \tilde{\mathbf{P}}_{\boldsymbol{\alpha}} \left[\mathbf{I}_n + U\right].
\end{aligned}
$$

Using this, we have

$$
\begin{aligned}
\mathbf{I}_n - \mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}} &= \mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}} + \mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}} \\
&= \mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}} + \mathbf{P}_{\boldsymbol{\alpha}} - \tilde{\mathbf{P}}_{\boldsymbol{\alpha}} \left[\mathbf{I}_n + U\right] \\
&= \mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}} + \mathbf{E}_{\boldsymbol{\alpha}}.
\end{aligned}
$$

We now analyze the last term in greater detail. We have

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{\alpha}} &= \mathbf{P}_{\boldsymbol{\alpha}} - \tilde{\mathbf{P}}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n + U\right] \\
&= \mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n - U + n^{-1}\mathbf{B}_{\boldsymbol{\alpha}}\right]\mathbf{P}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n + U\right] \\
&= \mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n - U + n^{-1}\mathbf{B}_{\boldsymbol{\alpha}}\right]\mathbf{P}_{\boldsymbol{\alpha}} - \tilde{\mathbf{P}}_{\boldsymbol{\alpha}}U \\
&= \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}} - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}} - \tilde{\mathbf{P}}_{\boldsymbol{\alpha}}U \\
&= \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}} - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}}\left[\mathbf{I}_n - U + n^{-1}\mathbf{B}_{\boldsymbol{\alpha}}\right]\mathbf{P}_{\boldsymbol{\alpha}}U \\
&= \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}}U + \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}}U - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}} - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}}U \\
&= \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\boldsymbol{\alpha}}(\mathbf{I}_n - U)\mathbf{P}_{\boldsymbol{\alpha}}U - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}}(\mathbf{I}_n + U) \\
&= -\mathbf{P}_{\boldsymbol{\alpha}}U(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}}) + \mathbf{P}_{\boldsymbol{\alpha}}U\mathbf{P}_{\boldsymbol{\alpha}}U - n^{-1}\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{B}_{\boldsymbol{\alpha}}\mathbf{P}_{\boldsymbol{\alpha}}(\mathbf{I}_n + U).
\end{aligned}$$

Since $\mathbf{E}_{\boldsymbol{\alpha}}$ is $\mathbf{P}_{\boldsymbol{\alpha}}$ times another matrix, we have $(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}})\mathbf{E}_{\boldsymbol{\alpha}} = 0$. Using this, we have

$$\begin{aligned}
T_{\boldsymbol{\alpha}} &= \mathrm{E}_{\mathrm{B}}||(\mathbf{I}_n - \mathbf{P}_{\mathcal{W},\boldsymbol{\alpha}})\mathbf{Y}||^2 \\
&= \mathrm{E}_{\mathrm{B}}\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}} + \mathbf{E}_{\boldsymbol{\alpha}})^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}} + \mathbf{E}_{\boldsymbol{\alpha}})\mathbf{Y} \\
&= Y^T(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}})\mathbf{Y} + \mathrm{E}_{\mathrm{B}}\mathbf{Y}^T\mathbf{E}_{\boldsymbol{\alpha}}^T\mathbf{E}_{\boldsymbol{\alpha}}\mathbf{Y}.
\end{aligned}$$

Thus, we have a very neat decomposition of $T_{\boldsymbol{\alpha}}$, with the first term capturing the squared residuals in model $\boldsymbol{\alpha}$, and the other term containing all the bootstrap related quantities.

Our next task is to compute $\mathrm{E}_{\mathrm{B}}(\mathbf{E}_{\boldsymbol{\alpha}}^T\mathbf{E}_{\boldsymbol{\alpha}})$. Note that all the bootstrap weights are in the matrix U. We have

$$\mathbf{E}_{\boldsymbol{\alpha}}^T\mathbf{E}_{\boldsymbol{\alpha}} = (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}})U\mathbf{P}_{\boldsymbol{\alpha}}U(\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}}) + \mathbf{E}_{R,\boldsymbol{\alpha}},$$

where $\mathbf{E}_{R,\boldsymbol{\alpha}}$ is a symmetric matrix, which is entirely negligible. The algebra for showing the different components of $\mathbf{E}_{R,\boldsymbol{\alpha}}$ is long and tedious, and it also is the source of much of our technical assumptions, so we skip the details here.

Consequently, we now need to evaluate

$$\begin{aligned}
\mathbf{V}_{\boldsymbol{\alpha}} &= \mathrm{E}_{\mathrm{B}} U \mathbf{P}_{\boldsymbol{\alpha}} U \\
&= ((\mathrm{E}_{\mathrm{B}} U_i U_j \mathbf{P}_{\boldsymbol{\alpha},ij})) \\
&= \begin{cases} \sigma_w^2 \mu_w^{-2} \mathbf{P}_{\boldsymbol{\alpha},ii} & \text{if } i = j \\ c_{w11} \sigma_w^2 \mu_w^{-2} \mathbf{P}_{\boldsymbol{\alpha},ij} & \text{if } i \neq j \end{cases}
\end{aligned}$$

For use later on, we define the diagonal matrix

$$\mathbf{G}_{\boldsymbol{\alpha}} = (1 - c_{w11}) \sigma_w^2 \mu_w^{-2} \mathrm{diag}(\mathbf{P}_{\boldsymbol{\alpha},ii}).$$

In terms of the subsequent analysis, the difference between $\mathbf{V}_{\boldsymbol{\alpha}}$ and $\mathbf{G}_{\boldsymbol{\alpha}}$ is negligible. Thus, the properties of $T_{\boldsymbol{\alpha}}$ are governed by

$$Z_{\boldsymbol{\alpha}} = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}}) [\mathbf{I}_n + \mathbf{G}_{\boldsymbol{\alpha}}] (\mathbf{I}_n - \mathbf{P}_{\boldsymbol{\alpha}}) \mathbf{Y}.$$

The difference between $T_{\boldsymbol{\alpha}}$ and $Z_{\boldsymbol{\alpha}}$ is negligible, although showing this involves a lot of algebra.

There is a $Z_{\boldsymbol{\alpha}}$ value corresponding to each model $\boldsymbol{\alpha}$. Also note that, at least in large samples, $Z_{\boldsymbol{\alpha}}$ is positive, which is very consistent with the fact that $T_{\boldsymbol{\alpha}}$ is always positive. We compare the various $T_{\boldsymbol{\alpha}}$'s, and establish that with probability tending towards one the true model has the smallest value. This is again a lot of algebra, but the main scheme for proving

this uses the moment-generating function for $T_{\boldsymbol{\alpha}}$. The results follow from a calculation based on derivatives of determinants of matrices like $A + tB$.

2.5 Simulation Study

To examine the finite-sample performance of the selection procedures, a simulation study was carried out based on bootstrapping pairs with different bootstrap sampling size m and various model settings. First, the similar results of the linear model from Shao's paper were replicated. The same data set used by Shao from the solid waste data example of Gunst and Mason (1980) was used again; and the same value of β, $(2, 9, 6, 4, 8)$, is chosen. Then the Fay-Herriot model was tested with the same data set. After that, two different selection procedures of the nested-error model based on the same data set were carried out.

2.5.1 Algorithms and Settings for Different Models

For the linear model, the size of the data set, n, is 40 and e_i, $i = 1, \ldots, n$, is iid standard normal errors. The algorithm is:

1. Generate data, Y, according to the defined linear model (2.1).
2. Bootstrap one sample with size $m = 15$ from the original data.
3. For each candidate model: (1) estimate the unknown parameters by (2.2); (2) calculate the loss function in (2.3).
4. Repeat step 2&3 $B = 100$ times. Choose the candidate model with the minimized mean of loss functions as the optimal model. Get the 95% confidence interval of fixed effect parameters from the 2.5% and 97.5% quantiles of B bootstrap samples.

5. Repeat steps 2, 3 & 4 with different bootstrap samples of size $m = 20, 25, 30, 40$.

6. Record the optimal model and corresponding confidence intervals. Repeat steps 1-5 1000 times.

7. Repeat the whole process with different true model settings.

For the Fay-Herriot model, the same x is used as the linear model and $n = 40$; for the variance components, chose $A = 1$ and $D = diag(0.7I_8, 0.6I_8, 0.5I_8, 0.4I_8, 0.3I_8)$, I_8 means an 8×8 identical matrix. The algorithm is:

1. Generate data, Y, based on the model (2.5).

2. Bootstrap one sample with size $m = 15$ from the original data.

3. For each candidate model: (1) estimate variance components A using (2.7); (2) estimate the fixed effect parameters with $\hat{A}$ using (2.6); (3) calculate the loss function using (2.8) and (2.9).

4. Repeat step 2&3 $B = 100$ times. Choose the candidate model with the minimized mean of loss functions as the optimal model. Get the 95% confidence interval of fixed effect parameters from the 2.5% and 97.5% quantiles of B bootstrap samples.

5. Repeat steps 2, 3 & 4 with different bootstrap samples of size $m = 20, 25, 30, 40$.

6. Record the optimal model and corresponding confidence intervals. Repeat steps 1-5 1000 times.

7. Repeat the whole process with different true model settings.

For the nested-error regression model, we use the same data set with size $n = 40$, divided into $t = 10$ groups with $n_i = (5, 6, 4, 5, 3, 2, 5, 3, 3, 4)$; choose the variance components as $\sigma_\nu^2 = 1$ and $\sigma_e^2 = 1$. The algorithm of the bootstrap with group level is:

1. Generate data, Y, based on model (2.10).

2. Bootstrap one sample of $m = 4$ groups from the original 10 groups of the data.

3. For each candidate model: (1) estimate variance components by Henderson's method 3 in (2.12); (2) estimate the fixed effect parameters based on $\hat{\sigma}_\nu^2$ and $\hat{\sigma}_e^2$ using (2.11) or (2.13); (3) calculate the loss function using (2.14) and (2.15).

4. Repeat steps 2&3 $B = 100$ times. Choose the candidate model with the minimized mean of loss functions as the optimal model. Get the 95% confidence interval of fixed effect parameters from the 2.5% and 97.5% quantiles of B bootstrap samples.

5. Repeat steps 2, 3 & 4 with different bootstrap samples of size $m = 5, 6, 8, 10$.

6. Record the optimal model and corresponding confidence intervals. Repeat steps 1-5 1000 times.

7. Repeat the whole process with different true model settings.

The algorithm of the bootstrap with element level of Nested-Error Regression Model:

1. Generate data, Y, based on model (2.10).

2. Estimate variance components with the whole data by Henderson's method 3 in (2.12).

3. Transfer the original data to linearized data which can be fitted by a linear model using the tranformation in (2.16).

4. Bootstrap one sample with size $m = 15$ from the new data.

5. For each candidate model: estimate the fixed effect parameters like a linear model and calculate the corresponding loss function.

6. Repeat steps 4&5 $B = 100$ times. Choose the candidate model with the minimized mean of loss functions as the optimal model. Get the 95% confidence interval of fixed effect parameters from the 2.5% and 97.5% quantiles of B bootstrap samples.

7. Repeat steps 4, 5 & 6 with different bootstrap samples of size $m = 20, 25, 30, 40$.

8. Record the optimal model and corresponding confidence intervals. Repeat steps 1-7 1000 times.

9. Repeat the whole process with different true model settings.

The bootstrap estimators $\hat{\Gamma}_{n,m}(\alpha)$ were computed with size $B = 100$. The empirical probabilities of selecting each model and the coverage probabilities of the true parameter value were based on 1,000 simulations. The lengths of the confidence interval were means of the 1,000 replications. For comparison, the results of empirical selection methods using AIC, BIC and adjusted R^2 were also reported. The whole procedure was programmed in R.

2.5.2 Results

At first the results of the Linear Model are reported. Table 2.1 is the selection probabilities of the optimal model; Table 2.2 is the coverage probabilities of each parameter and the lengths of the corresponding confidence intervals. From the results, the similar characters of this bootstrap selection procedure as Shao's results are supported. The consistency of the

modified bootstrap selection procedure is supported clearly. Also, the modified bootstrap selection procedure can be substantially better than other selection methods in most cases.

After that the results of the Fay-Herriot model are reported in Table 2.3 $\sim$ 2.6. From the results in the Tables 2.3 and 2.4, the application of this bootstrap selection procedure working with the Fay-Herriot model is supported. In Tables 2.5 and 2.6, the different loss functions with or without considering the random effects are compared. The selection probabilities of the loss function adding the random effects are higher than that of the loss function without the random effects in some cases with higher bootstrap sample size. The coverage probabilities do not have a significant difference.

Tables 2.7 and 2.8 report the empirical selecting probabilities of the Nested-Error regression model working on the group level. Tables 2.9 and 2.10 report the results of working on the element level. From the results, the selection probabilities of working on the element level are higher in most cases, as the process of bootstrapping with the element level after the transformation is applied is closer to the process of the Linear model. But the data size is only 40 in these cases and especially the number of groups is only 10, maybe this is part of the reason why the bootstrapping with group level is not as good as the work with element level. After that, the two different definitions of loss functions are compared in Tables 2.11 and 2.12 with a bigger data size $n = 160$ to test the effect of increasing data size. Similar to the case of the Fay-Herriot model, the probabilities of the function with random effect are higher when the bootstrap sample size is bigger.

In addition, except the case when the full model is the optimal model, the modified bootstrap model selection procedure with smaller bootstrap size m works better than the unmodified bootstrap model selection procedure, when bootstrap with size 40. From the

trend of probability changing, the optimal choice of m clearly depends on the parameter β. When the size of parameters of the optimal model is small, the modified bootstrap selection procedure with smaller bootstrap size is better than the procedure with bigger bootstrap size. If the size of parameters of the optimal model increases, the optimal choice of bootstrap size m will increase. This changing trend is tested in more detail by more bootstraps of sizes $m = 11, 12, 13, 14, 15, 20, 25, 30, 40$. The result is not listed here.

The empirical probabilities of the coverage of the confidence interval are the probabilities of the bootstrap confidence interval covering the true parameter values. From the results, the 95% confidence interval from the modified bootstrap selection procedure will cover the true parameter values with more than probability 0.9 in most cases. The coverage probabilities with smaller bootstrap sample size are higher, but the lengths of the confidence intervals with bigger bootstrap sample size are shorter.

In conclusion, the modified bootstrap model selection procedure can be applied to non-linear model cases, like linear mixed models. Bootstrapping pairs with size m works well in these models, where $m/n \to 0$ and $m \to \infty$.

Table 2.1: Selection probabilities for linear model

True β^T	Model	Bootstrap					AIC	BIC	R^2
		m=15	m=20	m=25	m=30	m=40			
(2,0,0,4,0)	1,4*	0.957	0.851	0.747	0.654	0.491	0.576	0.818	0.315
	1,2,4	0.016	0.046	0.079	0.101	0.134	0.094	0.050	0.116
	1,3,4	0.018	0.066	0.097	0.128	0.192	0.086	0.040	0.105
	1,4,5	0.008	0.027	0.055	0.072	0.091	0.122	0.062	0.163
	1,2,3,4	0.001	0.007	0.006	0.018	0.034	0.051	0.017	0.114
	1,2,4,5	0.000	0.001	0.005	0.008	0.019	0.027	0.005	0.076
	1,3,4,5	0.000	0.002	0.011	0.013	0.023	0.024	0.004	0.053
	1,2,3,4,5	0.000	0.000	0.000	0.006	0.016	0.020	0.004	0.058
(2,0,0,4,8)	1,4,5*	0.958	0.897	0.828	0.770	0.575	0.732	0.893	0.515
	1,2,4,5	0.021	0.047	0.067	0.100	0.153	0.098	0.051	0.163
	1,3,4,5	0.020	0.054	0.096	0.103	0.201	0.098	0.042	0.161
	1,2,3,4,5	0.001	0.002	0.009	0.027	0.071	0.072	0.014	0.161
(2,9,0,4,8)	1,4,5	0.012	0.003	0.000	0.001	0.000	0.000	0.000	0.000
	1,2,4,5*	0.973	0.961	0.904	0.863	0.741	0.811	0.931	0.673
	1,3,4,5	0.002	0.002	0.001	0.000	0.000	0.001	0.001	0.001
	1,2,3,4,5	0.013	0.034	0.095	0.136	0.259	0.188	0.068	0.326
(2,9,6,4,8)	1,2,3,5	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1,3,4,5	0.069	0.023	0.009	0.001	0.000	0.000	0.003	0.000
	1,2,3,4,5*	0.914	0.976	0.991	0.999	1.000	1.000	0.997	1.000

Table 2.2: The coverage probabilities and lengths of confidence intervals for linear model

True β^T	Para	m=15	m=20	m=25	m=30	m=40	Non Bootstrap
(2,0,0,4,0)	σ_e^2	0.974(1.320)	0.935(1.135)	0.922(0.989)	0.883(0.905)	0.810(0.759)	0.920(0.902)
	β_1	0.995(1.427)	0.983(1.201)	0.968(1.060)	0.957(0.958)	0.905(0.819)	0.948(0.855)
	β_4	0.999(2.106)	0.974(1.751)	0.945(1.579)	0.924(1.481)	0.841(1.397)	0.948(0.966)
(2,0,0,4,8)	σ_e^2	0.980(1.372)	0.950(1.167)	0.918(1.012)	0.884(0.912)	0.840(0.779)	0.933(0.918)
	β_1	0.996(1.500)	0.990(1.252)	0.970(1.089)	0.952(0.983)	0.927(0.838)	0.955(0.864)
	β_4	0.996(4.348)	0.985(3.431)	0.980(2.909)	0.944(2.556)	0.912(2.108)	0.944(1.700)
	β_5	0.998(4.799)	0.994(3.769)	0.978(3.164)	0.962(2.768)	0.915(2.269)	0.964(2.119)
(2,9,0,4,8)	σ_e^2	0.972(1.403)	0.940(1.178)	0.914(1.023)	0.885(0.922)	0.831(0.786)	0.927(0.927)
	β_1	0.994(1.576)	0.989(1.300)	0.971(1.116)	0.959(1.005)	0.932(0.852)	0.937(0.863)
	β_2	0.996(14.015)	0.990(10.255)	0.974(8.075)	0.958(6.867)	0.914(5.353)	0.951(4.108)
	β_4	0.999(6.292)	0.993(4.799)	0.986(3.890)	0.975(3.384)	0.937(2.632)	0.936(2.153)
	β_5	0.999(5.416)	0.993(4.072)	0.973(3.351)	0.971(2.945)	0.919(2.367)	0.946(2.151)
(2,9,6,4,8)	σ_e^2	0.971(1.555)	0.920(1.187)	0.907(1.050)	0.873(0.916)	0.814(0.791)	0.927(0.939)
	β_1	0.999(1.738)	0.988(1.365)	0.967(1.175)	0.959(1.029)	0.935(0.865)	0.940(0.873)
	β_2	0.909(17.204)	0.986(13.43)	0.973(10.983)	0.962(9.297)	0.931(7.400)	0.944(7.022)
	β_3	0.982(10.175)	0.988(7.186)	0.976(5.656)	0.971(4.735)	0.940(3.661)	0.957(3.277)
	β_4	0.978(6.962)	0.997(5.174)	0.983(4.196)	0.980(3.531)	0.939(2.750)	0.944(2.286)
	β_5	0.997(5.958)	0.995(4.365)	0.982(3.588)	0.970(3.061)	0.931(2.448)	0.945(2.203)

Table 2.3: Selection probabilities for Fay-Herriot model, loss function without random effect

True β^T	Model	Bootstrap					AIC	BIC	R^2
		m=15	m=20	m=25	m=30	m=40			
(2,0,0,4,0)	1,4*	0.966	0.902	0.812	0.673	0.487	0.558	0.806	0.326
	1,2,4	0.006	0.033	0.057	0.088	0.126	0.107	0.061	0.141
	1,3,4	0.003	0.013	0.025	0.057	0.084	0.092	0.057	0.108
	1,4,5	0.025	0.049	0.090	0.123	0.145	0.118	0.051	0.148
	1,2,3,4	0.000	0.001	0.009	0.025	0.072	0.063	0.012	0.108
	1,2,4,5	0.000	0.001	0.006	0.020	0.038	0.022	0.006	0.048
	1,3,4,5	0.000	0.001	0.001	0.010	0.031	0.019	0.005	0.059
	1,2,3,4,5	0.000	0.000	0.000	0.004	0.017	0.021	0.002	0.062
(2,0,0,4,8)	1,4,5*	0.995	0.966	0.901	0.822	0.660	0.688	0.857	0.498
	1,2,4,5	0.003	0.026	0.060	0.090	0.147	0.138	0.073	0.188
	1,3,4,5	0.002	0.008	0.032	0.063	0.117	0.111	0.060	0.174
	1,2,3,4,5	0.000	0.000	0.007	0.025	0.076	0.063	0.010	0.140
(2,9,0,4,8)	1,4,5	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5*	0.981	0.968	0.949	0.888	0.802	0.848	0.931	0.710
	1,3,4,5	0.015	0.008	0.007	0.003	0.002	0.004	0.009	0.003
	1,2,3,4,5	0.002	0.024	0.044	0.109	0.196	0.148	0.060	0.287
(2,9,6,4,8)	1,2,3,5	0.065	0.017	0.003	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.042	0.007	0.001	0.000	0.000	0.000	0.000	0.000
	1,3,4,5	0.203	0.069	0.029	0.013	0.002	0.000	0.009	0.000
	1,2,3,4,5*	0.690	0.907	0.967	0.987	0.998	1.000	0.991	1.000

Table 2.4: The coverage probabilities and lengths of confidence intervals for the Fay-Herriot model, loss function without random effect

True β^T	Para	m=15	m=20	m=25	m=30	m=40	Non Bootstrap
(2,0,0,4,0)	σ_e^2	0.967(1.927)	0.934(1.666)	0.919(1.476)	0.876(1.333)	0.820(1.145)	0.641(0.656)
	β_1	0.997(1.748)	0.988(1.473)	0.970(1.293)	0.960(1.160)	0.927(0.994)	0.944(1.042)
	β_4	0.992(2.570)	0.973(2.089)	0.956(1.878)	0.926(1.838)	0.853(1.719)	0.939(1.199)
(2,0,0,4,8)	σ_e^2	0.972(1.960)	0.940(1.696)	0.917(1.490)	0.876(1.349)	0.816(1.149)	0.625(0.658)
	β_1	0.996(1.837)	0.990(1.526)	0.976(1.332)	0.963(1.192)	0.926(1.014)	0.939(1.048)
	β_4	0.997(5.288)	0.991(4.232)	0.973(3.622)	0.959(3.197)	0.920(2.634)	0.946(2.140)
	β_5	0.998(5.861)	0.991(4.607)	0.979(3.890)	0.970(3.411)	0.934(2.788)	0.946(2.582)
(2,9,0,4,8)	σ_e^2	0.965(1.990)	0.930(1.711)	0.909(1.506)	0.871(1.356)	0.817(1.162)	0.656(0.667)
	β_1	0.997(1.936)	0.991(1.583)	0.980(1.364)	0.963(1.220)	0.927(1.028)	0.947(1.057)
	β_2	0.983(16.619)	0.984(12.204)	0.967(9.712)	0.954(8.265)	0.895(6.491)	0.950(5.209)
	β_4	0.998(7.626)	0.998(5.829)	0.983(4.755)	0.979(4.095)	0.948(3.208)	0.954(2.738)
	β_5	0.998(6.525)	0.991(4.982)	0.981(4.136)	0.967(3.574)	0.932(2.882)	0.950(2.639)
(2,9,6,4,8)	σ_e^2	0.968(2.219)	0.927(1.757)	0.891(1.519)	0.858(1.367)	0.788(1.164)	0.615(0.652)
	β_1	0.997(2.084)	0.992(1.657)	0.985(1.416)	0.969(1.251)	0.936(1.047)	0.938(1.054)
	β_2	0.787(17.268)	0.919(15.283)	0.957(12.815)	0.963(11.158)	0.926(8.877)	0.938(8.329)
	β_3	0.935(11.206)	0.961(8.563)	0.968(6.800)	0.959(5.696)	0.924(4.379)	0.927(3.847)
	β_4	0.924(7.675)	0.968(6.094)	0.974(5.038)	0.978(4.275)	0.947(3.332)	0.936(2.819)
	β_5	0.999(7.121)	0.995(5.327)	0.982(4.348)	0.974(3.719)	0.938(2.949)	0.942(2.650)

Table 2.5: Selection probabilities for Fay-Herriot model, loss function with random effect

True β^T	Model	m=15	m=20	m=25	m=30	m=40
(2,0,0,4,0)	1,4*	0.976	0.951	0.910	0.872	0.806
	1,2,4	0.002	0.009	0.022	0.038	0.054
	1,3,4	0.002	0.007	0.018	0.021	0.039
	1,4,5	0.020	0.033	0.048	0.066	0.087
	1,2,3,4	0.000	0.000	0.001	0.002	0.005
	1,2,4,5	0.000	0.000	0.001	0.001	0.007
	1,3,4,5	0.000	0.000	0.000	0.000	0.001
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.001
(2,0,0,4,8)	1,4,5*	0.993	0.988	0.964	0.940	0.898
	1,2,4,5	0.005	0.008	0.025	0.032	0.051
	1,3,4,5	0.002	0.003	0.011	0.025	0.045
	1,2,3,4,5	0.000	0.001	0.000	0.003	0.006
(2,9,0,4,8)	1,4,5	0.010	0.001	0.001	0.000	0.000
	1,2,4,5*	0.972	0.972	0.954	0.940	0.912
	1,3,4,5	0.017	0.011	0.017	0.012	0.008
	1,2,3,4,5	0.001	0.016	0.028	0.048	0.080
(2,9,6,4,8)	1,2,3,5	0.080	0.031	0.009	0.005	0.001
	1,2,4,5	0.051	0.016	0.007	0.002	0.003
	1,3,4,5	0.246	0.120	0.064	0.057	0.035
	1,2,3,4,5*	0.623	0.833	0.920	0.936	0.961

Table 2.6: The coverage probabilities and lengths of confidence intervals for the Fay-Herriot model, loss function with random effect

True β^T	Para	m=15	m=20	m=25	m=30	m=40
(2,0,0,4,0)	σ_e^2	0.966(1.891)	0.940(1.637)	0.917(1.458)	0.888(1.331)	0.834(1.134)
	β_1	0.996(1.738)	0.983(1.463)	0.969(1.284)	0.960(1.162)	0.910(0.991)
	β_4	0.993(2.517)	0.978(1.979)	0.952(1.716)	0.919(1.570)	0.862(1.331)
(2,0,0,4,8)	σ_e^2	0.976(1.926)	0.943(1.665)	0.914(1.476)	0.884(1.337)	0.817(1.141)
	β_1	0.994(1.835)	0.990(1.518)	0.978(1.326)	0.963(1.199)	0.934(1.015)
	β_4	0.998(5.279)	0.983(4.182)	0.965(3.511)	0.948(3.105)	0.900(2.515)
	β_5	0.998(5.775)	0.985(4.586)	0.972(3.858)	0.957(3.385)	0.924(2.777)
(2,9,0,4,8)	σ_e^2	0.970(2.021)	0.944(1.725)	0.910(1.529)	0.890(1.378)	0.829(1.171)
	β_1	0.998(1.938)	0.983(1.567)	0.972(1.369)	0.948(1.215)	0.910(1.032)
	β_2	0.973(16.438)	0.983(12.087)	0.965(9.504)	0.941(8.000)	0.884(6.120)
	β_4	0.995(7.678)	0.989(5.802)	0.976(4.762)	0.950(4.070)	0.913(3.189)
	β_5	0.998(6.587)	0.995(5.021)	0.985(4.143)	0.962(3.586)	0.921(2.886)
(2,9,6,4,8)	σ_e^2	0.963(2.267)	0.939(1.802)	0.901(1.546)	0.862(1.384)	0.790(1.172)
	β_1	0.994(2.099)	0.986(1.676)	0.974(1.422)	0.968(1.262)	0.928(1.054)
	β_2	0.729(16.369)	0.863(14.225)	0.918(12.238)	0.911(10.559)	0.892(8.551)
	β_3	0.929(11.030)	0.927(8.344)	0.942(6.633)	0.924(5.612)	0.903(4.316)
	β_4	0.908(7.533)	0.943(5.982)	0.968(4.967)	0.950(4.235)	0.916(3.312)
	β_5	0.994(7.162)	0.983(5.330)	0.979(4.347)	0.960(3.731)	0.934(2.976)

Table 2.7: Selection probabilities for Nested-Error regression model with group level

True β^T	Model	Bootstrap					AIC	BIC	R^2
		m=4	m=5	m=6	m=8	m=10			
(2,0,0,4,0)	1,4*	0.933	0.870	0.788	0.621	0.498	0.471	0.724	0.315
	1,2,4	0.017	0.032	0.049	0.090	0.103	0.124	0.072	0.129
	1,3,4	0.022	0.038	0.052	0.079	0.089	0.096	0.055	0.102
	1,4,5	0.026	0.048	0.072	0.122	0.146	0.142	0.090	0.152
	1,2,3,4	0.002	0.006	0.021	0.040	0.074	0.074	0.033	0.117
	1,2,4,5	0.000	0.001	0.008	0.021	0.029	0.022	0.006	0.046
	1,3,4,5	0.000	0.005	0.008	0.021	0.036	0.037	0.010	0.061
	1,2,3,4,5	0.000	0.000	0.002	0.006	0.025	0.034	0.010	0.078
(2,0,0,4,8)	1,4,5*	0.895	0.881	0.838	0.745	0.627	0.634	0.825	0.468
	1,2,4,5	0.054	0.053	0.071	0.096	0.141	0.126	0.073	0.164
	1,3,4,5	0.043	0.057	0.070	0.110	0.141	0.114	0.066	0.164
	1,2,3,4,5	0.009	0.010	0.022	0.049	0.090	0.126	0.036	0.204
(2,9,0,4,8)	1,4,5	0.021	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5*	0.782	0.811	0.822	0.800	0.730	0.746	0.888	0.655
	1,3,4,5	0.152	0.104	0.062	0.027	0.014	0.005	0.010	0.001
	1,2,3,4,5	0.044	0.084	0.116	0.173	0.256	0.249	0.102	0.344
(2,9,6,4,8)	1,2,3,5	0.102	0.022	0.006	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.066	0.031	0.011	0.001	0.000	0.000	0.002	0.000
	1,3,4,5	0.509	0.291	0.131	0.029	0.005	0.005	0.015	0.003
	1,2,3,4,5*	0.323	0.656	0.852	0.970	0.995	0.995	0.983	0.997

Table 2.8: The coverage probabilities and lengths of confidence intervals for for Nested-Error regression model with group level

True β^T	Para	m=4	m=5	m=6	m=8	m=10	Non Bootstrap
(2,0,0,4,0)	σ_v^2	0.914(2.673)	0.883(2.326)	0.855(2.105)	0.804(1.810)	0.768(1.597)	0.943(5.833)
	σ_e^2	0.964(1.408)	0.935(1.228)	0.917(1.116)	0.871(0.942)	0.826(0.841)	0.919(1.035)
	β_1	0.983(2.338)	0.971(2.044)	0.952(1.842)	0.928(1.570)	0.890(1.398)	0.918(1.515)
	β_4	0.982(3.145)	0.958(2.473)	0.930(2.112)	0.885(1.795)	0.824(1.661)	0.938(1.070)
(2,0,0,4,8)	σ_v^2	0.926(2.757)	0.881(2.345)	0.851(2.100)	0.794(1.790)	0.760(1.570)	0.934(5.781)
	σ_e^2	0.946(1.476)	0.917(1.279)	0.886(1.144)	0.838(0.957)	0.797(0.844)	0.908(1.044)
	β_1	0.981(2.535)	0.968(2.172)	0.951(1.918)	0.913(1.618)	0.879(1.426)	0.916(1.522)
	β_4	0.983(5.079)	0.973(4.255)	0.961(3.742)	0.919(3.037)	0.891(2.593)	0.934(1.814)
	β_5	0.992(6.799)	0.985(5.003)	0.974(4.080)	0.939(3.197)	0.921(2.722)	0.949(2.421)
(2,9,0,4,8)	σ_v^2	0.953(2.896)	0.903(2.406)	0.863(2.118)	0.814(1.794)	0.767(1.566)	0.929(5.598)
	σ_e^2	0.952(1.743)	0.924(1.402)	0.896(1.210)	0.843(1.002)	0.794(0.875)	0.921(1.081)
	β_1	0.989(2.822)	0.979(2.277)	0.953(2.012)	0.923(1.664)	0.895(1.454)	0.920(1.511)
	β_2	0.820(13.957)	0.885(11.292)	0.912(9.585)	0.897(7.139)	0.852(5.980)	0.943(4.417)
	β_4	0.950(6.866)	0.971(5.705)	0.962(4.929)	0.945(3.850)	0.911(3.242)	0.943(2.336)
	β_5	0.993(7.922)	0.984(5.554)	0.976(4.452)	0.937(3.304)	0.894(2.830)	0.928(2.474)
(2,9,6,4,8)	σ_v^2	0.968(3.220)	0.917(2.499)	0.862(2.123)	0.785(1.728)	0.736(1.514)	0.936(5.835)
	σ_e^2	0.964(2.110)	0.907(1.580)	0.879(1.283)	0.824(1.012)	0.793(0.874)	0.914(1.086)
	β_1	0.994(2.964)	0.972(2.378)	0.963(2.051)	0.924(1.678)	0.889(1.457)	0.928(1.527)
	β_2	0.438(11.579)	0.680(12.512)	0.847(12.343)	0.926(10.322)	0.915(8.705)	0.932(8.088)
	β_3	0.813(9.159)	0.835(7.749)	0.889(6.808)	0.928(5.367)	0.919(4.514)	0.929(3.951)
	β_4	0.794(6.134)	0.889(5.724)	0.918(5.057)	0.945(4.041)	0.922(3.367)	0.940(2.530)
	β_5	0.982(9.033)	0.987(6.423)	0.972(4.956)	0.944(3.642)	0.915(3.025)	0.937(2.660)

Table 2.9: Selection probabilities for Nested-Error regression model with element level

True β^T	Model	Bootstrap					AIC	BIC	R^2
		m=15	m=20	m=25	m=30	m=40			
(2,0,0,4,0)	1,4*	0.964	0.860	0.741	0.630	0.443	0.552	0.825	0.306
	1,2,4	0.010	0.023	0.049	0.066	0.090	0.100	0.051	0.114
	1,3,4	0.006	0.029	0.048	0.069	0.097	0.080	0.036	0.097
	1,4,5	0.020	0.077	0.118	0.148	0.183	0.125	0.056	0.170
	1,2,3,4	0.000	0.007	0.027	0.048	0.093	0.069	0.016	0.118
	1,2,4,5	0.000	0.002	0.007	0.011	0.024	0.023	0.007	0.050
	1,3,4,5	0.000	0.001	0.008	0.020	0.045	0.029	0.007	0.071
	1,2,3,4,5	0.000	0.001	0.002	0.008	0.025	0.022	0.002	0.074
(2,0,0,4,8)	1,4,5*	0.990	0.961	0.891	0.814	0.631	0.690	0.879	0.478
	1,2,4,5	0.010	0.022	0.044	0.072	0.131	0.104	0.043	0.160
	1,3,4,5	0.000	0.013	0.052	0.079	0.137	0.104	0.050	0.153
	1,2,3,4,5	0.000	0.004	0.013	0.035	0.101	0.102	0.028	0.209
(2,9,0,4,8)	1,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5*	0.968	0.943	0.909	0.852	0.750	0.813	0.908	0.688
	1,3,4,5	0.024	0.014	0.011	0.007	0.002	0.001	0.006	0.000
	1,2,3,4,5	0.008	0.043	0.080	0.141	0.248	0.186	0.086	0.312
(2,9,6,4,8)	1,2,3,5	0.018	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.010	0.004	0.000	0.001	0.000	0.000	0.001	0.000
	1,3,4,5	0.140	0.036	0.017	0.009	0.002	0.002	0.011	0.000
	1,2,3,4,5*	0.832	0.959	0.983	0.990	0.998	0.998	0.988	1.000

Table 2.10: The coverage probabilities and lengths of confidence intervals for Nested-Error regression model with element level

True β^T	Para	m=15	m=20	m=25	m=30	m=40	Non Bootstrap
(2,0,0,4,0)	β_1	0.990(2.390)	0.974(2.036)	0.953(1.815)	0.933(1.652)	0.898(1.417)	0.931(1.519)
	β_4	0.989(2.175)	0.966(1.852)	0.927(1.709)	0.891(1.635)	0.823(1.528)	0.941(1.065)
(2,0,0,4,8)	β_1	0.989(2.479)	0.977(2.090)	0.954(1.854)	0.932(1.681)	0.891(1.440)	0.928(1.519)
	β_4	0.999(4.363)	0.989(3.467)	0.961(2.943)	0.931(2.637)	0.876(2.203)	0.943(1.807)
	β_5	0.997(5.526)	0.983(4.264)	0.966(3.577)	0.947(3.129)	0.894(2.565)	0.931(2.412)
(2,9,0,4,8)	β_1	0.992(2.628)	0.983(2.164)	0.957(1.897)	0.935(1.709)	0.898(1.453)	0.918(1.527)
	β_2	0.975(13.728)	0.976(10.271)	0.959(8.354)	0.943(7.216)	0.885(5.945)	0.941(4.401)
	β_4	0.998(6.272)	0.993(4.752)	0.976(3.968)	0.960(3.428)	0.914(2.750)	0.939(2.328)
	β_5	0.999(6.294)	0.986(4.687)	0.974(3.837)	0.955(3.329)	0.904(2.663)	0.939(2.467)
(2,9,6,4,8)	β_1	0.997(2.814)	0.985(2.260)	0.960(1.946)	0.947(1.744)	0.905(1.475)	0.922(1.516)
	β_2	0.849(17.305)	0.955(14.590)	0.966(12.200)	0.951(10.542)	0.919(8.559)	0.918(8.039)
	β_3	0.944(10.208)	0.969(7.838)	0.970(6.359)	0.958(5.478)	0.922(4.368)	0.925(3.925)
	β_4	0.968(6.498)	0.990(5.052)	0.976(4.200)	0.965(3.614)	0.930(2.888)	0.929(2.516)
	β_5	1.000(7.140)	0.988(5.209)	0.976(4.223)	0.960(3.618)	0.916(2.868)	0.928(2.643)

Table 2.11: Compare of loss functions without and with random effect for Nested-Error model, data size $n = 160$.

True β^T	Model	m=15		m=20		m=25		m=30		m=40	
		w/o	w	w/o	w	w/o	w	w/o	w	w/o	w
(2,0,0,4,0)	1,4*	0.735	0.654	0.593	0.562	0.494	0.535	0.447	0.485	0.339	0.461
	1,2,4	0.065	0.083	0.094	0.092	0.079	0.081	0.094	0.097	0.099	0.092
	1,3,4	0.058	0.079	0.081	0.106	0.108	0.106	0.094	0.112	0.130	0.119
	1,4,5	0.081	0.097	0.112	0.106	0.148	0.119	0.146	0.121	0.160	0.128
	1,2,3,4	0.029	0.052	0.063	0.076	0.076	0.083	0.094	0.088	0.097	0.094
	1,2,4,5	0.002	0.011	0.013	0.020	0.020	0.025	0.034	0.031	0.047	0.036
	1,3,4,5	0.013	0.009	0.016	0.020	0.034	0.027	0.043	0.040	0.056	0.040
	1,2,3,4,5	0.016	0.016	0.027	0.018	0.040	0.025	0.047	0.025	0.072	0.029
(2,0,0,4,8)	1,4,5*	0.820	0.781	0.724	0.719	0.642	0.654	0.583	0.637	0.489	0.596
	1,2,4,5	0.066	0.101	0.115	0.122	0.138	0.146	0.152	0.150	0.178	0.168
	1,3,4,5	0.074	0.078	0.090	0.094	0.109	0.120	0.127	0.128	0.153	0.134
	1,2,3,4,5	0.041	0.040	0.071	0.066	0.111	0.079	0.138	0.085	0.180	0.102
(2,9,0,4,8)	1,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5*	0.870	0.881	0.807	0.842	0.752	0.798	0.717	0.789	0.664	0.769
	1,3,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3,4,5	0.130	0.119	0.193	0.158	0.248	0.202	0.283	0.211	0.336	0.231
(2,9,6,4,8)	1,2,3,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,3,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3,4,5*	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.12: The coverage probabilities and lengths of confidence intervals for Nested-Error regression model with different loss functions, data size $n = 160$.

		m=15		m=20		m=25	
True β^T	Para	w/o	w	w/o	w	w/o	w
(2,0,0,4,0)	σ_v^2	0.982(1.635)	0.982(1.638)	0.951(1.406)	0.951(1.409)	0.942(1.251)	0.946(1.255)
	σ_e^2	0.996(0.775)	0.996(0.773)	0.980(0.659)	0.975(0.658)	0.957(0.597)	0.957(0.595)
	β_1	0.996(1.200)	0.996(1.200)	0.987(1.030)	0.987(1.031)	0.969(0.922)	0.971(0.922)
	β_4	0.984(1.155)	0.982(1.215)	0.957(1.038)	0.955(1.051)	0.930(0.948)	0.928(0.922)
(2,0,0,4,8)	σ_v^2	0.977(1.658)	0.977(1.659)	0.954(1.424)	0.952(1.426)	0.939(1.262)	0.941(1.264)
	σ_e^2	0.984(0.776)	0.984(0.775)	0.977(0.668)	0.977(0.667)	0.959(0.595)	0.959(0.594)
	β_1	0.993(1.207)	0.993(1.208)	0.986(1.047)	0.986(1.046)	0.974(0.932)	0.974(0.933)
	β_4	0.992(1.894)	0.993(1.903)	0.974(1.539)	0.973(1.536)	0.952(1.331)	0.951(1.321)
	β_5	0.996(2.174)	0.996(2.175)	0.990(1.797)	0.989(1.794)	0.978(1.575)	0.978(1.568)
(2,9,0,4,8)	σ_v^2	0.979(1.644)	0.979(1.646)	0.953(1.407)	0.953(1.410)	0.936(1.245)	0.936(1.247)
	σ_e^2	0.981(0.787)	0.981(0.785)	0.966(0.673)	0.966(0.672)	0.949(0.596)	0.949(0.595)
	β_1	0.994(1.216)	0.994(1.217)	0.988(1.042)	0.987(1.042)	0.970(0.933)	0.969(0.934)
	β_2	0.992(4.316)	0.990(4.286)	0.981(3.725)	0.981(3.634)	0.958(3.339)	0.958(3.231)
	β_4	0.996(2.325)	0.997(2.323)	0.986(1.891)	0.988(1.886)	0.973(1.601)	0.971(1.595)
	β_5	0.997(2.205)	0.997(2.203)	0.991(1.853)	0.991(1.845)	0.976(1.602)	0.977(1.597)
(2,9,6,4,8)	σ_v^2	0.972(1.685)	0.972(1.685)	0.940(1.438)	0.940(1.438)	0.915(1.273)	0.915(1.273)
	σ_e^2	0.978(0.800)	0.978(0.800)	0.958(0.687)	0.958(0.687)	0.946(0.604)	0.946(0.604)
	β_1	0.994(1.229)	0.994(1.229)	0.981(1.050)	0.981(1.050)	0.968(0.932)	0.968(0.932)
	β_2	0.999(7.032)	0.999(7.032)	0.996(5.862)	0.996(5.862)	0.978(5.130)	0.978(5.130)
	β_3	0.997(3.531)	0.997(3.531)	0.987(2.926)	0.987(2.926)	0.965(2.542)	0.965(2.542)
	β_4	0.997(2.470)	0.997(2.470)	0.990(1.989)	0.990(1.989)	0.988(1.705)	0.988(1.705)
	β_5	0.991(2.376)	0.991(2.376)	0.984(1.950)	0.984(1.950)	0.977(1.718)	0.977(1.718)

Continued from Table 2.12

True β^T	Para	m=30		m=40	
		w/o	w	w/o	w
(2,0,0,4,0)	σ_v^2	0.933(1.133)	0.935(1.138)	0.876(0.982)	0.883(0.986)
	σ_e^2	0.948(0.542)	0.944(0.540)	0.924(0.466)	0.924(0.465)
	β_1	0.962(0.833)	0.962(0.831)	0.924(0.726)	0.919(0.726)
	β_4	0.903(0.876)	0.892(0.845)	0.840(0.790)	0.834(0.728)
(2,0,0,4,8)	σ_v^2	0.907(1.142)	0.913(1.144)	0.857(0.982)	0.866(0.986)
	σ_e^2	0.933(0.540)	0.934(0.539)	0.891(0.464)	0.895(0.464)
	β_1	0.956(0.848)	0.956(0.847)	0.928(0.728)	0.929(0.727)
	β_4	0.936(1.178)	0.934(1.153)	0.893(1.002)	0.888(0.971)
	β_5	0.959(1.420)	0.956(1.412)	0.929(1.207)	0.922(1.194)
(2,9,0,4,8)	σ_v^2	0.913(1.133)	0.920(1.135)	0.866(0.978)	0.869(0.980)
	σ_e^2	0.927(0.539)	0.927(0.538)	0.882(0.469)	0.882(0.468)
	β_1	0.960(0.847)	0.960(0.847)	0.916(0.730)	0.916(0.730)
	β_2	0.942(3.076)	0.941(2.918)	0.898(2.693)	0.890(2.513)
	β_4	0.959(1.415)	0.958(1.405)	0.922(1.176)	0.923(1.167)
	β_5	0.961(1.450)	0.961(1.441)	0.927(1.229)	0.927(1.219)
(2,9,6,4,8)	σ_v^2	0.914(1.152)	0.914(1.152)	0.860(0.994)	0.860(0.994)
	σ_e^2	0.934(0.547)	0.934(0.547)	0.902(0.471)	0.902(0.471)
	β_1	0.947(0.852)	0.947(0.852)	0.918(0.732)	0.918(0.732)
	β_2	0.965(4.684)	0.965(4.684)	0.924(3.961)	0.924(3.961)
	β_3	0.950(2.314)	0.950(2.314)	0.926(1.947)	0.926(1.947)
	β_4	0.955(1.508)	0.955(1.508)	0.914(1.254)	0.914(1.254)
	β_5	0.947(1.536)	0.947(1.536)	0.896(1.305)	0.896(1.305)

Chapter 3

Prediction Error of Small Area Predictors Shrinking both Mean and Variances

3.1 Introduction

The goal of this chapter is to introduce a methodology which develops a dual "shrinkage" estimation for both the small area means and variances in a unified framework. In this process, the smoothed variance estimators use information from direct point estimators and their sampling variances and consequently for the smoothed small area estimators. The modeling perspective is closely related to Wang and Fuller (2003), Rivest *et al.* (2003), You and Chapman (2006) and Hwang and Zhao (2009).

The conditional mean squared error of prediction (CMSEP) is used to evaluate the prediction error, which is more akin to Booth and Hobert (1998). Booth and Hobert argued

strongly for CMSEP as opposed to unconditional mean squared error of prediction (UMSEP). Recently, this technique has again been emphasized by Lohr and Rao (2009) in the context of nonlinear mixed effect models. These authors favor CMSEP particularly for non-normal models when the posterior variance of small area parameters depends on the area specific responses. Although they were interested only in generalized linear mixed models where the posterior variance depends on area specific responses, this property of posterior variance is perhaps true for a situation with posterior non-linearity.

A brief outline of the rest of the chapter is as follows. Section 3.2 contains the proposed model and estimation method of small area parameters and the structural parameters. In Section 3.3, we provide an estimation of prediction error in terms of the conditional mean squared error of prediction. Simulation is performed in Section 3.4. And the detailed forms of some matrix calculations are given in appendix 3.5 of this chapter.

3.2 Model and Estimation Method

3.2.1 Model with Assumption

As noted by Bell (2008), the previous researchers Wang and Fuller (2003) and Rivest and Vandal (2003) essentially used the direct survey estimates to estimate the parameters related to sampling variance modeling and did not use the direct survey variance estimates though they were available. We propose a hierarchical model that uses both the direct survey estimates and sampling variance estimates to estimate all the parameters that determines the stochastic system. In this process we exploit the mean-variance joint modelling via a hierarchical model so that the final estimator is based on shrinking both mean and variance.

We like to mention that Hwang et al. (2009) considered shrinking means and variances in the context of microarray data analysis and prescribed an important solution where they plugged in a shrinkage estimator of variance into the mean estimator. Thereby the inference regarding the mean does not take into account the variance estimation completely. Furthermore, their model does not include any covariate information.

Let (X_i, S_i^2) be the pair of direct survey estimates and their sampling variances for the i^{th} area, $i = 1, \cdots, n$. Let $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ip})^T$ be the set of covariates available at the estimation stage and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ be the associated regression coefficients. We propose the following hierarchical model:

$$\left.\begin{aligned} &X_i|\theta_i, \sigma_i^2 \sim \text{Normal}(\theta_i, \sigma_i^2) \\ &\theta_i \sim \text{Normal}(\boldsymbol{Z}_i^T\boldsymbol{\beta}, \tau^2); i = 1, \cdots, n, \end{aligned}\right\} \tag{3.1}$$

$$\left.\begin{aligned} &(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1} \\ &\sigma_i^{-2} \sim \text{Gamma}\{\alpha, \gamma\}; i = 1, \cdots, n, \end{aligned}\right\} \tag{3.2}$$

where $\boldsymbol{B} = (\alpha, \gamma, \boldsymbol{\beta}^T, \tau^2)^T$, referred to as the structural parameters, are unknown and n_i represents the sample size for a simple random sample (SRS) from the i^{th} area. For a complex survey design the degrees of freedom of the chi-square distribution need to be determined carefully (e.g., Maples and Huang 2009). Note that the chi-square distribution for the sample variance is valid for only a random sample. Note that σ_i^2 are the sampling variances of the X_i's and are usually estimated by S_i^2's. You and Chapman (2006) did not consider the second level of sampling variance modelling. Thus their model can be treated as the Bayesian version

of the models considered in Rivest and Vandal (2003) and Wang and Fuller (2003). The second level of (3.2) might be further extended as Gamma$\{\gamma, \exp(\boldsymbol{Z}_i^T\boldsymbol{\beta}_2)/\gamma\}$ to accommodate covaraite information in the variance modeling. The inference can be made from the conditional distribution of θ_i (the parameter of interest) given the data $(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, \cdots, n$. However, this does not have a closed form expression. We adopted rejection sampling to overcome the situation.

3.2.2 Estimation of the Small Area Parameters

If the parameters $\boldsymbol{B}$ are known, the joint distribution of $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$ is

$$
\begin{aligned}
\pi(X_i, S_i^2, \theta_i, \sigma_i^2|\boldsymbol{B}) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}\} \frac{1}{\Gamma\{(n_i-1)/2\}2^{(n_i-1)/2}} \\
&\quad \left[(n_i-1)\frac{S_i^2}{\sigma_i^2}\right]^{(n_i-1)/2-1} \exp\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\} \left(\frac{n_i-1}{\sigma_i^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \\
&\quad \exp\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\} \frac{1}{\Gamma(\alpha)\gamma^\alpha} \left(\frac{1}{\sigma_i^2}\right)^{\alpha+1} \exp\{-\frac{1}{\gamma\sigma_i^2}\} \\
&\approx \exp\left\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2} - \frac{(n_i-1)S_i^2}{2\sigma_i^2} - \frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2} - \frac{1}{\gamma\sigma_i^2}\right\} \\
&\quad \times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+\alpha+1} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(\alpha)\gamma^\alpha} \\
&= \exp\left[-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2} - \left\{\frac{(X_i-\theta_i)^2}{2} + \frac{(n_i-1)S_i^2}{2} + \frac{1}{\gamma}\right\}\frac{1}{\sigma_i^2}\right] \\
&\quad \times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+\alpha+1} \frac{1}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha}.
\end{aligned} \tag{3.3}
$$

Therefore the conditional distribution of σ_i^2 and θ_i given the data (X_i, S_i^2), $i = 1, \cdots, n$ and $\boldsymbol{B}$ are

$$\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B}) \propto \frac{\exp\left[-\frac{(X_i - Z_i^T\beta)^2}{2(\sigma_i^2+\tau^2)} - \left\{\frac{(n_i-1)}{2}S_i^2 + \frac{1}{\gamma}\right\}\frac{1}{\sigma_i^2}\right]}{(\sigma_i^2)^{(n_i-1)/2+\alpha+1}(\sigma_i^2+\tau^2)^{1/2}}, \tag{3.4}$$

$$\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) \propto \exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\}\psi_i^{-(n_i/2+\alpha)}, \tag{3.5}$$

where

$$\psi_i \equiv \left\{0.5(X_i - \theta_i)^2 + 0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\}. \tag{3.6}$$

Note that the above conditionals are obtained by integrating out θ_i and σ_i^2 respectively from the joint distribution (3.3).

From now on we will borrow the notations from Booth and Hobert (1998) for determining all related stochastic distributions. In this context, a meaningful point estimator for θ_i is its conditional mean,

$$\theta_i(\boldsymbol{B}; X_i, S_i^2) = E_{\boldsymbol{B}}(\theta_i|X_i, S_i^2), \tag{3.7}$$

where $E_{\boldsymbol{B}}$ represents the expection with respect to the conditional distribution of θ_i with known $\boldsymbol{B}$. A sensible prediction error can be measured by the posterior variance

$$\nu_i(\boldsymbol{B}; X_i, S_i^2) = var_{\boldsymbol{B}}(\theta_i|X_i, S_i^2). \tag{3.8}$$

Neither (3.7) nor (3.8) has any closed form expression. Therefore, we have numerically computed using the following approach.

1. Generate R random numbers from $\pi(\theta_i|X_i, S_i^2, \boldsymbol{B})$ with rejection sampling method:

   1a. Generate a candidate random sample $\theta_i^{(c)}$ from a Normal$(Z_i^T\beta, \tau^2)$;

   1b. Generate a uniform random number U;

   1c. If
   $$U < \left\{ 1 + \frac{(\frac{\theta_i - X_i}{Gx})^2}{n_i + 2\alpha - 1} \right\}^{-(n_i/2+\alpha)}$$
   where $G_x^2 = \{(n_i - 1)S_i^2 + 2/\gamma\}/(n_i + 2\alpha + 1)$, then accept the candiate random sample as $\theta_i^{(r)}$; otherwise go back to 1a;

   1d. Repeat 1a to 1c R times.

2. Approximate θ_i and ν_i by
   $$\tilde{\theta}_i = \frac{1}{R}\sum_{r=1}^{R} \theta_i^{(r)}$$
   and
   $$\tilde{\nu}_i = \frac{1}{R}\sum_{r=1}^{R} (\theta_i^{(r)} - \tilde{\theta}_i)^2$$

In pratice $\boldsymbol{B}$ is unknown. We estimate them by maximizing the marginal likelihood and the details are given in the next section. Let $\hat{\boldsymbol{B}}$ denote the corresponding estimator. Substituting $\boldsymbol{B}$ by $\hat{\boldsymbol{B}}$ in formulas (3.7) and (3.8) will produce the estimates of θ_i and ν_i:

$$\hat{\theta}_i = \theta_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = E_{\hat{\boldsymbol{B}}}(\theta_i|X_i, S_i^2) \tag{3.9}$$

$$\hat{\nu}_i = \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = var_{\hat{\boldsymbol{B}}}(\theta_i|X_i, S_i^2). \tag{3.10}$$

The estimator (3.9) is popularly known as the empirical Bayes estimator and (3.10) is the estimated Bayes risks. It is well known that the ν_i only consider the variability in the prediction procedure, but not the variability due to the parameter estimation ($\hat{\boldsymbol{B}}$). We accounted this additional variability by adapting the technique of Booth and Hobert (1998) in this set up. The details are discussed in Section 3.3.

3.2.3 Estimation of the Stuctural Parameters

In pracice, $\boldsymbol{B}$ is unknown. We obtain the maximum likelihood estimate of $\boldsymbol{B}$ by maximizing the marginal likelihood $L_M = \prod_{i=1}^{n} L_i^M$ of $\{(X_i, S_i^2, \boldsymbol{Z}_i)_{i=1}^n; \boldsymbol{B}\}$, where

$$L_i^M \propto \text{Constant}_i \cdot \frac{\Gamma(\frac{n_i}{2}+\alpha)}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha} \int \exp\left\{-\frac{(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i. \quad (3.11)$$

Therefore, the log-likelihood is

$$\log(L_M) = \sum_{i=1}^{n} \Bigg[\log Constant_i + \log\{\Gamma(\frac{n_i}{2}+\alpha)\} - \log\{\Gamma(\alpha)\} - \frac{1}{2}\log\tau^2 - \alpha\log(\gamma) + \log\left[\int exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i\right]\Bigg]$$

The maximizing marginal likelihood equation is

$$\frac{\partial \log(L_M)}{\partial \boldsymbol{B}} = 0$$

where L_M is the complete data likelihood and $L_M = \prod_{i=1}^{n} L_i^M$.

The detailed expression of the partial derivative corresponding to $\boldsymbol{\beta}$ is:

$$\begin{aligned}\frac{\partial \log(L_M)}{\partial \beta} &= \sum_{i=1}^{n} \frac{\int Z \frac{\theta_i - Z_i^T\beta}{\tau^2} exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i}{\int exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i} \\ &= \sum_{i=1}^{n} E\left(Z\frac{\theta_i - Z_i^T\beta}{\tau^2}\right) \end{aligned} \tag{3.12}$$

where the expectation corresponds to conditional distribution of θ_i, $\pi(\theta_i|X_i, S_i^2, \boldsymbol{B})$. The estimate of $\boldsymbol{\beta}$ is obtained by solving $\partial \log(L_M)/\partial\beta = 0$, and we obtain

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T)^{-1} \{\sum_{i=1}^{n} \boldsymbol{Z}_i E(\theta_i)\}. \tag{3.13}$$

The expression of the partial derivative corresponding to τ^2 is:

$$\begin{aligned}\frac{\partial \log(L_M)}{\partial \tau^2} &= -\frac{n}{2\tau^2} + \sum_{i=1}^{n} \frac{\int \frac{(\theta_i - Z_i^T\beta)^2}{2(\tau^2)^2} exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i}{\int exp\left\{-\frac{(\theta_i - Z_i^T\beta)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i} \\ &= -\frac{n}{2\tau^2} + \sum_{i=1}^{n} E\left(\frac{(\theta_i - Z_i^T\beta)^2}{2(\tau^2)^2}\right) \end{aligned} \tag{3.14}$$

Then the estimate of τ^2 is obtained by solving $\partial \log(L_M)/\partial\tau^2 = 0$, and we obtain

$$\hat{\tau}^2 = \sum_{i=1}^{n} E(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/n. \tag{3.15}$$

Similarly, we estimate α and γ by solving $S_\alpha = 0$ and $S_\gamma = 0$ where

$$\begin{aligned} S_\alpha &= \frac{\partial E log(L_M)}{\partial \alpha} \\ &= \sum_{i=1}^{n} log'\{\Gamma(\frac{n_i}{2}+\alpha)\} - n log'\{\Gamma(\alpha)\} - n log(\gamma) - \sum_{i=1}^{n} E\{log(\psi_i)\} \qquad (3.16) \\ S_\gamma &= \frac{\partial E log(L_M)}{\partial \gamma} \\ &= -\frac{n\alpha}{\gamma} + \frac{1}{\gamma^2}\sum_{i=1}^{n}(\frac{n_i}{2}+\alpha)E\left(\frac{1}{\psi_i}\right) \qquad (3.17) \end{aligned}$$

To solve equations (3.13), (3.15), (3.16) and (3.17), we use the EM algorithm. At the E-step of the t^{th} iteration, we calculate the expectation with respect to the conditional density of θ_i, $\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}^{(t-1)})$, $i = 1, \ldots, n$, with $\boldsymbol{B}^{(t-1)}$ denoting the estimate of $\boldsymbol{B}$ at the $(t-1)^{th}$ iteration. The rejection sampling method is used to generate samples from the conditional distribution of θ_i's. At the M-step of the t^{th} iteration, we maximize L_M by solving equations (3.13), (3.15), (3.16) and (3.17) conditional on the values of expectations obtained in the E-step. To solve these equations, we also need

$$\begin{aligned} S_{\alpha\alpha} &= \textstyle\sum_{i=1}^{n}\left[log''\{\Gamma(\frac{n_i}{2}+\alpha)\} - log''\{\Gamma(\alpha)\} + Var\{\log(\psi_i)\}\right] \\ S_{\alpha\gamma} &= \textstyle\sum_{i=1}^{n}\left[-\frac{1}{\gamma} + \frac{1}{\gamma^2}E\left(\frac{1}{\psi_i}\right) - (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}Cov\left\{\frac{1}{\psi_i}, \log(\psi_i)\right\}\right] \\ S_{\gamma\alpha} &= S_{\alpha\gamma} \qquad (3.18) \\ S_{\gamma\gamma} &= \textstyle\sum_{i=1}^{n}\left\{\frac{\alpha}{\gamma^2} - (n_i+2\alpha)\frac{1}{\gamma^3}E\left(\frac{1}{\psi_i}\right) + (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^4}E\left(\frac{1}{\psi_i^2}\right)\right. \\ &\quad \left. +(\frac{n_i}{2}+\alpha)^2\frac{1}{\gamma^4}Var\left(\frac{1}{\psi_i}\right)\right\} \end{aligned}$$

Then α and γ can be estimated by iterative Newton-Raphson method:

$$\begin{bmatrix} \alpha \\ \gamma \end{bmatrix}^{l} = \begin{bmatrix} \alpha \\ \gamma \end{bmatrix}^{(l-1)} - \begin{bmatrix} S_{\alpha\alpha} & S_{\alpha\gamma} \\ S_{\gamma\alpha} & S_{\gamma\gamma} \end{bmatrix}^{-1} \begin{bmatrix} S_{\alpha} \\ S_{\gamma} \end{bmatrix}$$

3.3 Prediction Error Calculation

3.3.1 Mean Squared Error of Prediction

Following the definition of conditional mean squared prediction error (CMSEP) of Booth and Hobert (1998), the prediction error variance is,

$$CMSEP(\boldsymbol{B}; X_i, S_i^2) = E_{\boldsymbol{B}}[\{\hat{\theta}_i - \theta_i(\boldsymbol{B}; X_i, S_i^2)\}^2 | X_i, S_i^2]$$

where $\hat{\theta}_i$ and $\theta_i(\boldsymbol{B}; X_i, S_i^2)$ are as defined in (3.9) and (3.7). Since $\theta_i(\boldsymbol{B}; X_i, S_i^2) - \theta_i$ and $\hat{\theta}_i - \theta_i(\boldsymbol{B}; X_i, S_i^2)$ are conditionally independent given X_i and S_i^2,

$$\begin{aligned} CMSEP(\boldsymbol{B}; X_i, S_i^2) &= var_{\boldsymbol{B}}(\theta_i | X_i, S_i^2) + E_{\boldsymbol{B}}[\{\hat{\theta}_i - \theta_i(\boldsymbol{B}; X_i, S_i^2)\}^2 | X_i, S_i^2] \\ &= \nu_i(\boldsymbol{B}; X_i, S_i^2) + c_i(\boldsymbol{B}; X_i, S_i^2) \end{aligned} \tag{3.19}$$

where $c_i(\boldsymbol{B}; X_i, S_i^2)$ is the correction term due to the estimation of unknown parameters $\boldsymbol{B}$. The correction contribution is of order $O_p(n^{-1})$. Note that the above measure is still not usable because it involved the unknown structural parameters $\boldsymbol{B}$. It's natural to plug-in the

estimate $\hat{\boldsymbol{B}}$ of $\boldsymbol{B}$ and get a usable measure of mean squared prediction error

$$\widehat{CMSEP}_i = MSE(\hat{\boldsymbol{B}}; X_i, S_i^2) = \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) + c_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = \hat{\nu}_i + \hat{c}_i$$

As it will be clear in the next section (as well as from small area estimation literature) the estimator (3.19) has considerable bias. Typically the order of bias is $O_p(n^{-1})$ due to estimation of ν_i by $\hat{\nu}_i$.

3.3.2 Bias Correction for $\nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2)$

For the bias correction of the estimated conditional variance $\nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2)$ we expand this about $\boldsymbol{B}$:

$$\begin{aligned}\hat{\nu}_i = \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = \nu_i(\boldsymbol{B}; X_i, S_i^2) + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \frac{\partial \nu_i(B; X_i, S_i^2)}{\partial \boldsymbol{B}} \\ + \frac{1}{2}(\hat{\boldsymbol{B}} - \boldsymbol{B})^T \frac{\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B} \partial \boldsymbol{B}^T}(\hat{\boldsymbol{B}} - \boldsymbol{B}) + O_p(n^{-2}). \qquad (3.20)\end{aligned}$$

Then the approximated bias involved in $\nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2)$ is

$$E(\hat{\boldsymbol{B}} - \boldsymbol{B}) \frac{\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B}} + \frac{1}{2}\mathrm{tr}\{\frac{\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B} \partial \boldsymbol{B}^T} \boldsymbol{I}^{-1}(\boldsymbol{B})\},$$

where $\boldsymbol{I}(\boldsymbol{B})$ is the Fisher's information matrix obtained from the marginal likelihood L_M,

$$I_{sr} = -E\left[\frac{\partial^2}{\partial \boldsymbol{B}_s \partial \boldsymbol{B}_r} log - likelihood(X_i, S_i^2)\right]$$

The second order derivative of the log-likelihood is given in section 3.2.3 (3.18) and appendix 3.5.1 and can be used to compute the information matrix.

Handling the bias analytically is difficult. Booth and Hobert (1998) adopted bootstrap bias correction instead. This is due to the fact that there is no closed form expressions available for this bias terms. The bootstrap method requires repeated estimation of model parameters based on re-sampled data. This often pauses practical and computational difficulties in this hierarchical model. As we will see in the next subsection handling the second term in (3.19) is also difficult due to the same difficulty of not having any closed form expression. Thus, if we have to do the bootstrap for the bias correction of ν_i, the estimation of c_i can also be done in the same run. It is not necessary to obtain any analytical approximations. The resampling techniques has been used in Jiang *et al.* (2002), Hall and Maiti (2006). We will be applying the rejection sampling method instead where there is no need of estimating model parameters repeatedly.

Let $\hat{\boldsymbol{B}}$ be the maximum likelihood estimator of $\boldsymbol{B}$ as proposed in the previous section. Following Cox & Snell (1968, Equation 20), we approximate $E\left(\hat{\boldsymbol{B}}-\boldsymbol{B}\right)$ up to $O(n^{-1})$. Define $I^{-1}=(I^{rs})$ with as the inverse of $I=(I_{rs})$, where $I_{rs}=E(-V_{rs}^{(\cdot)})$ and $V_{rs}^{(i)}=\partial^2 \log L_i^M/\partial \boldsymbol{B}_r \partial \boldsymbol{B}_s$. The bias in the s^{th} element of $\hat{\boldsymbol{B}}$ is

$$E(\hat{\boldsymbol{B}}_s-\boldsymbol{B}_s)\approx\frac{1}{2}\sum_r\sum_t\sum_u I^{rs}I^{tu}(K_{rtu}+2J_{t,ru}) \tag{3.21}$$

$$K_{rst}=E(W_{rst}^{(\cdot)}),\qquad W_{rst}^{(i)}=\frac{\partial^3 \log L_i^M}{\partial \boldsymbol{B}_r \partial \boldsymbol{B}_s \partial \boldsymbol{B}_t},$$

$$J_{r,st}=E\{\sum U_r^{(i)}V_{st}^{(i)}\},\qquad U_r^{(i)}=\frac{\partial \log L_i^M}{\partial \boldsymbol{B}_r}$$

where the L_i^M is the marginal likelihood of (X_i, S_i^2) defined in the previous section. The

detailed formulas are given in appendix.

With

$$\left(\frac{\partial \nu_i}{\partial \boldsymbol{B}}\right)^T = \left\{\frac{\partial \nu_i}{\partial \alpha}, \frac{\partial \nu_i}{\partial \gamma}, \frac{\partial \nu_i}{\partial \beta}, \frac{\partial \nu_i}{\partial \tau^2}\right\} \tag{3.22}$$

and $\nu_i(\boldsymbol{B}; X_i, S_i^2) = var_{\boldsymbol{B}}(\theta_i | X_i, S_i^2)$, we have

$$\begin{aligned}
\frac{\partial \nu_i}{\partial \alpha} &= -Cov^*\{\theta_i^2, \log(\psi_i)\} + 2E^*\theta_i Cov^*\{\theta_i, \log(\psi_i)\} \\
\frac{\partial \nu_i}{\partial \gamma} &= (\frac{n_i}{2} + \alpha)\frac{1}{\gamma^2}\left\{Cov^*(\theta_i^2, \frac{1}{\psi_i}) - 2E^*\theta_i Cov^*(\theta_i, \frac{1}{\psi_i})\right\} \\
\frac{\partial \nu_i}{\partial \beta} &= \frac{1}{\tau^2}\left\{Cov^*(\theta_i^2, \theta_i) - 2E^*\theta_i Cov^*(\theta_i, \theta_i)\right\} \\
\frac{\partial \nu_i}{\partial \tau^2} &= \frac{1}{2(\tau^2)^2}\left\{Cov^*\{\theta_i^2, (\theta_i - \beta)^2\} - 2E^*\theta_i Cov^*\{\theta_i, (\theta_i - \beta)^2\}\right\}
\end{aligned}$$

where the * means that the expection, varirance and covariance that are calculated with respect to the conditional distribution of θ_i at the estimated parameters' value. The approximated expression of $\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)/\partial \boldsymbol{B} \partial \boldsymbol{B}^T$ is given in the appendix. The expection, variance and covariance are computed based on the Monte Carlo approximation. For examples, $\hat{E^*}(\theta_i) = \sum_{r=1}^{R} \theta_i^{(r)}/R$, where $(\theta_i^{(1)}, \cdots, \theta_i^{(R)})$ are random numbers generated from the conditional distribution of θ_i. Therefore,

$$\hat{\nu_i} \approx \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) - (\hat{\boldsymbol{B}} - \boldsymbol{B})\frac{\partial \nu_i}{\partial \boldsymbol{B}} - \frac{1}{2} tr\{\frac{\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B} \partial \boldsymbol{B}^T} \boldsymbol{I}^{-1}(\boldsymbol{B})\}.$$

This expression is second order correct meaning the bias is of order $o_p(n^{-1})$.

3.3.3 Approximation of $c_i(\boldsymbol{B}; X_i, S_i^2)$

The definition of $c_i(\boldsymbol{B}; X_i, S_i^2)$ is given in the previous section by $c_i(\boldsymbol{B}; X_i, S_i^2) = E_{\boldsymbol{B}}[\{\hat{\theta}_i - \theta_i(\boldsymbol{B}; X_i, S_i^2)\}^2 | X_i, S_i^2]$ where $\hat{\theta}_i - \theta_i(\boldsymbol{B}; X_i, S_i^2) = \theta_i(\hat{\boldsymbol{B}}; X_i, S_i^2) - \theta_i(\boldsymbol{B}; X_i, S_i^2)$. Using the Taylor series expansion and ignoring the term $O_p(|\hat{\boldsymbol{B}} - \boldsymbol{B}|^2)$ we write

$$\theta_i(\hat{\boldsymbol{B}}; X_i, S_i^2) - \theta_i(\boldsymbol{B}; X_i, S_i^2) = A_i^T(\boldsymbol{B}; X_i, S_i^2)(\hat{\boldsymbol{B}} - \boldsymbol{B}) \tag{3.23}$$

where

$$\begin{aligned} A_i^T(\boldsymbol{B}; X_i, S_i^2) &= \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B}} \\ &= \left(\frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \alpha}, \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \gamma}, \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \beta}, \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \tau^2} \right) \end{aligned}$$

Since $\theta_i(\boldsymbol{B}; X_i, S_i^2) = E_{\boldsymbol{B}}(\theta_i | X_i, S_i^2)$, the components of A_i are

$$\begin{aligned} \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \alpha} &= E^*(\theta_i)E^*\{\log(\psi_i)\} - E^*\{\theta_i \log(\psi_i)\} = -Cov^*\{\theta_i, \log(\psi_i)\} \\ \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \gamma} &= (\frac{n_i}{2} + \alpha)\frac{1}{\gamma^2}\left\{ E^*\left(\frac{\theta_i}{\psi_i}\right) - E^*(\theta_i)E^*\left(\frac{1}{\psi_i}\right)\right\} = (\frac{n_i}{2} + \alpha)\frac{1}{\gamma^2} Cov^*(\theta_i, \frac{1}{\psi_i}) \\ \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \beta} &= \frac{1}{\tau^2}\left[E^*(\theta_i^2) - \{E^*(\theta_i)\}^2\right] = \frac{1}{\tau^2}\text{var}^*(\theta_i) \\ \frac{\partial \theta_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \tau^2} &= \frac{1}{2(\tau^2)^2}\left[E^*\{\theta_i(\theta_i - \beta)^2\} - E^*(\theta_i)E^*(\theta_i - \beta)^2\right] = \frac{1}{2\tau^2} Cov^*\{\theta_i, (\theta_i - \beta)^2\} \end{aligned}$$

As a consequence of (3.23)

$$c_i(\boldsymbol{B}; X_i, S_i^2) \approx A_i^T(\boldsymbol{B}; X_i, S_i^2)\boldsymbol{I}^{-1}(\boldsymbol{B})A_i(\boldsymbol{B}; X_i, S_i^2) \tag{3.24}$$

and the approximation is correct up to $O_p(n^{-1})$. Substituting $\hat{\boldsymbol{B}}$ into formula (3.24) will yield an estimate of the correction term,

$$c_i(\hat{\boldsymbol{B}}; X_i, S_i^2) \approx A_i^T(\hat{\boldsymbol{B}}; X_i, S_i^2)\boldsymbol{I}^{-1}(\hat{\boldsymbol{B}})A_i(\hat{\boldsymbol{B}}; X_i, S_i^2).$$

As the estimated information matrix is $\{\boldsymbol{I}(\hat{\boldsymbol{B}})\}^{-1} = O_p(n^{-1})$, the error in the approximation is $o_p(n^{-1})$. Similar to the case of ν_i, the items in $A_i(\hat{\boldsymbol{B}}; X_i, S_i^2)$ do not have closed forms. They can be approximated by the samples generated in the estimation procedure. Summing up all the derivations and approximations we obtain the following result.

Theorem. *The estimated conditional mean squared of prediction error for $\hat{\theta}_i$ is*

$$\begin{aligned}\widehat{CMSPE} = \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) - (\hat{\boldsymbol{B}} - \boldsymbol{B})\frac{\partial \nu_i}{\partial \boldsymbol{B}} - \frac{1}{2}\mathrm{tr}\{\frac{\partial \nu_i(\boldsymbol{B}; X_i, S_i^2)}{\partial \boldsymbol{B}\partial \boldsymbol{B}^T}\boldsymbol{I}^{-1}(\boldsymbol{B})\} \\ + A_i^T(\hat{\boldsymbol{B}}; X_i, S_i^2)I(\hat{\boldsymbol{B}})^{-1}A_i(\hat{\boldsymbol{B}}; X_i, S_i^2). \qquad (3.25)\end{aligned}$$

The formula is second order correct in the sense that it has a bias of $O_p(n^{-2})$.

3.4 Simulation Study

Simulation design To check the finite sample performance of the proposed estimators, a simulation set up closely related to Wang and Fuller (2003) was considered. To simplify the simulation, we do not choose any covariate Z, only (X_i, S_i^2) are generated. First, generate observations for each small area using the model

$$X_{ij} = \beta + u_i + e_{ij}, \quad j = 1, \ldots, n_i,\ i = 1, \ldots, n,$$

where $u_i \sim \text{Normal}(0, \tau^2)$ and $e_{ij} \sim \text{Normal}(0, \text{n}_\text{i}\sigma_\text{i}^2)$. Then the random effects model for the small area mean is

$$X_i = \beta + u_i + e_i, \quad i = 1, \ldots, n, \tag{3.26}$$

where $X_i = \bar{X_i}. = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$, $e_i = \bar{e_i}. = n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$. Therefore, $X_i \sim \text{Normal}(\theta_i, \sigma_i^2)$, where $\theta_i = \beta + u_i$ and $\theta_i \sim \text{Normal}(\beta, \tau^2)$, and $e_i \sim \text{Normal}(0, \sigma_i^2)$. We estimated the mean for each small area, θ_i, $i = 1, \ldots, n$. We estimated σ_i^2 with the unbiased estimator

$$S_i^2 = \frac{1}{n_i - 1} \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i.)^2 \tag{3.27}$$

It is to be noted that $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{(n_i - 1)}$. Like Wang and Fuller (2003), we set all n_i equal to m that eased our programming efforts. However, the sampling variances were still unequal by choosing one-third of the σ_i^2 equal to 1, one-third equal to 4, and one-third equal to 16. In the simulation, we set $\beta = 10$ and took three different values of τ^2, 0.5, 1, 4. For each of τ^2, we generated 1,000 samples for each of the combinations $(m, n) = (9, 36)$, $(18, 180)$. In table 3.1 we present the mean, empirical standard deviation and the estimated standard deviation of estimates of β and τ^2. The results are consistent with the large sample theory of maximum likelihood estimation.

Method of analysis We also compute the estimates based on the Wang and Fuller approach. Table 3.2 and Table 3.3 provide the numerical results for comparing two methods: (I) the proposed method and (II) Wang and Fuller (2003). All the results are averaged over areas within the group having the same sampling variances. We calculated the bias and prediction mean squared error of $\hat{\theta}_i$ based on 1,000 replications. We used empirical measures

Table 3.1: Results of the simulation study, and here we present estimate (Est.), empirical standard deviation (SD) for β and τ^2. We set $\beta = 10$.

		For β			
		Proposed Method		Wang and Fuller (2003)	
	τ^2	Est.	SD	Est.	SD
$n = 36$, $m = 9$	0.5	9.9989	0.3715	10.0031	0.3355
	1	10.0053	0.3772	10.0079	0.3684
	4	9.9978	0.4950	9.9992	0.4967
$n = 180$, $m = 18$	0.5	9.9978	0.1676	10.0015	0.1387
	1	9.9987	0.1660	10.0013	0.1532
	4	9.9993	0.2134	9.9996	0.2127
		For τ^2			
		Proposed Method		Wang and Fuller (2003)	
	τ^2	Est.	SD	Est.	SD
$n = 36$, $m = 9$	0.5	0.5978	0.3156	1.2227	0.9973
	1	1.1190	0.4983	1.6468	1.1591
	4	4.0449	1.5113	4.5154	2.0121
$n = 180$, $m = 18$	0.5	0.5211	0.0381	0.7767	0.3783
	1	1.0451	0.1190	1.2757	0.4618
	4	4.1092	0.5812	4.2707	0.8718

of relative bias and coefficient of variation to quantify the performaces of different MSE estimators. Relative bias of the MSE estimator was defined by

$$RB_i = \left| \frac{E\{\widehat{MSE}_i\} - MSE_i}{MSE_i} \right| \tag{3.28}$$

for $i = 1, ..., 3$, where $E\{\widehat{MSE}_i\}$ was estimated empirically as the average of values of $\widehat{MSE}_i$ over replications. MSE_i was defined as the average value of $\left(\hat{\theta}_i - \theta_i\right)^2$. The coefficient of variation of the MSE estimator was taken to be

$$CV_i = \frac{[E\{\widehat{MSE}_i - MSE_i\}^2]^{\frac{1}{2}}}{MSE_i} \tag{3.29}$$

for $i = 1, ..., 3$.

Table 3.2: Results of MSE estimator, $n = 36$, $m = 9$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
Bias	1	0.0018	-0.0024	0.0098	0.0157	0.0032	0.0070
	4	-0.0061	-0.0001	0.0001	-0.0011	0.0066	0.0056
	16	-0.0208	-0.0235	0.0173	0.0198	0.0365	0.0361
MSE	1	0.4281	0.6532	0.5588	0.6352	0.8786	0.8583
	4	0.6658	1.0389	0.9718	1.1982	2.2183	2.3032
	16	0.6785	1.0215	1.1576	1.4069	3.6549	3.8209
RB	1	0.2561	0.6044	0.2364	0.2885	0.2922	0.0009
	4	0.0016	1.1227	-0.0544	0.4850	-0.0557	0.0363
	16	0.0508	1.7802	-0.1246	0.6538	-0.1120	0.0450
CV	1	0.6327	0.9124	0.5246	0.5963	0.4925	0.4293
	4	0.6908	1.4339	0.4382	0.7895	0.3309	0.3697
	16	0.9411	2.2498	0.5212	1.0594	0.3272	0.3721

Table 3.3: Results of MSE estimator, $n = 180$, $m = 18$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
Bias	1	0.0009	0.0001	-0.0017	-0.0000	-0.0007	-0.0005
	4	-0.0034	-0.0053	0.0016	0.0040	0.0049	0.0053
	16	0.0004	-0.0014	0.0034	0.0053	0.0048	0.0045
MSE	1	0.3524	0.3931	0.5146	0.5383	0.8121	0.8119
	4	0.4742	0.5226	0.8434	0.8897	2.0940	2.1194
	16	0.5110	0.5292	0.9612	0.9866	3.2698	3.3052
RB	1	0.0696	0.3926	0.1237	0.1529	0.1847	0.0245
	4	-0.0216	0.5509	0.0044	0.2046	-0.0190	0.0231
	16	-0.0324	0.6202	0.0295	0.2717	-0.0043	0.0639
CV	1	0.1791	0.5482	0.2228	0.3309	0.3128	0.2939
	4	0.1142	0.7653	0.1323	0.3987	0.1833	0.2150
	16	0.1045	0.8875	0.1263	0.5028	0.1389	0.2006

Simulation results The proposed method has considerable lower prediction mean squared errors. The gain is a maximum when the ratio of sampling variance to model variance is the largest. The reduction is almost half when the number of small areas is 36. The relative bias is always less than Wang and Fuller (2003) except the case $\tau^2 = 4, \sigma_i = 1$ (high model variance, low sampling variance). The proposed method might have low underestimation (about 5%) in case of $n = 36$. On the other hand, the Wang Fuller method can have very large overestimation. In terms of coefficient of variation the proposed method outperforms Wang and Fuller (2003).

3.5 Appendix: Matrix Calculation Results

3.5.1 Computation of $\hat{B} - B$

In this section of the appendix, the detailed expression of equations (3.21) are given. The log-likelihood is already given in the section 3.2.3. The first order derivative terms, $U_r^{(i)}$, are given in the formulas (3.16), (3.17), (3.12) and (3.14) in section 3.2.3.

For the second order derivative, $V_{\alpha\alpha}^{(i)}$, $V_{\gamma\gamma}^{(i)}$, and $V_{\alpha\gamma}^{(i)}$ are given in equation (3.18) as $S_{\alpha\alpha}$, $S_{\gamma\gamma}$, and $S_{\alpha\gamma}$. The other terms are

$$
\begin{aligned}
V_{\beta\beta}^{(i)} &= -\frac{1}{\tau^2} + \frac{1}{(\tau^2)^2} Var(\theta_i - \beta) \\
V_{\tau^2\tau^2}^{(i)} &= \frac{1}{2(\tau^2)^2} - \frac{1}{(\tau^2)^3} E(\theta_i - \beta)^2 + \frac{1}{4(\tau^2)^4} Var((\theta_i - \beta)^2)
\end{aligned}
$$

The other cross terms of the second order derivative are

$$
\begin{aligned}
V^{(i)}_{\alpha\beta} &= -\frac{1}{\tau^2} Cov(\log(\psi_i), \theta_i - \beta) \\
V^{(i)}_{\alpha\tau^2} &= -\frac{1}{2(\tau^2)^2} Cov(\log(\psi_i), (\theta_i - \beta)^2) \\
V^{(i)}_{\gamma\beta} &= (\frac{n_i}{2} + \alpha) \frac{1}{\gamma^2 \tau^2} Cov(\frac{1}{\psi_i}, \theta_i - \beta) \\
V^{(i)}_{\gamma\tau^2} &= (\frac{n_i}{2} + \alpha) \frac{1}{2\gamma^2 (\tau^2)^2} Cov(\frac{1}{\psi_i}, (\theta_i - \beta)^2) \\
V^{(i)}_{\beta\tau^2} &= -\frac{1}{(\tau^2)^2} E(\theta_i - \beta) + \frac{1}{2(\tau^2)^3} Cov(\theta_i - \beta, (\theta_i - \beta)^2)
\end{aligned}
$$

The third derivative terms are:

$$
\begin{aligned}
W^{(i)}_{\alpha\alpha\alpha} &= \log\Gamma'''(\frac{n_i}{2} + \alpha) - \log\Gamma'''(\alpha) - Cov\{\log(\psi_i), \log^2(\psi_i)\} \\
&\quad -2E\log(\psi_i) Var\{\log(\psi_i)\} \\
W^{(i)}_{\gamma\gamma\gamma} &= -\frac{2\alpha}{\gamma^3} + (\frac{n_i}{2} + \alpha) \left(\frac{6}{\gamma^4} E\frac{1}{\psi_i} - \frac{6}{\gamma^5} E\frac{1}{\psi_i^2} + \frac{2}{\gamma^6} E\frac{1}{\psi_i^3} \right) \\
&\quad -(\frac{n_i}{2} + \alpha)^2 \left\{ \frac{6}{\gamma^5} Var(\frac{1}{\psi_i}) + \frac{1}{\gamma^6} Cov(\frac{1}{\psi_i}, \frac{1}{\psi_i^2}) \right\} \\
&\quad -(\frac{n_i}{2} + \alpha)^3 \frac{1}{\gamma^6} \left\{ Cov(\frac{1}{\psi_i}, \frac{1}{\psi_i^2}) - 2E\frac{1}{\psi_i} Var(\frac{1}{\psi_i}) \right\} \\
W^{(i)}_{\beta\beta\beta} &= \frac{1}{(\tau^2)^3} \left[Cov\{\theta_i - \beta, (\theta_i - \beta)^2\} - 2E(\theta_i - \beta) Var(\theta_i - \beta) \right] \\
W^{(i)}_{\tau^2\tau^2\tau^2} &= -\frac{1}{(\tau^2)^3} + \frac{3}{(\tau^2)^4} E(\theta_i - \beta)^2 - \frac{3}{2(\tau^2)^5} Var\{(\theta_i - \beta)^2\} + \frac{1}{8(\tau^2)^6} \cdot \\
&\quad \left[Cov\{(\theta_i - \beta)^4, (\theta_i - \beta)^2\} - 2E(\theta_i - \beta)^2 Var\{(\theta_i - \beta)^2\} \right]
\end{aligned}
$$

The other cross terms of the third order derivative are

$$
\begin{aligned}
W^{(i)}_{\alpha\alpha\gamma} &= (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\left[Cov\{\log^2(\psi_i),\frac{1}{\psi_i}\}-2E\log(\psi_i)Cov\{\log(\psi_i),\frac{1}{\psi_i}\}\right]\\
&\quad -\frac{2}{\gamma^2}Cov\{\log(\psi_i),\frac{1}{\psi_i}\}\\
W^{(i)}_{\alpha\alpha\beta} &= \frac{1}{\tau^2}\left[Cov\{\log^2(\psi_i),\theta_i-\beta\}-2E\log(\psi_i)Cov\{\log(\psi_i),\theta_i-\beta\}\right]\\
W^{(i)}_{\alpha\alpha\tau^2} &= \frac{1}{2(\tau^2)^2}\left[Cov\{\log^2(\psi_i),(\theta_i-\beta)^2\}-2E\log(\psi_i)Cov\{\log(\psi_i),(\theta_i-\beta)^2\}\right]\\
W^{(i)}_{\alpha\gamma\gamma} &= \frac{1}{\gamma^2}-\frac{2}{\gamma^3}E\frac{1}{\psi_i}-\frac{1}{\gamma^4}E\frac{1}{\psi_i^2}+(\frac{n_i}{2}+\alpha)\frac{1}{\gamma^4}\left[2Var(\frac{1}{\psi_i})-Cov\{\log(\psi_i),\frac{1}{\psi_i^2}\}\right]\\
&\quad -(\frac{n_i}{2}+\alpha)^2\frac{1}{\gamma^4}\left[Cov\{\log(\psi_i)\cdot\frac{1}{\psi_i},\frac{1}{\psi_i}\}-Cov\{\log(\psi_i),\frac{1}{\psi_i}\}E\frac{1}{\psi_i}\right.\\
&\quad \left.-E\log(\psi_i)Var(\frac{1}{\psi_i})\right]\\
W^{(i)}_{\alpha\gamma\beta} &= \frac{1}{\gamma^2\tau^2}Cov(\frac{1}{\psi_i},\theta_i-\beta)-(\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2\tau^2}\left[Cov\{\log(\psi_i)\cdot\frac{1}{\psi_i},\theta_i-\beta\}-\right.\\
&\quad \left.Cov\{\log(\psi_i),\theta_i-\beta\}E\frac{1}{\psi_i}-E\log(\psi_i)Cov(\frac{1}{\psi_i},\theta_i-\beta)\right]\\
W^{(i)}_{\alpha\gamma\tau^2} &= \frac{1}{2\gamma^2(\tau^2)^2}Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}-\frac{(\frac{n_i}{2}+\alpha)}{2\gamma^2(\tau^2)^2}\left[Cov\{\log(\psi_i)\cdot\frac{1}{\psi_i},(\theta_i-\beta)^2\}\right.\\
&\quad \left.-Cov\{\log(\psi_i),(\theta_i-\beta)^2\}E\frac{1}{\psi_i}-E\log(\psi_i)Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}\right]\\
W^{(i)}_{\alpha\beta\beta} &= -\frac{1}{(\tau^2)^2}\left[Cov\{\log(\psi_i),(\theta_i-\beta)^2\}-2E(\theta_i-\beta)Cov\{\log(\psi_i),\theta_i-\beta\}\right]\\
W^{(i)}_{\alpha\beta\tau^2} &= \frac{1}{(\tau^2)^2}Cov\{\log(\psi_i),\theta_i-\beta\}-\frac{1}{2(\tau^2)^3}\left[Cov\{\log(\psi_i)\cdot(\theta_i-\beta),(\theta_i-\beta)^2\}-\right.\\
&\quad \left.Cov\{\log(\psi_i),(\theta_i-\beta)^2\}E(\theta_i-\beta)-E\log(\psi_i)Cov\{\theta_i-\beta,(\theta_i-\beta)^2\}\right]\\
W^{(i)}_{\alpha\tau^2\tau^2} &= \frac{1}{4(\tau^2)^3}Cov\{\log(\psi_i),(\theta_i-\beta)^2\}-\frac{1}{4(\tau^2)^4}\left[Cov\{\log(\psi_i)\cdot(\theta_i-\beta)^2,(\theta_i-\beta)^2\}\right.\\
&\quad \left.-Cov\{\log(\psi_i),(\theta_i-\beta)^2\}E(\theta_i-\beta)^2-E\log(\psi_i)Var\{(\theta_i-\beta)^2\}\right]
\end{aligned}
$$

$$
\begin{aligned}
W^{(i)}_{\gamma\gamma\beta} &= -(\frac{n_i}{2}+\alpha)\frac{2}{\gamma^3\tau^2}Cov(\theta_i-\beta,\frac{1}{\psi_i}) + (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^4\tau^2}Cov(\theta_i-\beta,\frac{1}{\psi_i^2}) \\
&\quad +(\frac{n_i}{2}+\alpha)^2\frac{1}{\gamma^4\tau^2}\left\{Cov(\theta_i-\beta,\frac{1}{\psi_i^2}) - 2E\frac{1}{\psi_i}Cov(\theta_i-\beta,\frac{1}{\psi_i})\right\} \\
W^{(i)}_{\gamma\gamma\tau^2} &= -\frac{(\frac{n_i}{2}+\alpha)}{\gamma^3(\tau^2)^2}Cov\{(\theta_i-\beta)^2,\frac{1}{\psi_i}\} + \frac{(\frac{n_i}{2}+\alpha)}{2\gamma^4(\tau^2)^2}Cov\{(\theta_i-\beta)^2,\frac{1}{\psi_i^2}\} \\
&\quad +\frac{(\frac{n_i}{2}+\alpha)^2}{2\gamma^4(\tau^2)^2}\left[Cov\{(\theta_i-\beta)^2,\frac{1}{\psi_i^2}\} - 2E\frac{1}{\psi_i}Cov\{(\theta_i-\beta)^2,\frac{1}{\psi_i}\}\right] \\
W^{(i)}_{\gamma\beta\beta} &= (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2(\tau^2)^2}\left[Cov\{\frac{1}{\psi_i}\cdot(\theta_i-\beta),\theta_i-\beta\} - Cov(\frac{1}{\psi_i},\theta_i-\beta)E(\theta_i-\beta)\right. \\
&\quad \left.-E\frac{1}{\psi_i}Var(\theta_i-\beta)\right] \\
W^{(i)}_{\gamma\beta\tau^2} &= -\frac{(\frac{n_i}{2}+\alpha)}{\gamma^2(\tau^2)^2}Cov(\frac{1}{\psi_i},\theta_i-\beta) + \frac{(\frac{n_i}{2}+\alpha)}{2\gamma^2(\tau^2)^3}\left[Cov\{\frac{1}{\psi_i}\cdot(\theta_i-\beta),(\theta_i-\beta)^2\}\right. \\
&\quad \left.-Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}E(\theta_i-\beta) - E\frac{1}{\psi_i}Cov\{\theta_i-\beta,(\theta_i-\beta)^2\}\right] \\
W^{(i)}_{\gamma\tau^2\tau^2} &= -\frac{(\frac{n_i}{2}+\alpha)}{\gamma^2(\tau^2)^3}Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\} + \frac{(\frac{n_i}{2}+\alpha)}{4\gamma^2(\tau^2)^4}\left[Cov\{\frac{1}{\psi_i}\cdot(\theta_i-\beta)^2,(\theta_i-\beta)^2\}\right. \\
&\quad \left.-Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}E(\theta_i-\beta)^2 - E\frac{1}{\psi_i}Var\{(\theta_i-\beta)^2\}\right] \\
W^{(i)}_{\beta\beta\tau^2} &= \frac{1}{(\tau^2)^2} - \frac{2}{(\tau^2)^3}Var(\theta_i-\beta) + \frac{1}{2(\tau^2)^4}\left[Var\{(\theta_i-\beta)^2\}\right. \\
&\quad \left.-2E(\theta_i-\beta)Cov\{(\theta_i-\beta)^2,\theta_i-\beta\}\right] \\
W^{(i)}_{\beta\tau^2\tau^2} &= \frac{2}{(\tau^2)^3}E(\theta_i-\beta) - \frac{2}{(\tau^2)^4}Cov\{(\theta_i-\beta)^2,\theta_i-\beta\} \\
&\quad +\frac{1}{4(\tau^2)^5}\left[Cov\{(\theta_i-\beta)^3,(\theta_i-\beta)^2\} - Var\{(\theta_i-\beta)^2\}E(\theta_i-\beta)\right]
\end{aligned}
$$

All the other terms are equal to the above values according to symmetry.

3.5.2 Second Order Correction of ν_i

In the equation (3.20), the second order derivative of ν_i is included. The detailed formula is given in the following:

$$\frac{\partial^2 \nu_i}{\partial \alpha^2} = -\frac{\partial}{\partial \alpha} Cov(\theta_i^2, \log\psi_i) + 2\frac{\partial}{\partial \alpha} E\theta_i \cdot Cov(\theta_i, \log\psi_i) + 2E\theta_i \frac{\partial}{\partial \alpha} Cov(\theta_i, \log\psi_i)$$
$$\frac{\partial}{\partial \alpha} Cov(\theta_i^2, \log\psi_i) = -Cov(\theta_i^2 \log\psi_i, \log\psi_i) + Cov(\theta_i^2, \log\psi_i) E\log\psi_i + E\theta_i^2 Cov(\log\psi_i, \log\psi_i)$$
$$\frac{\partial}{\partial \alpha} E\theta_i = -Cov(\theta_i, \log\psi_i)$$
$$\frac{\partial}{\partial \alpha} Cov(\theta_i, \log\psi_i) = -Cov(\theta_i \log\psi_i, \log\psi_i) + Cov(\theta_i, \log\psi_i) E\log\psi_i + E\theta_i Cov(\log\psi_i, \log\psi_i)$$

$$\frac{\partial^2 \nu_i}{\partial \alpha \partial \gamma} = -\frac{\partial}{\partial \gamma} Cov(\theta_i^2, \log\psi_i) + 2\frac{\partial}{\partial \gamma} E\theta_i \cdot Cov(\theta_i, \log\psi_i) + 2E\theta_i \frac{\partial}{\partial \gamma} Cov(\theta_i, \log\psi_i)$$
$$\frac{\partial}{\partial \gamma} Cov(\theta_i^2, \log\psi_i) = (n_i/2 + \alpha)\frac{1}{\gamma^2} \Big\{ Cov(\theta_i^2 \log\psi_i, \frac{1}{\psi_i}) - Cov(\theta_i^2, \frac{1}{\psi_i}) E\log\psi_i$$
$$-E\theta_i^2 Cov(\log\psi_i, \frac{1}{\psi_i}) \Big\} - \frac{1}{\gamma^2} Cov(\theta_i^2, \frac{1}{\psi_i})$$
$$\frac{\partial}{\partial \gamma} E\theta_i = (n_i/2 + \alpha)\frac{1}{\gamma^2} Cov(\theta_i, \frac{1}{\psi_i})$$
$$\frac{\partial}{\partial \gamma} Cov(\theta_i, \log\psi_i) = (n_i/2 + \alpha)\frac{1}{\gamma^2} \Big\{ Cov(\theta_i \log\psi_i, \frac{1}{\psi_i}) - Cov(\theta_i, \frac{1}{\psi_i}) E\log\psi_i$$
$$-E\theta_i Cov(\log\psi_i, \frac{1}{\psi_i}) \Big\} - \frac{1}{\gamma^2} Cov(\theta_i, \frac{1}{\psi_i})$$

$$\frac{\partial^2 \nu_i}{\partial \alpha \partial \beta} = -\frac{\partial}{\partial \beta} Cov(\theta_i^2, \log\psi_i) + 2\frac{\partial}{\partial \beta} E\theta_i \cdot Cov(\theta_i, \log\psi_i) + 2E\theta_i \frac{\partial}{\partial \beta} Cov(\theta_i, \log\psi_i)$$
$$\frac{\partial}{\partial \beta} Cov(\theta_i^2, \log\psi_i) = \frac{1}{\tau^2} \Big\{ Cov(\theta_i^2 \log\psi_i, \theta_i - \beta) - Cov(\theta_i^2, \theta_i - \beta) E\log\psi_i$$
$$-E\theta_i^2 Cov(\log\psi_i, \theta_i - \beta) \Big\}$$
$$\frac{\partial}{\partial \beta} E\theta_i = \frac{1}{\tau^2} Var(\theta_i)$$
$$\frac{\partial}{\partial \beta} Cov(\theta_i, \log\psi_i) = \frac{1}{\tau^2} \big\{ Cov(\theta_i \log\psi_i, \theta_i - \beta) - Cov(\theta_i, \theta_i - \beta) E\log\psi_i$$
$$-E\theta_i Cov(\log\psi_i, \theta_i - \beta) \big\}$$

$$\frac{\partial^2 \nu_i}{\partial\alpha\partial\tau^2} = -\frac{\partial}{\partial\tau^2}Cov(\theta_i^2, \log\psi_i) + 2\frac{\partial}{\partial\tau^2}E\theta_i \cdot Cov(\theta_i, \log\psi_i) + 2E\theta_i\frac{\partial}{\partial\tau^2}Cov(\theta_i, \log\psi_i)$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i^2, \log\psi_i) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i^2\log\psi_i, (\theta_i-\beta)^2\} - Cov\{\theta_i^2, (\theta_i-\beta)^2\}E\log\psi_i$$
$$-E\theta_i^2 Cov\{\log\psi_i, (\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial}{\partial\tau^2}E\theta_i = \frac{1}{2(\tau^2)^2}Cov\{\theta_i, (\theta_i-\beta)^2\}$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i, \log\psi_i) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i\log\psi_i, (\theta_i-\beta)^2\} - Cov\{\theta_i, (\theta_i-\beta)^2\}E\log\psi_i$$
$$-E\theta_i Cov\{\log\psi_i, (\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial^2 \nu_i}{\partial\gamma^2} = -(\frac{n_i}{2}+\alpha)\frac{2}{\gamma^3}\left\{Cov(\theta_i^2, \frac{1}{\psi_i}) - 2E\theta_i Cov(\theta_i, \frac{1}{\psi_i})\right\} + (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\cdot$$
$$\left\{\frac{\partial}{\partial\gamma}Cov(\theta_i^2, \frac{1}{\psi_i}) - 2\frac{\partial}{\partial\gamma}E\theta_i \cdot Cov(\theta_i, \frac{1}{\psi_i}) - 2E\theta_i\frac{\partial}{\partial\gamma}Cov(\theta_i, \frac{1}{\psi_i})\right\}$$

$$\frac{\partial}{\partial\gamma}Cov(\theta_i^2, \frac{1}{\psi_i}) = (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\left\{Cov(\theta_i^2\frac{1}{\psi_i}, \frac{1}{\psi_i}) - Cov(\theta_i^2, \frac{1}{\psi_i})E\frac{1}{\psi_i} - E\theta_i^2 Var(\frac{1}{\psi_i})\right\}$$
$$+\frac{1}{\gamma^2}Cov(\theta_i^2, \frac{1}{\psi_i^2})$$

$$\frac{\partial}{\partial\gamma}E\theta_i = (n_i/2+\alpha)\frac{1}{\gamma^2}Cov(\theta_i, \frac{1}{\psi_i})$$

$$\frac{\partial}{\partial\gamma}Cov(\theta_i, \frac{1}{\psi_i}) = (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\left\{Cov(\theta_i\frac{1}{\psi_i}, \frac{1}{\psi_i}) - Cov(\theta_i, \frac{1}{\psi_i})E\frac{1}{\psi_i} - E\theta_i Var(\frac{1}{\psi_i})\right\}$$
$$+\frac{1}{\gamma^2}Cov(\theta_i, \frac{1}{\psi_i^2})$$

$$\frac{\partial^2 \nu_i}{\partial\gamma\partial\beta} = (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\left\{\frac{\partial}{\partial\beta}Cov(\theta_i^2, \frac{1}{\psi_i}) - 2\frac{\partial}{\partial\beta}E\theta_i \cdot Cov(\theta_i, \frac{1}{\psi_i}) - 2E\theta_i\frac{\partial}{\partial\beta}Cov(\theta_i, \frac{1}{\psi_i})\right\}$$

$$\frac{\partial}{\partial\beta}Cov(\theta_i^2, \frac{1}{\psi_i}) = \frac{1}{\tau^2}\left\{Cov(\theta_i^2\frac{1}{\psi_i}, \theta_i-\beta) - Cov(\theta_i^2, \theta_i-\beta)E\frac{1}{\psi_i} - E\theta_i^2 Cov(\frac{1}{\psi_i}, \theta_i-\beta)\right\}$$

$$\frac{\partial}{\partial\beta}E\theta_i = \frac{1}{\tau^2}Var(\theta_i)$$

$$\frac{\partial}{\partial\beta}Cov(\theta_i, \frac{1}{\psi_i}) = \frac{1}{\tau^2}\left\{Cov(\theta_i\frac{1}{\psi_i}, \theta_i-\beta) - Cov(\theta_i, \theta_i-\beta)E\frac{1}{\psi_i} - E\theta_i Cov(\frac{1}{\psi_i}, \theta_i-\beta)\right\}$$

$$\frac{\partial^2 \nu_i}{\partial\gamma\partial\tau^2} = (\frac{n_i}{2}+\alpha)\frac{1}{\gamma^2}\left\{\frac{\partial}{\partial\tau^2}Cov(\theta_i^2,\frac{1}{\psi_i}) - 2\frac{\partial}{\partial\tau^2}E\theta_i \cdot Cov(\theta_i,\frac{1}{\psi_i}) - 2E\theta_i\frac{\partial}{\partial\tau^2}Cov(\theta_i,\frac{1}{\psi_i})\right\}$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i^2,\frac{1}{\psi_i}) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i^2\frac{1}{\psi_i},(\theta_i-\beta)^2\} - Cov\{\theta_i^2,(\theta_i-\beta)^2\}E\frac{1}{\psi_i} - E\theta_i^2 Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial}{\partial\tau^2}E\theta_i = \frac{1}{2(\tau^2)^2}Cov\{\theta_i,(\theta_i-\beta)^2\}$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i,\frac{1}{\psi_i}) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i\frac{1}{\psi_i},(\theta_i-\beta)^2\} - Cov\{\theta_i,(\theta_i-\beta)^2\}E\frac{1}{\psi_i} - E\theta_i Cov\{\frac{1}{\psi_i},(\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial^2 \nu_i}{\partial\beta^2} = \frac{1}{\tau^2}\left\{\frac{\partial}{\partial\beta}Cov(\theta_i^2,\theta_i-\beta) - 2\frac{\partial}{\partial\beta}E\theta_i \cdot Cov(\theta_i,\theta_i-\beta) - 2E\theta_i\cdot\frac{\partial}{\partial\beta}Cov(\theta_i,\theta_i-\beta)\right\}$$

$$\frac{\partial}{\partial\beta}Cov(\theta_i^2,\theta_i-\beta) = \frac{1}{\tau^2}\Big[Cov\{\theta_i^2(\theta_i-\beta),\theta_i-\beta\} - Cov(\theta_i^2,\theta_i-\beta)E(\theta_i-\beta) - E\theta_i^2 Var(\theta_i-\beta)\Big]$$

$$\frac{\partial}{\partial\beta}E\theta_i = Cov(\theta_i,\theta_i-\beta)$$

$$\frac{\partial}{\partial\beta}Cov(\theta_i,\theta_i-\beta) = \frac{1}{\tau^2}\left[Cov\{\theta_i(\theta_i-\beta),\theta_i-\beta\} - Cov(\theta_i,\theta_i-\beta)E(\theta_i-\beta) - E\theta_i Var(\theta_i-\beta)\right]$$

$$\frac{\partial^2 \nu_i}{\partial\beta\partial\tau^2} = -\frac{1}{(\tau^2)^2}\left\{Cov(\theta_i^2,\theta_i-\beta) - 2E\theta_i\cdot Cov(\theta_i,\theta_i-\beta)\right\} + \frac{1}{\tau^2}\left\{\frac{\partial}{\partial\tau^2}Cov(\theta_i^2,\theta_i-\beta) - 2\frac{\partial}{\partial\tau^2}E\theta_i\cdot Cov(\theta_i,\theta_i-\beta) - 2E\theta_i\cdot\frac{\partial}{\partial\tau^2}Cov(\theta_i,\theta_i-\beta)\right\}$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i^2,\theta_i-\beta) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i^2(\theta_i-\beta),(\theta_i-\beta)^2\} - Cov\{\theta_i^2,(\theta_i-\beta)^2\}E(\theta_i-\beta) - E\theta_i^2 Cov\{\theta_i-\beta,(\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial}{\partial\tau^2}E\theta_i = \frac{1}{2(\tau^2)^2}Cov\{\theta_i,(\theta_i-\beta)^2\}$$

$$\frac{\partial}{\partial\tau^2}Cov(\theta_i,\theta_i-\beta) = \frac{1}{2(\tau^2)^2}\Big[Cov\{\theta_i(\theta_i-\beta),(\theta_i-\beta)^2\} - Cov\{\theta_i,(\theta_i-\beta)^2\}E(\theta_i-\beta) - E\theta_i Cov\{\theta_i-\beta,(\theta_i-\beta)^2\}\Big]$$

$$\frac{\partial^2 \nu_i}{\partial {\tau^2}^2} = -\frac{1}{(\tau^2)^3}\left[Cov\{\theta_i^2, (\theta_i-\beta)^2\} - 2E\theta_i Cov\{\theta_i, (\theta_i-\beta)^2\}\right] +$$
$$\frac{1}{2(\tau^2)^2}\cdot\left[\frac{\partial}{\partial\tau^2}Cov\{\theta_i^2, (\theta_i-\beta)^2\} - 2\frac{\partial}{\partial\tau^2}E\theta_i \cdot Cov\{\theta_i, (\theta_i-\beta)^2\}\right.$$
$$\left.-2E\theta_i \cdot \frac{\partial}{\partial\tau^2}Cov\{\theta_i, (\theta_i-\beta)^2\}\right]$$
$$\frac{\partial}{\partial\tau^2}Cov\{\theta_i^2, (\theta_i-\beta)^2\} = \frac{1}{2(\tau^2)^2}\left[Cov\{\theta_i^2(\theta_i-\beta)^2, (\theta_i-\beta)^2\} - E\theta_i^2 Var\{(\theta_i-\beta)^2\}\right.$$
$$\left.-Cov\{\theta_i^2, (\theta_i-\beta)^2\}E(\theta_i-\beta)^2\right]$$
$$\frac{\partial}{\partial\tau^2}E\theta_i = \frac{1}{2(\tau^2)^2}Cov\{\theta_i, (\theta_i-\beta)^2\}$$
$$\frac{\partial}{\partial\tau^2}Cov\{\theta_i, (\theta_i-\beta)^2\} = \frac{1}{2(\tau^2)^2}\left[Cov\{\theta_i(\theta_i-\beta)^2, (\theta_i-\beta)^2\} - E\theta_i Var\{(\theta_i-\beta)^2\}\right.$$
$$\left.-Cov\{\theta_i, (\theta_i-\beta)^2\}E(\theta_i-\beta)^2\right]$$

The other terms are equal to the symetric terms.

Chapter 4

Confidence Interval Estimation of Small Area Parameters Shrinking both Mean and Variances

4.1 Introduction

The new approach to small area estimation based on joint modeling of mean and variances is given in the previous chapter. The conditional mean squared error of prediction is also estimated to evaluate the prediction error. In this chapter, we will obtain confidence intervals of small area means. The small area estimation literature is dominated by point estimation and their standard errors. It is well known that the standard practice of (pt. est. $\pm\, q$s.e.), q is Z (standard normal) or t cut-off point, does not produce accurate intervals. See, Hall and Maiti (2006) and Chatterjee et al. (2008) for more details. The previous works are based on the bootsrap procedure and has limited use due to repeated estimation of model parameters.

The confidence intervals produced in this chapter are from a decision theory perspective.

The rest of the chapter is organized as follows. In section 4.2, the proposed model is repeated. Section 4.3 gives the definition of the confidence intervals. Theoretical justification and an alternative model are given in section 4.4 and 4.5. Section 4.6 contains a simulation study. A real data analysis is included in Section 4.7.

4.2 Proposed Model

The hierarchical model adopted here is already introduced in section 3.2. We want to point out again that this model used both the direct survey estimates and sampling variance estimates to estimate all the parameters that determines the stochastic system. From the simulation study, this mean-variance joint modelling performed better than Hwang et al. (2009), which does not take into account the variance estimation completely.

Let X_i be the direct survey estimates and S_i^2 be their sampling variances for the i^{th} area, $i = 1, \cdots, n$. Let $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ip})^T$ be the set of covariates available at the estimation stage and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ be the associated regression coefficients. We propose the following hierarchical model:

$$\left.\begin{aligned} &X_i|\theta_i, \sigma_i^2 \sim \text{Normal}(\theta_i, \sigma_i^2) \\ &\theta_i \sim \text{Normal}(\boldsymbol{Z}_i^T\boldsymbol{\beta}, \tau^2); i = 1, \cdots, n, \end{aligned}\right\} \tag{4.1}$$

$$\left.\begin{aligned} &(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i - 1} \\ &\sigma_i^{-2} \sim \text{Gamma}\{\alpha, \gamma\}; i = 1, \cdots, n, \end{aligned}\right\} \tag{4.2}$$

where $\boldsymbol{B} = (\alpha, \gamma, \boldsymbol{\beta}^T, \tau^2)^T$, referred to as the structural parameters, are unknown and n_i represents the sample size for a simple random sample (SRS) from the i^{th} area. Note that σ_i^2 are the sampling variances of X_i's and are usually estimated by S_i^2's. Note that the chi-square distribution for the sample variance is valid for only a random sample. For a complex survey design the degrees of freedom of the chi-square distribution need to be determined carefully (e.g., Maples and Huang 2009). The second level of (3.2) might be further extended as Gamma$\{\gamma, \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2)/\gamma\}$ to accommodate covaraite information in the variance modeling. The inference can be made from the conditional distribution of θ_i (the parameter of interest) given the data $(X_i, S_i^2, \boldsymbol{Z}_i), i = 1, \cdots, n$. Under our model set up the conditional distribution of θ_i does not have a closed form, and for handling a non-standard distribution we use Monte Carlo methods.

4.3 Confidence Interval

4.3.1 Definition

Following Joshi (1969), Casella and Hwang (1991), Hwang et al. (2009), consider the loss function

$$(k/\sigma)L(C) - I_C(\theta)$$

where k is a tuning parameter, independent of the model parameters and $L(C)$ is the length of a confidence interval C and $I_C(\theta)$ is 1 or 0 depending on $\theta \in C$ or not. Then the decision Bayes confidence interval for θ_i is obtained by minimizing $E\left\{[(k/\sigma)L(C) - I_C(\theta)]|X_i, S_i^2)\right\}$

and is given by

$$C_i(\boldsymbol{B}) = \{\theta : kE(\sigma_i^{-1}|X_i, S_i^2, \boldsymbol{B}) < \pi(\theta_i|X_i, S_i^2, \boldsymbol{B})\}. \tag{4.3}$$

When k and $\boldsymbol{B}$ are known we follow the following steps to calculate the above CI. Note that the conditional distribution of σ_i^2 and θ_i given the data and $\boldsymbol{B}$ in equation (3.4) and (3.5) are

$$\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B}) \propto \frac{\exp[-0.5(X_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/(\sigma_i^2+\tau^2) - \{0.5(n_i-1)S_i^2 + 1/b\}(1/\sigma_i^2)]}{(\sigma_i^2)^{(n_i-1)/2+a+1}(\sigma_i^2+\tau^2)^{1/2}}$$

$$\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) \propto \exp\{-0.5(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2/\tau^2\}\psi_i^{-(n_i/2+a)}$$

where $\psi_i = 0.5(X_i - \theta_i)^2 + 0.5(n_i-1)S_i^2 + 1/\gamma$.

Therefore, we calculate $E(\sigma_i^{-1}|X_i, S_i^2, \boldsymbol{B})$ using the Monte Carlo method. Suppose that σ_{ik}^2, $k = 1, \cdots, N$ are N observations from $\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B})$. Then $E(\sigma_i^{-1}|X_i, S_i^2, \boldsymbol{B})$ will be calculated by

$$\frac{1}{N}\sum_{k=1}^{N}\frac{1}{\sigma_{ik}}.$$

For drawing random numbers from $\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B})$ we use the rejection sampling method (Robert and Casella, 2004). Next we determine the values of θ_i by solving $kE(\sigma_i^{-1}|X_i, S_i^2, \boldsymbol{B}) - \pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) = 0$. Note that for solving the above equation we require the normalizing constant of the conditional density $\pi(\theta_i|X_i, S_i^2, \boldsymbol{B})$, and that is calculated by a numerical method.

In practice, $\boldsymbol{B}$ is unknown. We propose to estimate it by maximizing the marginal likelihood $L_M = \prod_{i=1}^n L_i^M$ of $\{(X_i, S_i^2, \boldsymbol{Z}_i)_{i=1}^n; \boldsymbol{B}\}$. The detailed prodecure is described

in section 3.2.3.

4.3.2 Choice of the Tuning Parameter

We discuss in this section the choice of the tuning parameter k in (4.3). As a first step when a is known, consider

$$k = k(\boldsymbol{B}) = u_0 \, \phi \left(t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right) \tag{4.4}$$

where ϕ is the standard normal distribution, $t_{\alpha/2}$ is $(1-\alpha/2)$-th percentile of t distribution with $(n_i - 1)$ degrees of freedom, and

$$u_0 = \sqrt{1 + \frac{\sigma_i^2}{\tau^2}}. \tag{4.5}$$

This choice of k involves components of $\boldsymbol{B}$ which will be assumed to be known for the moment. In the case when $\boldsymbol{B}$ is unknown, as is in most practical situations, $\boldsymbol{B}$ will be replaced by its corresponding estimate $\hat{\boldsymbol{B}}$ derived using the estimation method described in Section 3.2.3. The definition of u_0 in (4.5) further involves σ_i^2 which is not a component of $\boldsymbol{B}$. Thus, u_0 is replaced by $\hat{u}_0$ obtained by replacing σ_i^2 with its maximum a posteriori estimate

$$\hat{\sigma}_i^2 = \hat{\sigma}_i^2(\boldsymbol{B}) = \arg\max_{\sigma_i^2} \pi(\sigma_i^2 \mid X_i, S_i^2, \boldsymbol{B}), \tag{4.6}$$

the value of σ_i^2 that maximizes the posterior density of σ_i^2 given X_i, S_i^2 and $\boldsymbol{B}$.

We demonstrate that the coverage probability of $C_i(\boldsymbol{B})$ with this choice of k is close to $1-\alpha$. Theoretical justifications are provided in Section 4.4.

In Hwang et al. (2009) the choice of k is ad-hoc where they equate the Bayes interval to the t-interval and solve for k.

Note that without any hierarchical model S_i and X_i are independent as S_i^2 and X_i are ancillary and complete sufficient statistics for θ_i, respectively. However, under models (3.1) and (3.2) the conditional distribution of σ_i^2 and θ_i involve both X_i and S_i^2 which is seen from (3.4) and (3.5).

We would like to mention that in Hwang et al. (2009) the shrinkage estimators for σ_i^2 was based only on the information on S_i^2, but not using both X_i and S_i^2. They then plugged-in the Bayes estimator of σ_i into the Bayes estimators of small area parameters. Thus if $\hat{\sigma}^2_{B,i}$ is the Bayes estimator of σ_i^2, then Hwang et al.'s small area estimator can be written as $E(\theta_i|X_i, \sigma_i^2)|_{\sigma_i^2=\hat{\sigma}^2_{B,i}}$.

Remark 1. As it is mentioned previously the d.f. associated with the χ^2 distribution for sampling variance modeling need not be simply $n_i - 1$, n_i being the sample size for i-th area. There is no sound theoretical result for determining the d.f. in the case of complex survey design. Wang and Fuller (2003) approximated the χ^2 to the normal based on the Wilson-Hilferty approximation. If one knows the exact sampling design then the simulation based guideline of Maples et al. (2009) could be useful. In the case for county level estimation from the American Community Survey, they suggested the estimated d.f. $= 0.36 \times \sqrt{n_i}$.

4.4 Theoretical Justification of Tuning Parameter

We present some theoretical justification for the choice of k according to equations (4.4), (4.5) and (4.6). Assume $\boldsymbol{B}$ is fixed and known for the moment. The conditional distribution

of θ_i can be approximated as

$$\begin{aligned}\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) &= \textstyle\int_0^\infty \pi(\theta_i \,|\, X_i, S_i^2, \boldsymbol{B}, \sigma_i^2)\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B})\, d\sigma_i^2 \\ &\approx \pi(\theta_i \,|\, X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2)\end{aligned} \tag{4.7}$$

where $\hat{\sigma}_i^2$ as defined in (4.6). In a similar way, approximate $E(\sigma_i^{-1} \,|\, X_i, S_i^2, \boldsymbol{B})$ by

$$E(\sigma_i^{-1}|X_i, S_i^2, \boldsymbol{B}) \approx \hat{\sigma}_i^{-1}. \tag{4.8}$$

Based on (4.7) and (4.8), we have $C_i(\boldsymbol{B}) \approx \tilde{C}_i(\boldsymbol{B})$ where $\tilde{C}_i(\boldsymbol{B})$ is the confidence interval for θ_i given by

$$\tilde{C}_i(\boldsymbol{B}) = \left\{ \theta_i \,:\, \pi(\theta_i \,|\, X_i, S_i^2, \boldsymbol{B}, \hat{\sigma}_i^2) \geq k\, \hat{\sigma}_i^{-1} \right\}, \tag{4.9}$$

with σ_i^2 replaced by $\hat{\sigma}_i^2$. From (3.1), it follows that the conditional density $\pi(\theta_i \,|\, X_i, S_i^2, \boldsymbol{B}, \sigma_i^2)$ is normal with mean μ_i and variance v_i, where μ_i and v_i are given by the expressions

$$\mu_i = w_i\, X_i + (1 - w_i)\boldsymbol{Z}_i^T\boldsymbol{\beta} \;\text{ and }\quad v_i = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} = \sigma_i^2\left(1 + \frac{\sigma_i^2}{\tau^2}\right)^{-1}, \tag{4.10}$$

and $w_i = \dfrac{1/\sigma_i^2}{(1/\sigma_i^2 + 1/\tau^2)}$. Now, choosing $k = \hat{u}_0\, \phi\left(\, t_{\alpha/2}\sqrt{\frac{n_i+2a+2}{n_i-1}}\,\right)$ as discussed, the confidence interval $\tilde{C}_i(\boldsymbol{B})$ becomes

$$\tilde{C}_i(\boldsymbol{B}) = \left\{ \theta_i : \hat{u}_0 \frac{|\theta_i - \hat{\mu}_i|}{\hat{\sigma}_i} \leq\, t_{\alpha/2}\sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right\}, \tag{4.11}$$

where $\hat{\mu}_i$ is the expression for μ_i in (4.10) with σ_i^2 replaced by $\hat{\sigma}_i^2$. Now consider the behavior of $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\boldsymbol{B})$ as τ^2 ranges between 0 and ∞. When $\tau^2 \to \infty$, $\hat{\sigma}_i^2$ converges to

$$\hat{\sigma}^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, \infty) = \frac{\frac{(n_i-1)}{2} S_i^2 + \frac{1}{b}}{\frac{n_i-1}{2} + a + 1} = \frac{(n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 1}.$$

Similarly, when $\tau^2 \to 0$, $\hat{\sigma}_i^2$ converges to

$$\hat{\sigma}^2(0) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, 0) = \frac{(X_i - \boldsymbol{Z}_i^T \boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 2}.$$

For all intermediate values of τ^2, we have $\min\{\hat{\sigma}^2(0), \hat{\sigma}^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}^2(0), \hat{\sigma}^2(\infty)\}$. Therefore, it is sufficient to consider the following two cases: (i) $\hat{\sigma}_i^2 \geq \hat{\sigma}^2(\infty)$, where it follows that $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2 + \frac{2}{b} + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2$, and (ii) $\hat{\sigma}_i^2 \geq \hat{\sigma}^2(0)$, where it follows that $(n_i+2a+2)\hat{\sigma}_i^2 = (X_i - \boldsymbol{Z}_i^T \boldsymbol{\beta})^2 + (n_i-1)S_i^2 + \frac{2}{b} \geq (n_i-1)S_i^2$. So, in both cases (i) and (ii),

$$(n_i + 2a + 2)\, \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \tag{4.12}$$

Since $\theta_i - \mu_i \sim N\left(0, \sigma_i^2\tau^2/(\sigma_i^2 + \tau^2)\right)$ and $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i-1}$, the confidence interval

$$D_i = \left\{ \theta_i \,:\, u_0 \, \frac{|\theta_i - \mu_i|}{S_i} \leq t_{\alpha/2} \right\} \tag{4.13}$$

has coverage probability $1 - \alpha$. Thus, if u_0 and μ_i are replaced by $\hat{u}_0$ and $\hat{\mu}_i$, it is expected that the resulting confidence interval $\tilde{D}_i$, say, will have coverage probability of approximately $1 - \alpha$. From (4.12), we have

$$P(\tilde{C}(\boldsymbol{B})) \geq P(\tilde{D}_i) \approx 1 - \alpha, \tag{4.14}$$

establishing an approximate lower bound of $1 - \alpha$ for the confidence level of $\tilde{C}(\boldsymbol{B})$.

In (4.14), $\boldsymbol{B}$ was assumed fixed and known. In the case when $\boldsymbol{B}$ is unknown, we replace $\boldsymbol{B}$ by its marginal maximum likelihood estimate $\hat{\boldsymbol{B}}$. Since (4.14) holds regardless of the true value of $\boldsymbol{B}$, substituting $\hat{\boldsymbol{B}}$ for $\boldsymbol{B}$ in (4.14) will involve an order $O(1/\sqrt{N})$ of error where $N = \sum_{i=1}^{n} n_i$. Compared to each single n_i, this pooling of n_is is expected to reduce the error significantly so that $\tilde{C}(\hat{\boldsymbol{B}})$ is sufficiently close to $\tilde{C}(\boldsymbol{B})$ to satisfy the lower bound of $1 - \alpha$ in (4.14).

4.5 Alternative Model for Confidence Interval

It is possible to reduce the width of the confidence interval $\tilde{C}(\boldsymbol{B})$ based on an alternative hierarchical model for small area estimation. The constant term $n_i + 2a + 2$ in (4.12) in the alternative model becomes $n_i + 2a$. The model is

$$X_i \,|\, \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \tag{4.15}$$

$$\theta_i \,|\, \sigma_i^2 \sim N(\boldsymbol{Z}_i\boldsymbol{\beta}, \lambda\, \sigma_i^2), \tag{4.16}$$

$$(n_i - 1)(S_i^2/\sigma_i^2) \sim \chi^2_{n_i - 1}, \tag{4.17}$$

$$\sigma_i^2 \sim \text{Inverse-Gamma}(a, b), \tag{4.18}$$

for $i = 1, 2, \cdots, n$. Note that in the above alternative formulation, it is assumed that the variability of θ_i is proportional to σ_i^2 while in the previous model, the variance of θ_i was τ^2

independent of σ_i^2; in other words the model variance was constant. The set of all unknown parameters in the current hierarchical model is $\boldsymbol{B} = (a, b, \boldsymbol{\beta}, \lambda)$. By re-parametrizing the variance in (4.16), some simplications can be obtained in the derivation of the posteriors of θ_i and σ_i given X_i, S_i^2 and $\boldsymbol{B}$. We have

$$\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B}) = IG\left(\frac{n_i}{2} + a, \left[\frac{(n_i-1)S_i^2}{2} + \frac{(X_i - \boldsymbol{Z}_i\boldsymbol{\beta})^2}{2(1+\lambda)} + \frac{1}{b}\right]^{-1}\right).$$

Given $\boldsymbol{B}$ and σ_i^2, the conditional distribution of θ_i is normal with mean μ_i and variance v_i as in (4.10) with τ^2 replaced by $\lambda\,\sigma_i^2$: $\pi(\theta_i|X_i, \sigma_i^2, \boldsymbol{B}) = N(\mu_i, \frac{\lambda\sigma_i^2}{1+\lambda})$. By integrating out σ_i^2, it follows that the conditional distribution of θ_i given X_i, S_i^2 and $\boldsymbol{B}$ is

$$\begin{aligned}\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) &= \int_0^\infty \pi(\theta_i|X_i, \sigma_i^2, B)\pi(\sigma_i^2|X_i, S_i^2, \boldsymbol{B})d\sigma_i^2 \\ &\propto \left[\frac{(1+\lambda)}{2\lambda}(\theta_i - \mu_i)^2 + \frac{\delta^2}{2}\right]^{-(n_i+2a+1)/2}, \end{aligned} \tag{4.19}$$

where $\delta^2 = (n_i-1)S_i^2 + (X_i - \boldsymbol{Z}_i\boldsymbol{\beta})^2/(1+\lambda) + 2/b$. We can rewrite (4.19) as

$$\pi(\theta_i|X_i, S_i^2, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2 + a)\sqrt{1+\lambda}}{\delta^*\,\Gamma(n_i/2+a)\sqrt{(n_i+2a)\lambda\,\pi}}\left\{1 + \frac{(\theta_i - \mu_i)^2}{(n_i+2a)\delta^{*2}\,\lambda/(1+\lambda)}\right\}^{-\frac{(n_i+2a+1)}{2}}$$

which can be seen to be a scaled t-distribution with $n_i + 2a$ degrees of freedom and scale parameter $\delta^*\sqrt{\frac{\lambda}{1+\lambda}}$ where $\delta^{*2} = \frac{\delta^2}{(n_i+2a)}$. Also

$$E(\sigma_i^{-1}|X_i, S_i, \boldsymbol{B}) = \frac{\Gamma((n_i+1)/2+a)(\delta^2/2)^{-\{(n_i+1)/2+a\}}}{\Gamma(n_i/2+a)(\delta^2/2)^{-(n_i/2+a)}}$$
$$= \frac{\Gamma((n_i+1)/2+a)}{\Gamma(n_i/2+a)} \frac{\sqrt{2}}{\delta^* \sqrt{n_i+2a}}.$$

In this context, choosing $k = k(\boldsymbol{B})$ as

$$k = \left\{ 1 + \frac{t^2_{\alpha/2}}{n_i - 1} \right\}^{-(n_i+2a+1)/2} \cdot \sqrt{\frac{1+\lambda}{\lambda}} \cdot \frac{1}{\sqrt{2\,\pi}},$$

the confidence interval in (4.3) simplifies to

$$C_i(\boldsymbol{B}) \equiv \left\{ \theta_i : \frac{|\,\theta_i - \mu_i\,|}{\sqrt{\frac{\lambda}{1+\lambda} \frac{(n_i+2a)\delta^{*2}}{n_i-1}}} \leq t_{\alpha/2} \right\}. \tag{4.20}$$

Using similar arguments as before and noting that $(n_i + 2a)\delta^{*2} \geq (n_i - 1)S_i^2$, we have $P(C_i(\boldsymbol{B}) \geq D_i) = 1 - \alpha$ where D_i is the confidence interval in (4.13). Again here, $\boldsymbol{B}$ was assumed fixed and known. In the case when $\boldsymbol{B}$ is unknown, we replace $\boldsymbol{B}$ by its marginal maximum likelihood estimate $\hat{\boldsymbol{B}}$. It is expected that the pooling technique will result in an error small enough so that $P(C_i(\hat{\boldsymbol{B}})) \approx P(C_i(\boldsymbol{B}))$, and thus, enable the confidence level of $C_i(\hat{\boldsymbol{B}})$ to be greater than $1 - \alpha$.

4.6 Simulation Study

Simulation design We considered a simulation setting similar to Wang and Fuller (2003), which is the same as the set up in section 3.4. Each sample in the simulation study was generated through the same steps as the section 3.4. To simplify the simulation, we still do not choose any covariate Z, only (X_i, S_i^2) are generated. The observations for each small area is first generated as

$$X_{ij} = \beta + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \ i = 1, \ldots, n,$$

where $u_i \sim \text{Normal}(0, \tau^2)$ and $e_{ij} \sim \text{Normal}(0, n_i\sigma_i^2)$. Then, the model for (X_i, S_i^2) is the same as equations (3.26) and (3.27):

$$\begin{aligned} X_i &= \beta + u_i + e_i, \quad i = 1, \ldots, n, \\ S_i^2 &= \frac{1}{n_i - 1}\frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 \end{aligned}$$

where $X_i = \bar{X_{i\cdot}} = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$, $e_i = \bar{e_{i\cdot}} = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Therefore, $X_i \sim \text{Normal}(\theta_i, \sigma_i^2)$, where $\theta_i = \beta + u_i$ and $\theta_i \sim \text{Normal}(\beta, \tau^2)$, and $e_i \sim \text{Normal}(0, \sigma_i^2)$. It is to be noted that $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{(n_i-1)}$. Then the quantities to be predicted are the mean for each small area, θ_i, $i = 1, \ldots, n$.

Like Wang and Fuller (2003), we set all n_i equal to m which eased our programming efforts. However, the sampling variances were still unequal by choosing one-third of the σ_i^2 equal to 1, one-third equal to 4, and one-third equal to 16. In the simulation, we set $\beta = 10$ and took three different values of τ^2, 0.25, 1, 4. For each τ^2, we generated 200 samples for

each of the combinations $(m, n) = (9, 36)$, $(18, 180)$.

In this simulation study we compare the proposed method with the method of Wang and Fuller (2003), Hwang *et al.* (2009) and Qiu and Hwang (2007) which are referred to as I, II, III, and IV, respectively. Note that the estimator proposed in Qiu and Hwang (2007) was adjusted for the Fay-Herriot model (1979).

The methods are judged based on bias, mean squared error (MSE), asymptotic coverage probability (ACP) of the confidence intervals and the length of the confidence intervals (ALCI).

Simulation results In Table 4.1 we present the mean and empirical standard deviation of estimates of β and τ^2. The numerical results indicate good performance of the EM algorithm based maximum likelihood estimate of the model parameters.

Table 4.1: Estimation of model parameter. The left panel is for β and the right panel is for τ^2

	$\beta = 10$					$\tau^2 = 0.25, 1, 4$			
	$n = 36$, $m = 9$		$n = 180$, $m = 18$			$n = 36$, $m = 9$		$n = 180$, $m = 18$	
τ^2	Mean	SD	Mean	SD	τ^2	Mean	SD	Mean	SD
0.25	10.0071	0.3618	9.9951	0.1853	0.25	0.2558	0.0605	0.2575	0.0097
1	10.0142	0.3311	9.9970	0.1743	1	0.9418	0.3333	1.0426	0.1264
4	10.0282	0.4639	10.0048	0.2254	4	3.5592	1.3316	4.0817	0.5551

SD: standard deviation over 200 replicates

The following Tables 4.2, 4.3 and 4.4 provide the numerical results averaged over areas within the group (having the same sampling variances). We calculated the relative bias, mean squared error, coverage rate and average confidence intervals of the small area estimators based on 200 replications.

In most cases, the bias of the four methods are comparable. In the case of high sampling variance, the method IV outperformed other methods. High sampling variance gives more

Table 4.2: Simulation results for prediction when $\tau^2 = 0.25$

		$n = 36$, $m = 9$				$n = 180$, $m = 18$			
	σ_i^2	I	II	III	IV	I	II	III	IV
Bias	1	0.0048	0.0198	0.0272	0.0018	-0.0051	-0.0086	-0.0112	-0.0111
	4	-0.0033	-0.0061	-0.0145	-0.0158	-0.0130	-0.0109	-0.0065	-0.0116
	16	0.0126	0.0370	0.0369	0.0096	-0.0046	-0.0045	-0.0080	-0.0061
MSE	1	0.3066	0.3955	0.6861	0.3805	0.2258	0.2757	0.4470	0.2922
	4	0.3281	0.5119	1.3778	0.7285	0.2595	0.3010	0.5805	0.3748
	16	0.3715	0.4623	1.6749	1.9316	0.2815	0.2856	0.4856	0.6383
ALCI	1	2.1393	2.5485	4.4906	3.0528	1.9220	1.6006	3.6466	2.4811
	4	2.2632	3.9574	6.8887	5.6842	2.0557	2.1524	5.2472	4.2160
	16	2.3221	4.5619	9.3335	11.1363	2.1046	2.3308	6.5273	7.8492
ACP	1	0.9468	0.8958	0.9771	0.9708	0.9564	0.8160	0.9851	0.9631
	4	0.9468	0.9433	0.9829	0.9917	0.9555	0.8478	0.9967	0.9967
	16	0.9365	0.9375	0.9933	0.9975	0.9529	0.8472	0.9998	0.9999

Table 4.3: Simulation results for prediction when $\tau^2 = 1$

		$n = 36$, $m = 9$				$n = 180$, $m = 18$			
	σ_i^2	I	II	III	IV	I	II	III	IV
Bias	1	-0.0152	0.0205	0.0255	0.0051	-0.0064	-0.0085	-0.0111	-0.0101
	4	-0.0167	-0.0164	-0.0151	-0.0219	-0.0151	-0.0121	-0.0133	-0.0164
	16	-0.0323	0.0508	0.0515	0.0216	-0.0028	-0.0017	-0.0073	-0.0039
MSE	1	0.5645	0.6300	0.7238	0.6260	0.5288	0.5555	0.5673	0.6336
	4	0.8566	1.0746	1.5396	1.0992	0.8159	0.8707	0.9415	0.8948
	16	1.0482	1.2406	2.1059	2.3156	0.9786	1.0043	1.1024	1.1878
ALCI	1	3.4550	3.1822	4.4938	3.2117	3.1088	2.5094	3.6763	2.8676
	4	4.0321	5.8733	6.8984	5.7909	3.7844	4.2908	5.3323	4.5543
	16	4.4082	7.4286	9.3555	11.1555	4.1187	5.1590	6.6785	7.8937
ACP	1	0.9704	0.8800	0.9762	0.9275	0.9660	0.8771	0.9786	0.8879
	4	0.9633	0.9308	0.9812	0.9808	0.9627	0.9464	0.9918	0.9740
	16	0.9533	0.9325	0.9912	0.9938	0.9613	0.9560	0.9974	0.9979

Table 4.4: Simulation results for prediction when $\tau^2 = 4$

		$n = 36$, $m = 9$				$n = 180$, $m = 18$			
	σ_i^2	I	II	III	IV	I	II	III	IV
Bias	1	-0.0024	0.0248	0.0229	0.0180	-0.0084	-0.0098	-0.0122	-0.0106
	4	-0.0343	-0.0310	-0.0210	-0.0340	-0.0110	-0.0092	-0.0174	-0.0132
	16	-0.0147	0.0702	0.0767	0.0467	0.0016	0.0024	-0.0059	0.0012
MSE	1	0.8822	0.8786	0.8579	1.0559	0.8359	0.8334	0.8541	0.8605
	4	2.0577	2.2315	2.1818	2.2422	2.0424	2.0735	2.0935	2.1130
	16	3.4516	3.7401	3.9267	3.8981	3.3153	3.3516	3.3939	3.3631
ALCI	1	4.6318	4.1936	4.5369	3.7677	4.0256	3.5346	3.9626	3.7499
	4	6.2015	10.9093	7.0376	6.4314	5.9000	9.0913	6.2217	6.1540
	16	7.7221	18.0039	9.6718	11.3341	7.4430	14.6665	8.3908	8.7537
ACP	1	0.9791	0.9067	0.9733	0.9029	0.9674	0.9135	0.9600	0.9468
	4	0.9556	0.9850	0.9725	0.9496	0.9592	0.9919	0.9633	0.9573
	16	0.9510	0.9958	0.9796	0.9858	0.9573	0.9990	0.9718	0.9776

weight to the population mean by a construction that makes the estimator closer to the mean at the second level. On the other hand, methods I-III use shrinkage estimators of the sampling variances which would be less than the maximum of all sampling variances. This would tend to have a little more bias. However, due to shrinkage in sampling variance, method I would expect gain in the variance of the estimators meaning the mean squared error (MSE) would tend to smaller. Among the methods I-III, method I performed better than method II and method III, method II and III are much closer to each other. The maximum gain in method I compared to method II is 99%.

In terms of the mean squared error (MSE), method I performed consistently better than the other three methods in most cases, except the case when the ratio of sampling variance to model variance is lowest: $(\sigma_i^2 = 1)/(\tau^2 = 4) = 0.25$. In this case, the variance between small areas (model variance) is much higher than the variance within each small area(sampling variance). When using our method to estimate the mean of each small area, the information

"borrowed" from other areas will misdirect the estimation. The maximum and minimum gain compared to methods II and III are 30%, -9% and 77%, -11% respectively.

ACP: We calculated 95% confidence intervals. Methods I and III do not have any under coverage. This is expected by their optimal interval construction. Method I meets the nominal coverage rate more frequently than any other method. Method II has some under coverage. This could go as low as 82%.

ALCI: Method I produced considerably shorter confidence intervals in general. Method IV produced comparable lengths as in other methods, but the length was considerably higher in case of high sampling variance. The reason is the method IV truncates $M = \tau^2/(\tau^2+\sigma_i^2)$ with a positive number $M_1 = 1 - Q_\alpha/(N-2)$, where Q_α is the α-quantile of a chi-squared distribution with N degrees freedom. When the ratio of sampling variance to model variance, σ_i^2/τ^2, is high, M_1 is much greater than M. For example, in the case of $(\sigma_i^2, \tau^2) = (16, 0.25)$, the average length is 11.13 in method IV whereas this is only 2.78 in method I and 4.56 in method II.

4.7 A Real Data Analysis

We illustrate our methodology with a widely studied example. The data set is from the U.S. Department of Agriculture and was first analyzed by Battese, Harter and Fuller (1988). The data set is about crop and soybeans in 12 Iowa counties. The sample sizes for these areas are small, ranging from 1 to 5. We shall consider corn only to save space. For the proposed model, the sample size of each area requires $n_i > 1$. Therefore the same modified data from You and Chapman (2006) is used, which only includes the areas with sample size 2 or greater. The mean reported crop hectares for corn (x_i) comprise the direct survey estimates.

Table 4.5: Crop data from You and Chapman(2006)

		Corn			
County	n_i	x_i	z_1	z_2	$\sqrt{S_i^2}$
Franklin	3	158.623	318.21	188.06	5.704
Pocahontas	3	102.523	257.17	247.13	43.406
Winnebago	3	112.773	291.77	185.37	30.547
Wright	3	144.297	301.26	221.36	53.999
Webster	4	117.595	262.17	247.09	21.298
Hancock	5	109.382	314.28	198.66	15.661
Kossuth	5	110.252	298.65	204.61	12.112
Hardin	5	120.054	325.99	177.05	36.807

Table 4.6: Estimation results of corn

	I: Proposed method		II: Wang and Fuller(2003)	
County	$\hat{\theta}_i$	Confidence Interval	$\hat{\theta}_i$	Confidence Interval
Franklin	131.8106	104.085, 159.372(55.287)	155.4338	124.151, 193.094(68.943)
Pocahontas	108.7305	80.900, 136.436(55.536)	102.3682	-38.973, 244.019(282.993)
Winnebago	109.0559	81.430, 136.646(55.216)	115.9093	-53.768, 279.314(333.083)
Wright	131.6113	103.736, 159.564(55.828)	131.0674	8.330, 280.263(271.932)
Webster	113.1484	92.805, 133.348(40.543)	109.4795	32.514, 202.675(170.161)
Hancock	129.4279	111.781, 147.193(35.412)	124.1028	56.750, 162.013(105.262)
Kossuth	121.0071	103.451, 138.626(35.175)	116.7147	68.049, 152.454(84.405)
Hardin	130.2520	112.373, 148.114(35.741)	137.7983	51.734, 188.373(136.638)
	III: Hwang et al.(2009)		IV: Qiu and Hwang(2007)	
County	$\hat{\theta}_i$	Confidence Interval	$\hat{\theta}_i$	Confidence Interval
Franklin	158.4677	128.564, 188.370(59.805)	157.7383	146.999, 168.477(21.478)
Pocahontas	100.1276	-44.039, 244.295(288.334)	101.1661	19.444, 182.887(163.442)
Winnebago	114.1473	0.065, 228.228(228.163)	113.7746	56.263, 171.286(115.022)
Wright	140.3717	-24.119, 304.862(328.982)	143.2244	41.559, 244.889(203.330)
Webster	115.7865	50.297, 181.275(130.978)	115.2224	75.124, 155.320(80.196)
Hancock	111.3087	66.213, 156.403(90.189)	113.1766	83.691, 142.661(58.970)
Kossuth	110.9585	74.366, 147.550(73.184)	112.3239	89.520, 135.127(45.607)
Hardin	126.6093	40.040, 213.178(173.137)	123.9049	54.607, 193.202(138.594)

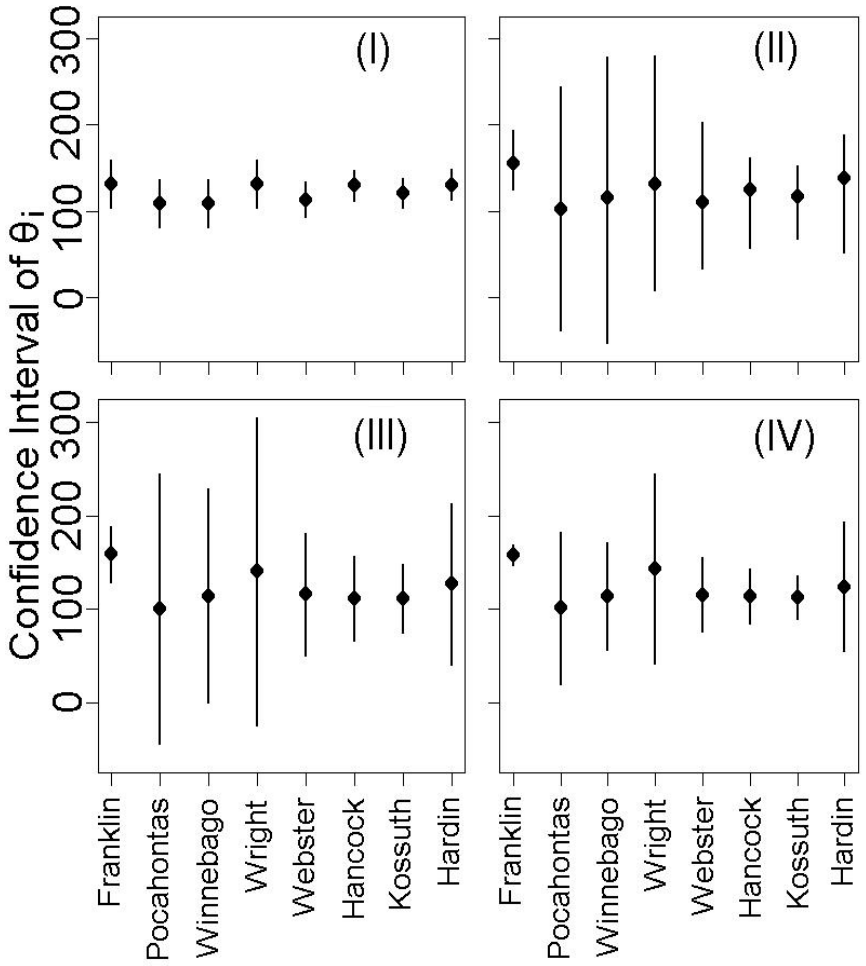

Figure 4.1: Corn hectares estimation. The vertical line for each county displays the confidence interval of $\hat{\theta}_i$, with $\hat{\theta}_i$ marked by the circle, for (I) Proposed method, (II)Wang and Fuller (2003), (III)Hwang *et al.* (2009) and (IV)Qiu and Hwang (2007).

The sample variances are calculated based on the original data assuming a simple random sampling. The sample s.d. varies widely from 5.704 to 53.999 (coefficient of variation varies from 0.036 to 0.423). The means of number of pixels from LANDSAT satellite data (z_i) are the covariates in the estimation procedure. z_1 is the mean of pixels of corn and z_2 is the mean of pixels of soybean. These covariates are used to fit the model (4.1) and (4.2). The detail of the modified data can be found in You and Chapman (2006)and tabulated in the Table 4.5.

The small area estimates and their confidence intervals are summarized in Fig 4.1. The

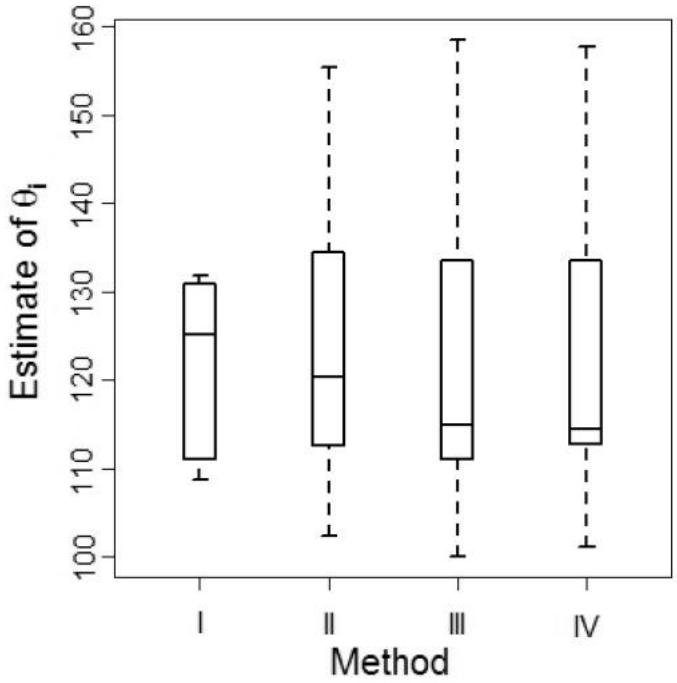

Figure 4.2: Boxplot of estimates of corn hectares for each county. (I) to (IV) are the 4 methods corresponding to Figure 4.1.

numerical figures are also provided in Table 4.6. The point estimates in all 4 methods are comparable. The summary measures, mean, median, and range of the parameter estimates for the methods (I,II,II,IV) are respectively $(121.9, 124.1, 122.2, 122.6)$, $(125.2, 120.4, 115.0, 114.5)$ and $(23.1, 53.0, 58.4, 56.6)$. The distributions are summarized in Fig 4.2. The central locations are similar under all the methods but there is a significant difference in their variability. If the assumption of constant model variance is true and when sample size varies from only from 3 to 5, then under simple random sampling, one would expect relatively low variation among the small area estimates. Method I is superior in this sense.

Further, smoothing sampling variances has strong implications in measuring uncertainty and hence in the interval estimation. The proposed method has the shortest confidence interval on average compared to all other methods. Methods II and III provide intervals with negative lower limit. This seems unrealistic because the direct average of area under corn is

high positive for all the counties. Moreover, the crude confidence intervals ($x_i \pm t_{0.025} S_i$), the widest, do not contain zero for any of the areas. Note that method II does not have any theoretical support on its confidence intervals. Method II and method III produce wider confidence intervals when the sampling variance is high. For example, the sample size for both Franklin county and Pocahontas county is three, but the sampling standard deviations are 5.704 and 43.406. Although the confidence interval under method I is comparable, they are wide apart for methods II and III. This is because, although these methods consider the uncertainty in sampling variance estimates, the smoothing did not use the information from direct survey estimates, resulted in the underlying sampling variance estimates remain highly variable (due to small sample size). In effect, the variance of the variance estimator (of the point estimates) is bigger compared to that in method I. This is further confirmed by the fact that the intuitive standard deviations of the "smoothed" small area estimates (one fourth of the interval) are smaller and less variable under method I compared to others.

In addition to the confidence interval, we also calculated the BIC for the proposed method and method III (Hwang et al.(2009)) as these two methods have the same numbers of parameters and their model structure are very similar to each other. The difference is in the level-2 model for the variance part. The BIC of proposed method is 210.025, method III's is 227.372. The BIC of the proposed method is lower. We could not compare Wang and Fuller (2003) since the explicit likelihood has not been used. The BIC criteria support our data analysis.

Chapter 5

Clustering Based Small Area Estimation: An Application to MEAP Data

5.1 Introduction

As we mentioned in the previous chapters, small area estimation problems exist in many application fields. In this chapter, we examine small area estimation in educational assessment. In particular, we analyze data from the Michigan Educational Assessment Program (MEAP).

For educational accountability purpose, the results of high stakes tests are used as an indicator of school districtperformance. Often times the average test scores across grade are reported for each school district. However, the number of students in each school distric are quite different from each other. Table 5.1 summarizes the number of school districts with a

Table 5.1: Number of public school 4th graders in some school districts in Michigan state

Range of Numbers of Students	Number of School Districts
$1 \sim 100$	210
$100 \sim 200$	124
$200 \sim 300$	62
$300 \sim 500$	39
$500 \sim 1000$	30
$1000 \sim$	11

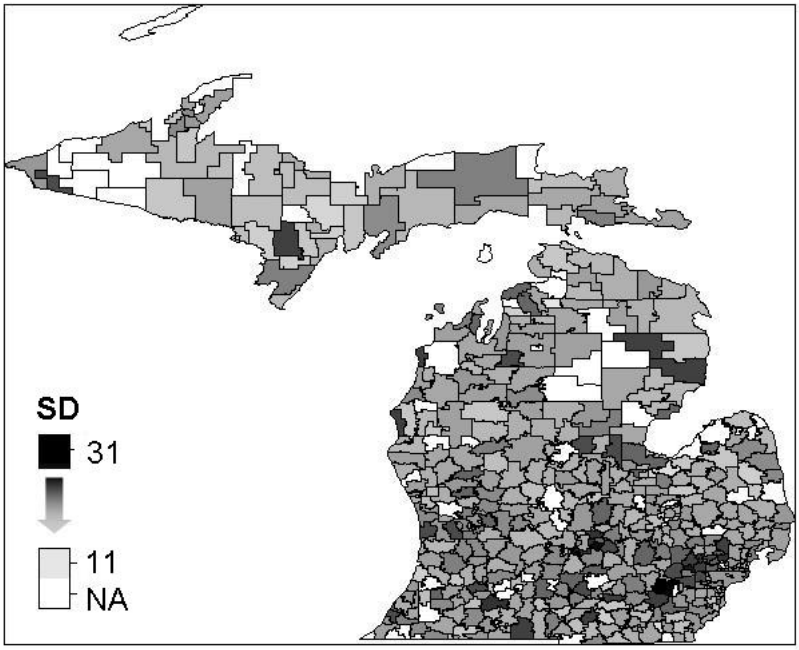

Figure 5.1: Standard deviation of public school students' assessment scores in available school districts in Michigan state.

given size for Grade 4 in Michigan. We can see that a large school district in a large city might have thousands of students in a grade, while a small school district in a rural area might have less than 100 students in a grade. Figure 5.1 gives the standard deviation of the assessment scores of students in each school district. From the map, it can be seen that there is a large amount of variation in the standard deviations of the school districts. Therefore, the direct comparison of average scores between large and small school districts may not be appropriate. A comparison of model based "smoothed" results is more appropriate.

The general approach for this type of analysis is to use the Fay-Herriot model. Let Y_i be the average score reported, X_i be the available covariate information for the regression stage, and β be the regression parameter, $i = 1, \ldots, n$ for the total number of districts n. Then the model is given by (1.1) or (2.5),

$$Y_i = X_i^T \beta + v_i + e_i \tag{5.1}$$

where v_i is the random effect for each school district, which follows an independent and identical distribution, usually a normal distribution; e_i is the sampling error; and the variances of e_i are known.

When we use the model based approach in equation (5.1), we "borrow strength" from all other school districts universally since all school districts sharing the same regression parameter β and the random effects v_i are iid. However, the actual geographical and socioeconomic characteristics of school districts in a large region are quite diverse. Poverty levels for the school districts in Figure 5.1 are given in Figure 5.2. From the map, it can be seen that the school districts in upper Michigan have higher poverty levels. Although the school districts in lower Michigan have lower poverty levels in general, some school districts have higher poverty levels than others.

Therefore, it is more appropriate to divide the school districts into several groups (clusters). In each group, the school districts are similar in some meaningful ways, namely they are from the same cluster. Then the estimation for a school district only borrows strength from similar school districts and avoids the misleading information from other school districts. Since we do not have any restriction on the spatial characteristic of school districts in our study, we only focus on the clustering based small area estimation. However, the

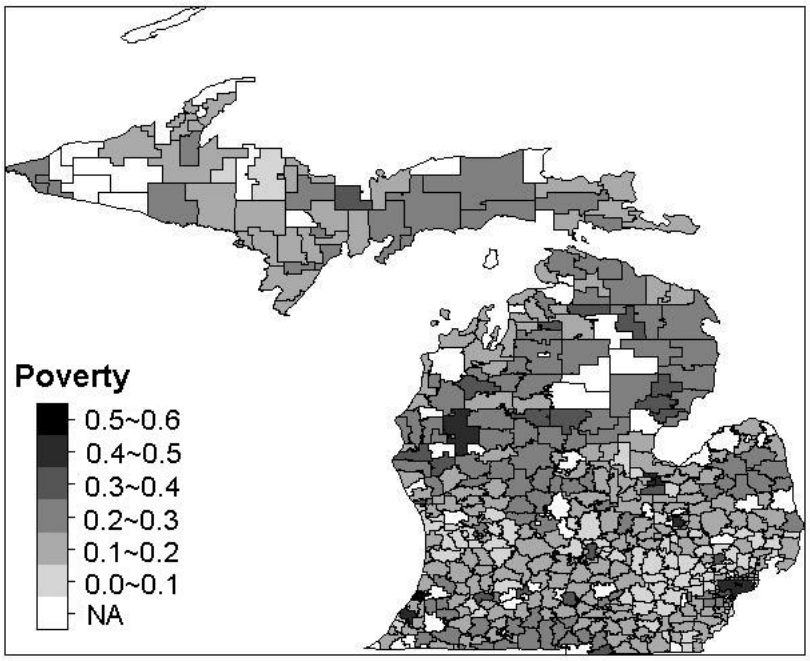

Figure 5.2: Poverty level of available school districts in Michigan state.

partition of clusters was not known. Cluster analysis had to be conducted before we can estimate the small areas.

There are many methods for cluster analysis (clustering). In one widely used class of methods in cluster analysis, two clusters are chosen to be merged based on the optimization of some criterion. Popular criteria include the sum of within-group sums of squares (Ward, 1963) and the shortest distance between groups. Iterative relocation (partitioning) is another common class of methods. In each iteration, data points are moved from one cluster to another depending on whether improvement is achieved with respect to some criterion. K-means clustering (MacQueen, 1967) is a method of iterative relocation with the sum of squares criterion. K-means clustering is used to generate an initial partition in this dissertation.

Clustering algorithms can also be based on probability models. Literature shows that

some of the heuristic methods, such as K-means clustering, can also be treated as approximate estimation methods of certain probability models. In the context of model-based clustering, finite mixture models are often proposed. Each component probability distribution in finite mixture models corresponds to a cluster. It has been shown that finite mixture models can be used to solve the practical questions that arise when applying clustering methods. A review of model-based clustering can be found in Fraley and Raftery (2002).

However, the finite mixture model approach of clustering does not explicitly include the partition as a parameter and involves independent and identically distributed structures. In addition, there is usually no restriction on the mean structure in this class of models. Information from covariates in the mean profile is often necessary in many applications. Therefore, the idea of modeling the mean via regression and keeping the ability to detect clusters at the same time has gained more attention recently. Booth et al. (2008) proposed a new clustering methodology based on a multilevel linear mixed model. A cluster-specific random effect is included in the model, which allows the departure of the cluster means from the assumed base model. An objective function is constructed based on the posterior distribution of the undergoing partition. The partition that maximizes the objective function is chosen as the "optimal" clustering partition. A stochastic search algorithm is also proposed to find such a posterior probability.

A similar model from Booth, Casella and Hobert (2008) is used to describe the relationship between the assessment performance and other covariate information in this study. The actual undergoing partition of districts is used as a parameter in the model. An objective function is defined as the posterior distribution of the partition parameter based on the given data. The posterior distribution is known up to a normalizing constant. The partitions with

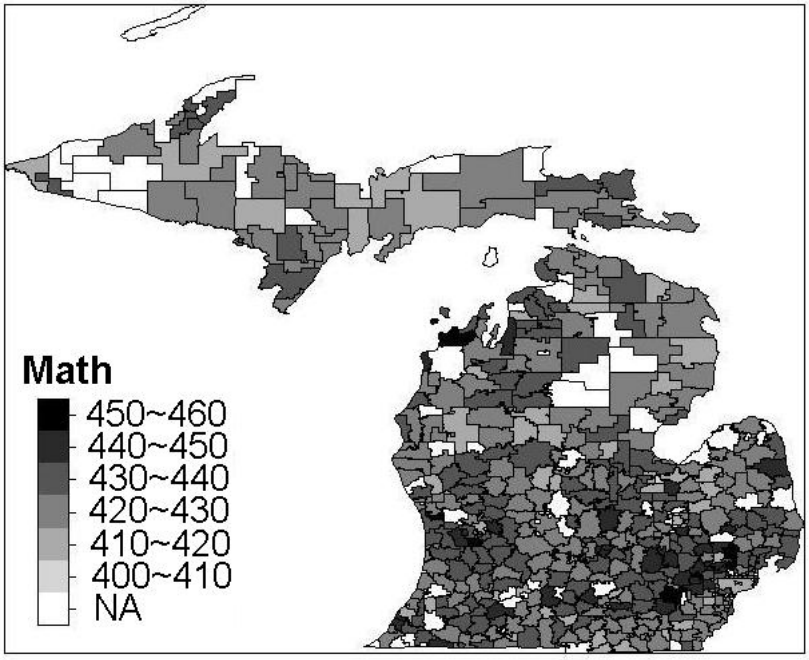

Figure 5.3: Math scores of 4th graders in public schools in available school districts in Michigan state.

the highest posterior probability will be the final result. The stochastic search procedure is such an algorithm constructed by a mixture of two Metropolis-Hastings algorithms: one makes small scale changes to individual objects and another performs large scale moves involving entire clusters. The details are given in the following sections.

The data set is introduced in Section 5.2, and the proposed model is given in Section 5.3. The stochastic search algorithm is described in Section 5.4. In the Section 5.5, the results of the data analysis are presented.

5.2 Data Set

The data used is from the Michigan Educational Assessment Program (MEAP). MEAP is a standardized test that was started by the State Board of Education. The test is taken by all public school students in the state of Michigan from elementary school to middle/junior

high school. The results at the district level are public accessible and can be downloaded at http://www.michigan.gov/mde/. Our study considered the 4th grade math score in Fall 2009. There are about 700 districts in the MEAP's data set. We wanted to analyze the cluster partition of districts with respect to the poverty level in each district, the data reported by Small Area Income and Poverty Estimates (SAIPE) program. SAIPE is under the U.S. Census Bureau and reports model-based estimation results of sample surveys in noncensal years. A more detailed description of SAIPE can be found at its website http://www.census.gov//did/www/saipe/. There are about 550 school districts in SAIPE's data set. Since MEAP and SAIPE use different codes to represent the same school districts, we only had 476 districts left after merging these two data sets by matching the names of school districts, that is, some school districts with data were not included in our merged data set due to non-uniform data coding. No statistical methods were used in this study to manipulate the data sets. In the future, if other information can be accessed so that all the schools districts' data can be matched, there would be no missing data and a more complete analysis could be conducted. There is a lot of information reported in the original data set, but we only kept the average math score, standard deviation, number of students, and poverty level in our data set.

The map in Figure 5.3 shows the distribution of math scores of public school 4th graders in available districts in Michigan state. Comparing the maps in Figures 5.1, 5.2 and 5.3, we can see there are some patterns between the math score and the poverty level, but there is not a simple linear relationship.

5.3 Proposed Model

We used w to represent a partition of the districts, and $c = c(w)$ clusters in this partition, which are denoted by $C_1, \ldots, C_c$. Let (Y_i, S_i^2) be the observed score and the variances for the i-th district, $i = 1, \cdots, n$, where n is the total number of available districts. Let $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^T$ be the set of available covariates and $\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kp})^T$ be the associated regression coefficients if the district i belongs to cluster k. For a fixed partition w, we consider the following hierarchical model:

$$
\begin{aligned}
Y_i &= \boldsymbol{X}_i^T \boldsymbol{\beta}_k + u_k + v_i + \varepsilon_i, \quad i \in C_k \\
u_k &\sim N(0, \lambda\sigma_k^2), \quad \lambda > 0 \\
v_i &\sim N(0, \sigma_k^2) \\
\varepsilon_i &\sim N(0, S_i^2)
\end{aligned}
\tag{5.2}
$$

where $\boldsymbol{X}_i^T \boldsymbol{\beta}_k$ is the linear part of the mean specification for each cluster, u_k are the random effects at the cluster level, v_i are the random effects from the districts within each cluster, and S_i^2, $i = 1, \ldots, n$ are treated as known and reported by MEAP. The data-driven tuning parameter λ can be determined by analysis of variance.

5.3.1 Prior Information

The priors of the model are set as

$$\boldsymbol{\beta}_k \sim uniform\ on\ \mathbb{R}^p \tag{5.3}$$

$$\sigma_k^2 \sim IG(a, b) \tag{5.4}$$

where $\mathbb{R}^p$ is the p-dimensional real space. We choose an improper prior for $\boldsymbol{\beta}_k$, and no specific prior information is given to $\boldsymbol{\beta}_k$. $IG(a, b)$ is the inverse gamma distribution with parameters a and b.

The prior for w is chosen as

$$\pi_n(w) = \frac{\Gamma(m)m^{c(w)}}{\Gamma(n+m)} \prod_{k=1}^{c(w)} \Gamma(n_k) \tag{5.5}$$

where $n_k = \#(C_k)$ and m is a parameter, $m > 0$. This prior distribution was used in Crowley (1997) and Booth et al. (2008). From this prior, more weight is put on the partition with smaller numbers of clusters when m is decreased. If w follows the density function in (5.5), then

$$Pr\{c(w) = k\} = \frac{\Gamma(m)m^k}{\Gamma(n+m)} \sum_{w:c(w)=k} \prod_{j=1}^{k} \Gamma(n_j) \tag{5.6}$$

and

$$E\{c(w)\} = m \sum_{i=0}^{n-1} \frac{1}{m+i} \tag{5.7}$$

If $m \to 0$, $E\{c(w)\} \to 1$. If $m \to \infty$, $E\{c(w)\} \to n$.

5.3.2 Model-Based Objective Functions

Based on the model and prior information described previously, we can determine the model-based objective functions, which mimics the procedure in Booth et al. (2008). For a fixed w, let θ_k denote the parameters of each cluster and $\theta = (\theta_k)_{k=1}^{c(w)}$. Then the joint density

function of $Y = (Y_1, \ldots, Y_n)$ is given by

$$f(Y|\theta, w) = \prod_{k=1}^{c(w)} \int \left[\prod_{i \in C_k} \int f(Y_i|u_k, v_i, \theta_k) f(v_i|\theta_k) dv_i \right] f(u_k|\theta_k) du_k \tag{5.8}$$

If we use $\boldsymbol{Y}_k^*$, an $n_k \times 1$ vector, to represent the Y_is in cluster k, then $\boldsymbol{Y}_k^*$ follows a normal distribution with mean $\boldsymbol{X}_k^* \boldsymbol{\beta}_k$ and variance $\boldsymbol{V}_k$. $\boldsymbol{X}_k^* = \{X_i^T, i \in C_k\}$ and

$$\boldsymbol{V}_k = \begin{bmatrix} (1+\lambda)\sigma_k^2 + S_{i_1}^2 & \lambda\sigma_k^2 & \cdots & \lambda\sigma_k^2 \\ \lambda\sigma_k^2 & (1+\lambda)\sigma_k^2 + S_{i_2}^2 & & \lambda\sigma_k^2 \\ \cdots & & \ddots & \vdots \\ \lambda\sigma_k^2 & \lambda\sigma_k^2 & \cdots & (1+\lambda)\sigma_k^2 + S_{i_{n_k}}^2 \end{bmatrix} \tag{5.9}$$

Then the density function of $\boldsymbol{Y}_k^*$ is

$$f(\boldsymbol{Y}_k^*|\theta, w) = (2\pi)^{-\frac{n_k}{2}} |\boldsymbol{V}_k|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^* \boldsymbol{\beta}_k)^T \boldsymbol{V}_k^{-1} (\boldsymbol{Y}_k^* - \boldsymbol{X}_k^* \boldsymbol{\beta}_k)\} \tag{5.10}$$

and

$$f(Y|\theta, w) = \prod_{k=1}^{c(w)} f(\boldsymbol{Y}_k^*|\theta, w) \tag{5.11}$$

Integrating the product of the joint density function of Y and the prior of $\boldsymbol{\beta}_k, \sigma_k^2$, and multiplying by the prior $\pi_n(w)$ yields the objective function $\pi(w|y)$,

$$\pi(w|y) \propto \pi_n(w) \cdot \prod_{k=1}^{c(w)} \int f(\boldsymbol{Y}_k^*|\theta, w) \pi(\sigma_k^2) \pi(\beta) d\sigma_k^2 d\beta_k \tag{5.12}$$

where $\pi(\cdot)$ represents the prior distributions of σ_k^2 and $\boldsymbol{\beta}_k$. Since we include individual district variance S_i^2 for each district, the objective function does not have a closed form for the final integrating result, and we compute the integral numerically.

5.3.3 Estimation of Mean Scores

For a fixed partition w, the model (5.2) for cluster k, $k = 1, \ldots, c(w)$ can be rewritten in the form of a general linear mixed model as (1.4),

$$\boldsymbol{Y}_k^* = \boldsymbol{X}_k^* \boldsymbol{\beta}_k + \boldsymbol{Z}_k \boldsymbol{v}_k + \boldsymbol{e}_k$$

where $\boldsymbol{Y}_k^* = (Y_i, i \in C_k)^T$, $\boldsymbol{X}_k^* = (X_i^T, i \in C_k)^T$, $\boldsymbol{e}_k = (\varepsilon_i, i \in C_k)^T$, and

$$\boldsymbol{Z}_k = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & & 1 \end{bmatrix}_{n_k \times (n_k+1)} \qquad \boldsymbol{v}_k = \begin{bmatrix} u_k \\ v_{i_1} \\ \vdots \\ v_{i_{n_k}} \end{bmatrix}_{(n_k+1)\times 1} \tag{5.13}$$

Then the variance matrices of $\boldsymbol{v}_k$ and $\boldsymbol{e}_k$ are $\boldsymbol{G}_k = diag(\lambda\sigma_k^2, \sigma_k^2, \ldots, \sigma_k^2)$ and $\boldsymbol{R}_k = diag(S_{i_1}^2, \ldots, S_{i_{n_k}}^2)$. Therefore, the variance matrix is $\boldsymbol{V}_k = \boldsymbol{R}_k + \boldsymbol{Z}_k \boldsymbol{G}_k \boldsymbol{Z}_k^T$.

Using the general theory of Henderson (1975) for a mixed linear model and equation (1.5), the best linear unbiased estimator (BLUE) of $\hat{Y}_i = \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_k + \hat{u}_k + \hat{v}_i$ is given by

$$\hat{Y}_i = \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_k + \boldsymbol{m}_i^T \boldsymbol{G}_k \boldsymbol{Z}_k^T \hat{\boldsymbol{V}}_k^{-1} (\boldsymbol{Y}_k^* - \boldsymbol{X}_k^* \hat{\boldsymbol{\beta}}_k) \tag{5.14}$$

where $\hat{\boldsymbol{\beta}}_k$ is the general least squared estimator,

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k^{*T}\hat{\boldsymbol{V}}_k^{-1}\boldsymbol{X}_k^*)^{-1}(\boldsymbol{X}_k^{*T}\hat{\boldsymbol{V}}_k^{-1}\boldsymbol{Y}_k^*) \tag{5.15}$$

where $\hat{\boldsymbol{V}}_k$ is the variance matrix with estimator of σ_k^2, $\hat{\sigma}_k^2$. Then $\boldsymbol{m}_i = (1, 0, \ldots, 1, \ldots, 0)^T$ is a $(n_k + 1) \times 1$ vector with 1 at the first and $i+1$ positions for ith district and 0 for other positions, and $\hat{\sigma}_k^2$ is obtained by maximizing the log of the marginal likelihood of σ_k^2,

$$\begin{aligned}
\log f(\sigma_k^2|\boldsymbol{Y}_k^*, \boldsymbol{X}_k^*, \theta) &= \log \int f(y_k^*|\theta, w) d\boldsymbol{\beta} \\
&= (1 - \frac{n_k}{2})\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}_k| + \frac{1}{2}\log|(\boldsymbol{X}_k^{*T}\boldsymbol{V}_k^{-1}\boldsymbol{X}_k^*)^{-1}| \\
&\quad -\frac{1}{2}\boldsymbol{Y}_k^{*T}\boldsymbol{V}_k^{-1}\boldsymbol{Y}_k^* \\
&\quad +\frac{1}{2}\boldsymbol{Y}_k^{*T}\boldsymbol{V}_k^{-1}\boldsymbol{X}_k^{*T}(\boldsymbol{X}_k^{*T}\boldsymbol{V}_k^{-1}\boldsymbol{X}_k^*)^{-1}\boldsymbol{X}_k^{*T}\boldsymbol{V}_k^{-1}\boldsymbol{Y}_k^*
\end{aligned} \tag{5.16}$$

The maximization is found by numeric methods.

5.3.4 MSE of Estimator

A MSE estimator from (6.2.37) in Rao (2003) for the BLUE given in the previous section is given as

$$\begin{aligned}
MSE(\hat{Y}_i) &= g_1(\hat{\sigma}_k^2) - \boldsymbol{b}^T_{\hat{\sigma}_k^2} \bigtriangledown g_1(\hat{\sigma}_k^2) + g_2(\hat{\sigma}_k^2) + 2g_3(\hat{\sigma}_k^2) \\
g_1(\hat{\sigma}_k^2) &= \boldsymbol{m}_i^T(\boldsymbol{G}_k - \boldsymbol{G}_k\boldsymbol{Z}^T\hat{\boldsymbol{V}}_K^{-1}\boldsymbol{Z}\boldsymbol{G}_k)\boldsymbol{m}_i^T \\
g_2(\hat{\sigma}_k^2) &= d^T(\boldsymbol{X}_k^{*T}\hat{\boldsymbol{V}}_k^{-1}\boldsymbol{X}_k^*)^{-1}d \\
g_3(\hat{\sigma}_k^2) &= tr\left[\left(\frac{\partial b^T}{\partial \sigma_k^2}\right)\hat{\boldsymbol{V}}\left(\frac{\partial b^T}{\partial \sigma_k^2}\right)^T \bar{\boldsymbol{V}}(\hat{\sigma}_k^2)\right]
\end{aligned} \tag{5.17}$$

where $d^T = \boldsymbol{x}_i^T - \boldsymbol{m}_i^T \boldsymbol{G}_k \boldsymbol{Z}^T \hat{\boldsymbol{V}}_K^{-1} \boldsymbol{X}_k^{*T}$, and $\boldsymbol{m}_i^T$, $\boldsymbol{G}_k$, $\boldsymbol{Z}^T$, $\hat{\boldsymbol{V}}_K$ and $\boldsymbol{X}_k^*$ are given in the previous section. $\boldsymbol{b}^T_{\hat{\sigma}^2_k} \bigtriangledown g_1(\hat{\sigma}_k^2)$ is the bias correction of $g_1(\hat{\sigma}_k^2)$ and calculated by bootstrap.

5.4 Stochastic Search

The estimator of mean score and the corresponding MSE can be calculated if the cluster partition is given, but the partition of the school districts is not known. The optimal cluster partition can be found by maximizing the objective function through the following stochastic search procedure.

Similar to Booth et al. (2008), the stochastic search procedure included two different behaviors: a small scale change as a biased random walk of one district and a large scale change as a split-merge of clusters. The two different scale moves are combined together with different probabilities. Let w' be the current partition and c is the number of clusters. At each iteration, if $c = 1$, a split move is proposed automatically. If $c > 1$, the biased random walk is proposed with probability p_b, and the split-merge move is proposed with probability $1 - p_b$.

5.4.1 Biased Random Walk

As with the nearest neighbor random walk, one district is moved each time, and the move $w' \to w$ has a positive probability if and only if w' and w share the same partition but only have one district in different clusters.

In Booth et al. (2008), the algorithm was applied to genetic data. A cluster may contain a single gene only because the observed data are replicated measurements of genes. However, we only have one piece of observed data for each school district. If the cluster contains one

school district only, the model (5.2) can not be estimated or determined for that cluster. Therefore, each cluster must contain at least two school districts. The detailed moving rules are described as follows. First, one district is chosen uniformly and randomly from the districts inside the clusters with $n_k >= 2$. Then it will move to one of the other $c-1$ clusters with probability $1/(c-1)$. The acceptance probability of $w' \to w$ is $min\{1, \pi(w)/\pi(w')\}$.

5.4.2 Split-Merge Moves

Splitting or merging of clusters are large scale moves. In the biased random walk, only one district is moved at a time. A large scale change would require a lot of steps of the biased random walk and is very unlikely to occur. Therefore, it is necessary to add the large scale change to the algorithm.

At each iteration, a merge move or split move will be proposed randomly. A merge move is proposed with probability $p_m \in (0, 1)$. Two clusters are chosen uniformly at random from the current partition and form a new cluster together. A split move is proposed with probability $1 - p_m$. Since each cluster includes by at least two school districts, one cluster with at least four school districts in the current partition is chosen and then split into two clusters with each cluster containing at least two school districts.

Suppose the current partition is w', and w is the partition after merging two clusters in w'. Then

$$\begin{aligned} P(w' \to w) &= \frac{p_m}{\frac{c(w')(c(w')-1)}{2}} \\ P(w \to w') &= \frac{1-p_m}{(2^{n^*-1}-n^*-1)\sum_{k=1}^{c(w)} I[\#\{C_k(w)\} \geq 4]} \end{aligned} \tag{5.18}$$

where n^* is the number of districts of the cluster in w which is formed by merging two clusters in w'. The acceptance probability of move $w' \to w$ is $min\{1, R\}$ where

$$R = \frac{\pi(w)P(w \to w')}{\pi(w')P(w' \to w)} \tag{5.19}$$

and the acceptance probability of move $w \to w'$ is $min\{1, 1/R\}$.

5.5 Data Analysis

In the analysis of our data set, we chose settings similar to those of Booth et al. (2008) for the proposed probabilities of different moves: $p_b = 0.9$ and $p_m = 0.5$. For the prior, we chose $a = 3$, $b = 40$ and kept the range of cluster-level variation to a smaller range to the variance between school districts in the model (5.1).

First, we applied the K-means cluster algorithm in R to the data without any covariates. We chose the initial number of clusters as 10 and used the numbers $1, \ldots, 10$ to denote the clusters. Then we started the stochastic search procedure with this initial partition. In each iteration, the biased random walk or split-merge moves is proposed and accepted with respect to the acceptance probability. The numbers $1, \ldots, c(w)$ were used to denote the clusters in the new partition. The total number of iterations was 10^5.

Figure 5.4 gives the trace plot of the number of clusters vs. iterations. Four lines are drawn with respect to different m values: $0.5, 0.1, 0.0001, 0.00001$. From Figure 5.4, we can see that the final number of clusters decreases as the value of m decreases. This follows the property of the prior of the partition parameter and resembles the results of Booth et al. (2008). We chose the m value as 0.00001 so that the expected number of clusters would be

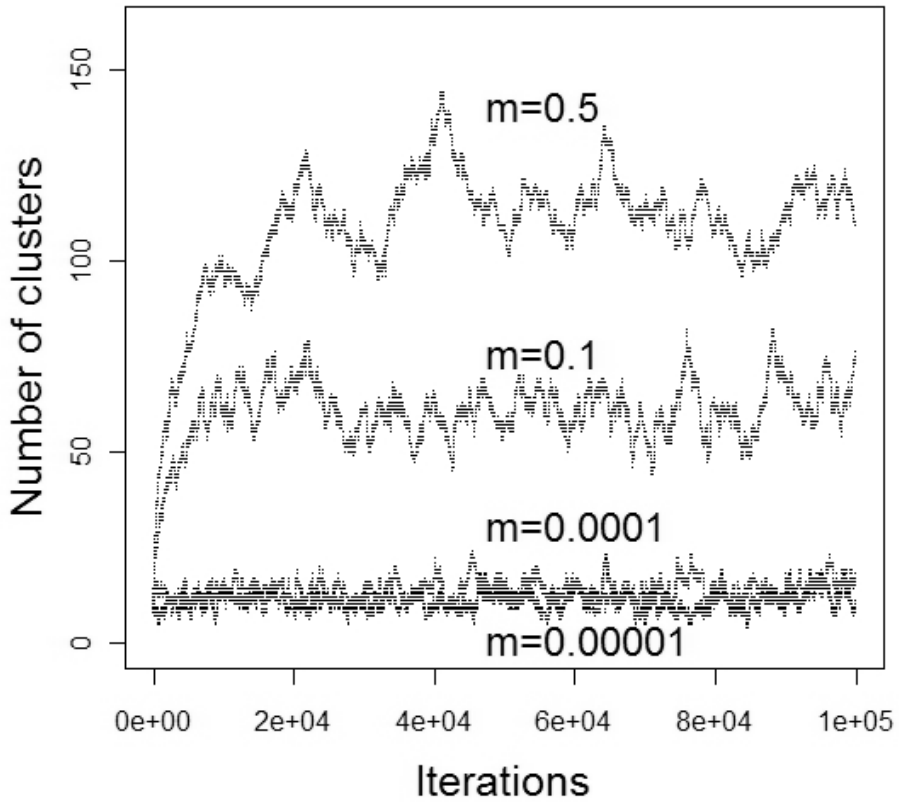

Figure 5.4: Trace plot of different m values.

close to 1. We ran the whole procedure again with double the number of iterations $2 \cdot 10^5$ and used the first 10^5 iterations as a burn-in period. The cluster index of each district was recorded.

The average number of clusters in the last 10^5 iterations was 9.4. When the stochastic search procedure converged, all the clustering results had closed posterior probability. One clustering result with 7 clusters is shown in Figure 5.5. Scatter plots of mean scores vs. poverty levels of the school districts in the 7 clusters are given in Figure 5.6. The reason for choosing this partition is that all the sizes of the clusters are greater than 10 in this clustering result, which is more similar to the real world since there are a total of 476 school

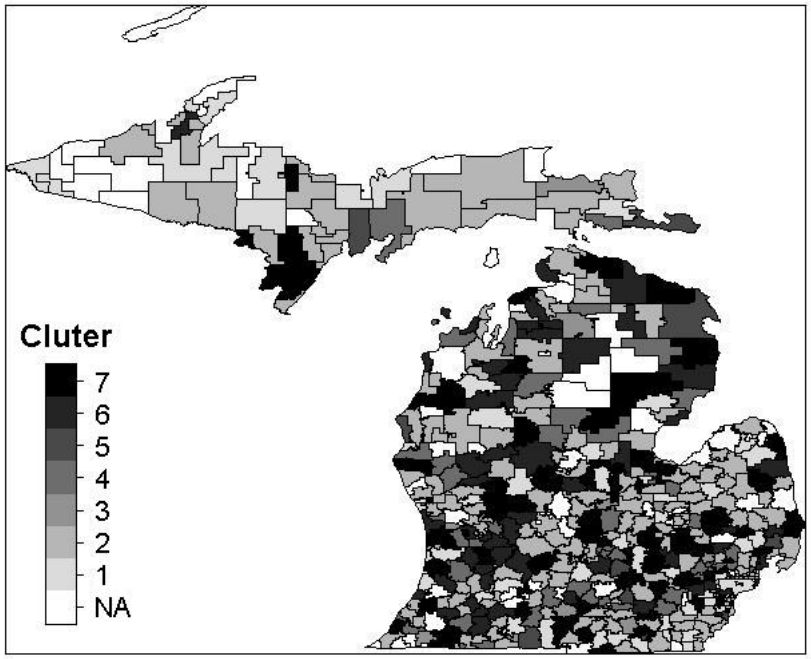

Figure 5.5: Map of clusters.

districts.

From the map of clusters in Figure 5.5, we can see that all the available school districts are divided into 7 clusters, and each cluster is formed from several geographically separated small groups. However, most districts in each small group are neighbors of each other. This agrees with our intuitive image of school districts in that adjacent school districts with similar poverty levels may perform similarly on the state assessment, but all the school districts in a county/city maybe not be placed in one performance level category.

The scatter plots in Figure 5.6 indicate that the relationship between the mean scores of school districts and their poverty levels are close to linearity. However, the R^2 is only 0.3 if we fit all the data with a simple linear regression. The R^2s are $(0.26, 0.37, 0.26, 0.89, 0.58, 0.42, 0.74)$ if we fit the data within each cluster with simple linear regression. But if we fit the data with model (5.2), the R^2s are greater than 0.9 for all the clusters. This indicates the neces-

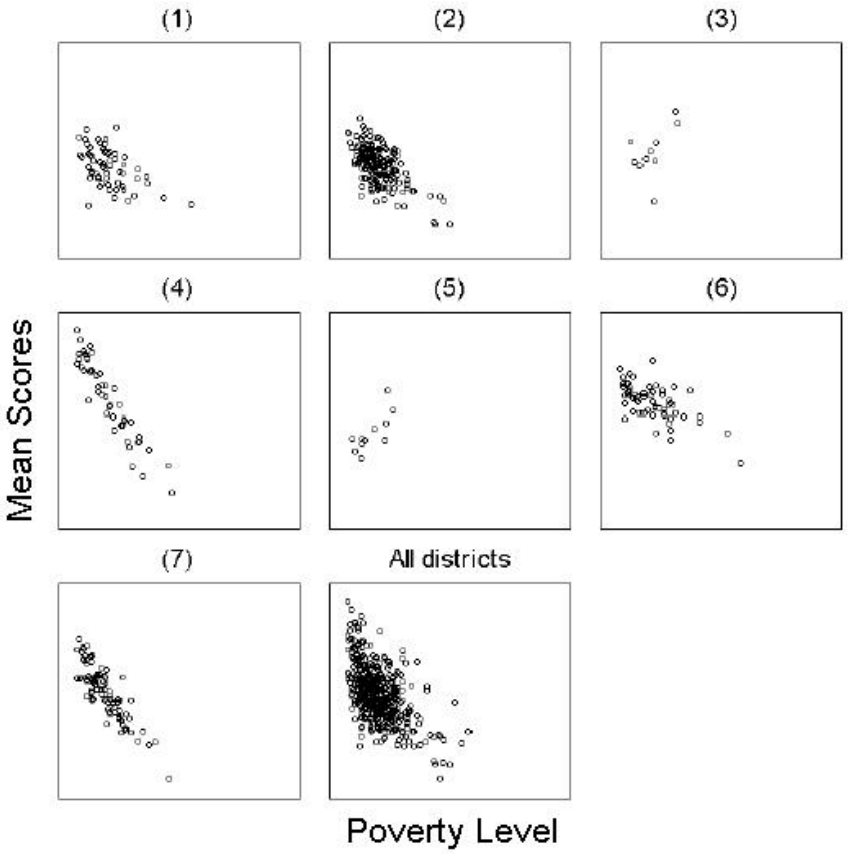

Figure 5.6: Mean scores vs. poverty levels for school districts in the 7 clusters and all the districts.

sity of fitting with a more complex model that accounts for clustering. The mixed model is appropriate since it allows the departure of mean structures for each school district. We also analyzed the data set with model (5.1). The BIC of model (5.1) is 3197, and model (5.2) is 2857. The BIC of the model with clustering is lower.

Based on model (5.1) and (5.2), the estimates of the mean scores can be obtained. The results show that 90% of the model based (with clustering) estimates are within 2 points of the observed scores. However, there are 10 school districts that have more than 4-point differences. The mean scores for all school districts range from 403.6 to 456.8 and 4 points is 7.5 % of this score range. Therefore, it is necessary to make adjustments if we want to

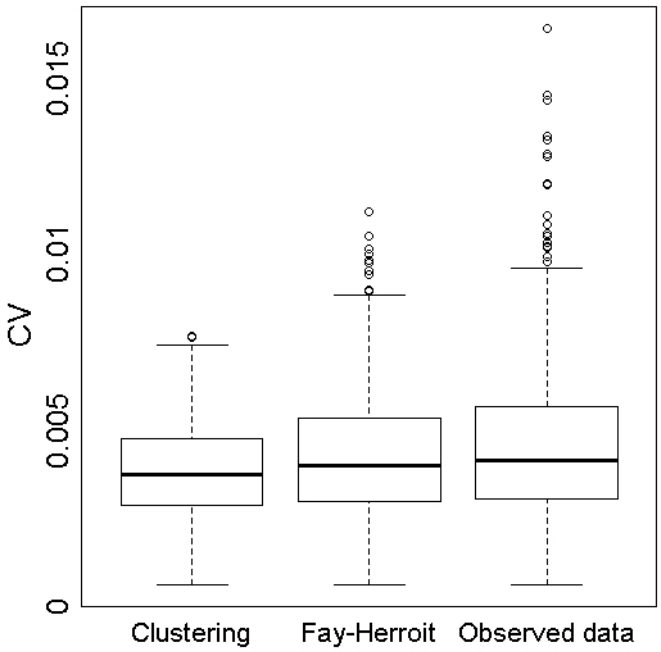

Figure 5.7: Coefficient of variation for the model-based estimates and observed mean scores.

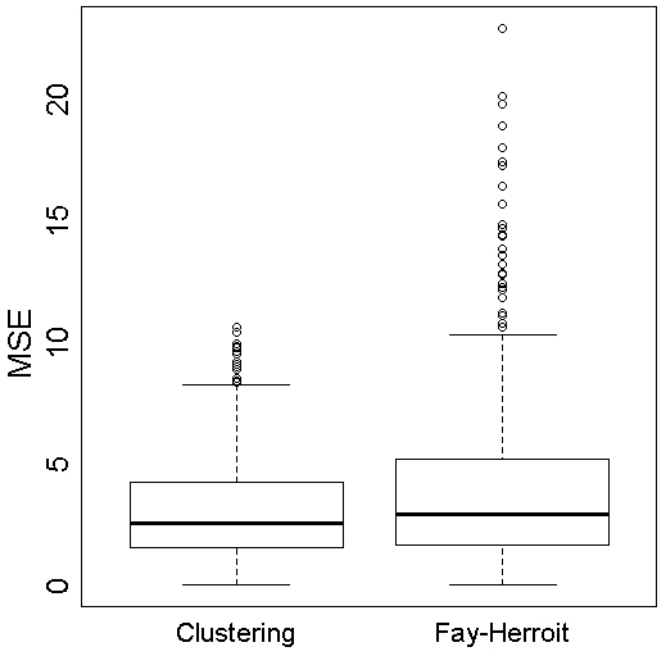

Figure 5.8: MSE of estimators for models with clustering (5.2) or no clustering (5.1).

compare the mean scores of these school districts directly, and we need to include more information to make the comparison more comprehensive.

Boxplots of coefficient of variation for the proposed model, the Fay-Herriot model, and observed data are given in Figure 5.7. The MSE's boxplots for estimators based on model (5.1) and (5.2) are given in Figure 5.8. From the comparisons of coefficient of variation and MSE, the model with clustering is more stable and can provide more accurate estimates.

From the results, we can see that cluster based small area estimation works very well for our study purpose and reaches a reasonable and stable clustering partition of school districts. In this study, the common prior was used and no specific background information was emphasized. Only the poverty level was included as covariate. If more information is available, then meaningful inferences for policymaking can be obtained easily using small area estimation methods similar to those used in the study. Additional information that could be used includes meaningful prior information such as spacial characteristic about school districts, or socioeconomic inferences such as population percentages of speficic races.

Chapter 6

Discussion

In this dissertation, we studied several topics of small area estimation. Small area estimation research has become popular in recent years because of increasing demand of small area statistics from both public and private sectors. Small area estimation uses small area models to link the direct estimators and covariates explicitly.

It is important to note that the quality of model based estimator for small area depends on the availability of good covariates. That is the reason we conducted the study of model selection in Chapter 2 of this dissertation. Bootstrap method was adopted since it has several advantages compared to other competitive methods. As indicated in (Shao, 1996), a modification with smaller bootstrap sample size is needed to result in a consistent model selection. This modified bootstrap selection procedure was extended to linear mixed effects models, which is the backbone of small area models. In this dissertation, we have established that the modified bootstrap selection procedure worked very well for linear mixed effects models.

For small area models, the typical modeling strategy in area level model is to assume

the sampling variances as known. However, the assumption of known and fixed sampling variances does not take into account the uncertainty of estimation into the overall small area estimation strategy. In Chapter 3, a new approach of small area estimation is introduced, which is a dual "shrinkage" estimation for both small area means and variances in a unified framework. Conditional mean squared error of prediction (CMSEP) was used to evaluate the performance of the proposed estimator. Since the model parameters are not known in practice, and they have to be estimated, the estimator of CMSEP was derived which is second order correct.

It is well known that the standard practice of confidence interval does not produce accurate intervals in case of small area estimation. Therefore, a confidence interval from a decision theory perspective was derived in Chapter 4. Both the theoretical justification and the simulation study were presented. In conclusion, the proposed estimator of small area parameters and corresponding estimator of CMSEP and confidence intervals outperform other methods from literatures under most cases.

In the Chapter 5, we applied small area estimation to 4th grade math score from Michigan Educational Assessment Program (MEAP). We wanted to conduct comparisons between different school districts. Because of the diverse standard deviations and varying sample sizes, school districts were identified as small areas in this study. When we used small area models to "borrow strength" from other school districts, we first conducted a cluster analysis to decide which school districts we should "borrow strength" from. The final estimate results were calculated based on the "optimal" partition of clusters. And the results outperformed those of the Fay-Herriot model.

All the computing in this dissertation was carried out in R and FORTRAN. The structure

of the programming was under R and the heavily computing part was executed by calling FORTRAN from R. The data analyzed in Chapter 4 and 5 are open access to public. The code for simulation illustrations and data analysis is available from the author upon request.

Issues Needed to be addresses for Practical Applications

Small area estimation can play an very important role in many applications, such as public health and education. In practice, there are many issues needed to pay attention to when we apply small area estimation to the real applications. The first issue is about the availability of covariates. It may happen that the good covariates are only available for some small areas, not for others. For example, in the MEAP study, Chapter 5, the data set contained more than 700 school districts, but the data set from Small Area Income & Poverty Estimates (SAIPE) program only contained around 550 school districts and used a different code system to represent the same school districts. After combined the two data sets by the names of school districts, only 476 school districts left. In our study, we did not adapt any imputation method to recover the missing data. But in practice, it is probably necessary to do so. There are many methods available for imputation of missing data, such as nearest-neighbor imputation, hot deck imputation, regression imputation, or using experts' judgments etc. Sometimes another model can also be fitted for missing values. For more detailed description of missing data and small area estimation, one can found in Longford (2005).

Another issue could be the estimates of sampling variances. Our study showed the advantages of modeling small area means and variances simultaneously. However in practice, one might still choose to model only the basic area level model, which is a simpler model and assume the sampling variances as known. In that case, the estimation methods of sampling

variance are very important for the accuracy of final estimation. The sampling variances are usually estimated quantities and those are subject to substantial errors due to the fact that they are often based on equivalent sample sizes as the direct estimates are being calculated. There are many literatures on different estimate methods of sampling variance. The detail variance estimation can be found in Wolter (1985).

There might be other issues in the application of small area estimation. However, by using a explicit model to link the direct estimators and covariants, diagnostics for Small area models is possible. Important results from other fields can also be applied to small area models directly. Therefore, small area estimation can be very useful to provide more stable and accurate estimation results in many applications. Extension of small area estimation to other related areas is of further interest.

BIBLIOGRAPHY

Bibliography

Adkins, L. and Hill, R. (1990). An improved confidence ellipsoid for the linear regression models. *Journal of Statistical computation and Simulations*, 36:9–18.

Battese, G., Harter, R., and Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83:28–36.

Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. *Tech. Report U.S. Census Bureau*.

Booth, J., Casella, G., and J.P., H. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, B*, 70:119–139.

Booth, J. and Hobert, J. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93:262–272.

Brackstone, G. (1987). Small area data: policy issues and technical challenges. *In Small Area Statistics R. Platek, J.N.K. Rao, C.E.Sarndal and M.P.Singh eds.*, Wiley, New York:3–20.

Casella, G. and Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1:159–173.

Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of eblup and related prediction intervals in linear mixed models. *Annals of Statistics*, 36:1221–1245.

Cho, M., Eltinge, J., Gershunskaya, J., and Huff, L. (2002). Evaluation of generalized variance function estimators for the u.s. current employment survey. In *Proc. Amer. Statist. Assoc. Surv. Res. Meth. Sec.*, pages 534–539.

Cox, D. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30:248–275.

Crowley, E. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, 92:192–198.

Das, K., Jiang, J., and Rao, J. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32:818–840.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Fay, R. and Herriot, R. (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.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Gershunskaya, J. and Lahiri, P. (2005). Variance estimation for domains in the u.s. current employment statistics program. *Proc. Amer. Statist. Assoc. Surv. Res. Meth. Sec.*, pages 3044–3051.

Ghosh, M. and Rao, J. (1994). Small arear estimation: An appraisal. *Statistical Science*, 9:54–76.

Gonzalez, M. (1973). Use and evaluation of synthetic estimates. In *Proceedings of the Social Statistics Section*, pages 33–36. American Statistical Association.

Gunst, G. and Mason, R. (1980). *Regression Analysis and Its Applications.* New York: Marcel Dekker.

Hall, P. (1989). Unusual properties of bootstrap confidence intervals in regression problem. *Probability Theory and Related Fields*, 81:247–273.

Hall, P. and Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression moels. *Annals of Statistics*, 34:1733–1750.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Herrador, M., Morales, D., Esteban, M. D., Snchez, n., Santamara Arana, L., Marhuenda, Y., and Prez, A. (2008). Sampling design variance estimation of small area estimators in the spanish labour force survey. *Statistics and Operations Research Transactions*, 32:177–198.

Huff, L., Eltinge, J., and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the u.s. current employment survey. *Proc. Amer. Statist. Assoc. Surv. Res. Meth. Sec.*, pages 1519–1524.

Hwang, J., Qiu, J., and Zhao, Z. (2009). Empirical bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society, B*, 71:265–285.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley:361–379.

Jiang, J., Lahiri, P., and Wan, S. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, 30:1782–1810.

Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40:1042–1067.

Lohr, S. *Sampling: Design and Analysis*. Duxbury, Pacific Grove, CA.

lohr, S. and Rao, J. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, 96:457–468.

Longford, N. T. (2005). *Missing Data and Small-Area Estimation*. Modern Analytical Equipment for the Survey Statistician Series. Springer, 1st edition.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, eds. L.M.L.Cam and J.Neyman, Berkley, CA: University of California Press, 281-297.

Maples, J., Bell, W., and Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the american community survey. *Proceeding of the Survey Research Methods Section, ASA*, pages 5056–5067.

Marker, D. (1999). Organization of small area estimatiors using a generalized linear regression framework. *Journal of Official Statistics*, 15:1–24.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Liner Models, 2nd ed.* London: Chapman and Hall.

Otto, M. and Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *American Statistical Association, Proceedings of the Section on Government Statistics*, pages 160–165.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, 70:125–143.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85:163–171.

Qiu, J. and Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63:767–776.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey methodology*, 25:175–186.

Rao, J. (2001). Small area estimation with applications to agriculture. *in Proceedings of the Second Conference on Agricultural and Environmental Statistical Applications*, ISTAT,Rome,Italy.

Rao, J. (2003a). *Small Area Estimation*. Wiley, New York.

Rao, J. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2:145–169.

Rao, J. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4):811–822.

Rivest, L.-P. and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proc. Internat. Conf. Recent adv. in surv. samp.*

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods (second edition).*

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, 91:655–665.

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82:499–508.

Wang, J. and Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98:716–723.

Ward, J. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58:234–244.

Wolter, K. (1985). *Introduction to Variance Estimation.* Statistics for Social and Behavioral Sciences. Springer, New York, second edition.

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32:97–103.