A BAYESIAN APPROACH TO JOINT SMALL AREA ESTIMATION

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yanping Qu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Glen D. Meeden, Adviser

June, 2012

Acknowledgments

My deepest gratitude is to my advisor, Professor Glen Meeden. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Professor Meeden taught me how to question thoughts and express ideas. His profound understanding of numerous areas has been a great resource for me and has made my Ph.D. research such a pleasant journey. I am also indebted to Professor Meeden for serving as a reference for my internship applications. He is not only an academic advisor, but also a mentor, a friend and inspiration. Professor Glen Meeden, being passionate, energetic and intellectual, has set up a solid professional model that I will be pursuing.

I am thankful to Professor Charlie Gayer for his serving as my defense committee chair and reviewing my thesis. My thanks also go to Professor Chatterjee and Professor Reilly for their time and effort for reviewing my thesis, and for their valuable suggestion regarding my research. My thanks go to Professor Weisberg for supervising me through my consulting work, which was a tremendous experience for me.

I wish to extend my thanks to Professor Tiefeng Jiang, Galin Jones, Peihua Qiu, Xiaotong Shen, and all other professors who give me encouragement and important guidance during my Ph.D. studies.

I am also grateful to my best friend Shanshan Ding for helping me all these years. Your friendship is a gift I will cherish forever.

Finally, I would like to thank my parents and my husband for their endless love and

comforting family support. I dedicate this thesis to my family.

Abstract

In small area estimation problems focus has been put on how to borrow strength across areas in order to develop a reliable estimator when auxiliary information is in hand. Some traditional methods for small area problems borrow strength through linear models that provide links to related areas, which may not be appropriate for some survey data. We propose a new approach to small area estimation, which borrows strength through a noninformative Bayesian prior without any assumption of linearity between variables. This approach results in a generalized constrained Dirichlet posterior estimator when auxiliary information is available for small areas. It is not only able to utilize the auxiliary information within small areas but also able to utilize the auxiliary information across small areas, which is usually impossible to take into account by traditional methods. When information about auxiliary variables is present, the proposed approach allows either estimates for a given area or, simultaneously, for several areas depending on the form of auxiliary information. The Bayes like character of the posterior allows one to prove the admissibility of the point estimator of interest suggesting that inferential procedures based on our approach will tend to have good frequentist properties. The form of our prior distribution allows us to assign a weight to each member of the sample and these weights allow us to find interval estimates for the small area means. This makes our methods easy to use in practice. Simulation studies and an application to a real study are given in this thesis to examine the performance of various approaches.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sample surveys have long been recognized as cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. They are widely used in providing estimates not only for the entire population of interest but also for a variety of subpopulations (domains or small areas). Domains may be defined by geographic areas or socio-demographic groups or other subpopulations. Examples of a geographic domain (area) include a state/province, county, unemployment insurance region and health service area. Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. It is now widely recognized that direct survey estimates for small areas are likely to yield unacceptably large standard errors due to the smallness of sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas.

In chapter 2, we give a literature review of the existing approaches to small area estimation. A common characteristic of these approaches is that the small area means are usually assumed to be related through some type of linear model. Drawing on linear model theory one can derive estimators, which borrows strength by using data from related areas to estimate the mean of interest. Some of these linear models rely on the

direct estimator, thus the corresponding estimators will become problematic when the director estimator is either not available or not reliable. An example is discussed to show the influence of a direct estimator on an indirect estimator in chapter 2. Finding a good estimate of the precision of these indirect estimators and model-based estimators is often difficult due to the small sample size in the domain and the complexity of the procedure, which make them less appealing to researchers. In practice, the assumption of an explicit linking model between variables may not be appropriate for some complicated situations.

In chapter 3, we introduce a novel approach to small area estimation, which is able to borrow strength from related small areas in a Bayesian way. The advantage of our approach is that it needs no assumption of explicit models but it can still make use of both unit-level and area-level auxiliary information. The noninformative prior distribution used in our method leads to an objective posterior distribution, which is appropriate when we believe the observed units are roughly exchangeable with the unobserved units. This will allow us to prove the admissibility of the resulting estimators. This suggests that inferential procedures based on our approach will tend to have good frequentist properties. When information about auxiliary variables is present, the proposed approach allows either estimates for a given area or, simultaneously, for several areas depending on the form of auxiliary information. In addition, our method is able to utilize across-area auxiliary information to help to improve the precision of point estimates, which cannot usually be taken into account by the standard methods. The form of our prior distribution will allow us to assign a weight to each member of the sample. We will incorporate these weights into a Dirchlet distribution which will allow us to find interval estimates for the small area means. This should also make our methods easy to use in practice.

In Chapter 4, we illustrate how to utilize the Metroplis-Hastings algorithm to obtain solutions to integration problems associated with our Bayesian analysis. Given the existence of auxiliary information, we need to construct high-dimensional Markov chains

on a convex polytope that has empty interiors, and then use Markov chain Monte Carlo(MCMC) techniques to compute approximately the desired posterior expectations.

The performance of our method is demonstrated through simulations in chapter 5. Diverse situations are explored to compare our method with the alternatives. The influence of hyperparameter in our model is also studied and a suggestion about selecting the hyperparameter is given. An application on real data from the Behavioral Risk Factor Surveillance System (BRFSS) is explored in Chapter 6 to examine the performance of several methods. Chapter 7 completes the thesis by summarizing the main findings of the thesis.

Chapter 2

Small Area Estimation

2.1 The Basic Idea

The term "small area" or "local area" is commonly used to denote a small geographical area, such as a county, a municipality or a census division. They may also describe a "small domain", i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area. An example of "other domains" is the set of business firms belonging to a census division by industry group. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide specific accuracy at much higher level of aggregation than that of small areas. Thus, the use of survey data in developing reliable small area statistics, possibly in conjunction with the census and administrative data, has received more and more attention.

In the context of sample surveys, we refer to a domain estimator as "direct" if it is based only on the domain-specific sample data. A direct estimator may also use known auxiliary information, such as the total of an auxiliary variable, x related to the variable of interest, y. It includes standard weighted survey estimators, and the associated inferences are based on the probability distribution induced by the sampling design

with the population values held fixed. (A more extensive treatment of direct estimation can be found in the textbook of Lohr and Rao(1999)). Direct estimators typically have good design properties: they are unbiased and can provide valid confidence intervals without any statistical model. "Model assisted" direct estimators that make use of "working" models are also design based, aiming at making the inference "robust" to possible model misspecification. However, in some applications, direct estimators are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area. In fact, no sample units may be selected from some small domains. In this thesis, we use the term "small area" to denote any domain for which direct estimates of adequate precision cannot be produced.

In making estimates for small areas with an adequate level of precision, it is often necessary to use "indirect" estimators that "borrow strength" by using values of the variable of interest, y, from related areas and thus increase the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to the related areas through the use of supplementary information related to y, such as administrative data or data from the last census. The essence of these models is the use of auxiliary data. These data are used to construct predictor variables for use in a statistical model that can be used to predict the estimate of interest for all small areas. The effectiveness of small area estimation depends initially on the availability of good predictor variables that are uniformly measured over the total area. It next depends on the choice of a good prediction model. One key distinction in model construction is between situations where the auxiliary data are available for the individual units in the population and those where they are available only at the aggregate level for each small area. In the former case the data can be used in unit level models, whereas in the latter they can be used only in area level models. Most of the existing models are not able to make use of both unit level auxiliary information and area level auxiliary information simultaneously. In addition, a linearity assumption is often made by these models, which is not always true in practice. In this chapter,

we review the standard techniques in the small area estimation and then discussed the advantages and disadvantages of them.

2.2 Traditional Indirect Estimators

Traditional indirect domain estimators are based on implicit models that provide a link to related areas. Such estimators include synthetic estimators, composite estimators and James-Stein estimators. Gonzalez(1973) proposed a synthetic estimator by assuming a linear model for the data so that the values of the areas that have not been sampled are estimated from the model using only information for available covariates. For the mean, the synthetic estimator is based on the following model

$$\bar{Y}_j = \beta \bar{X}_j + u_j$$

where $\bar{Y}_j$ is the population mean of variable of interest for jth small area and $j \in \{1, \cdots, M\}$, $\bar{X}_j = (\bar{X}_{j1}, \cdots, \bar{X}_{jm})$ is the corresponding mean vector of m auxiliary variables and u_j is an area-based random error, which is normally distributed with zero mean and variance σ_u^2.

The synthetic estimator can be obtained by using the estimate of β from linear regression of the individual level sample data and computing

$$\hat{\bar{Y}}_{j,STH} = \hat{\beta} \bar{X}_j \tag{2.1}$$

as the synthetic estimate in area j. Note that this estimator doesn't make use of any random effect, say u_i, and for this reason it may lead to biased estimates of the area means.

A natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is a composite estimator. The composite estimator is a weighted average of a direct estimator and an indirect estimator, such as the synthetic

estimator. The weights are defined so that if the sample size is "large" the direct estimate is given more weights than the synthetic one and when the sample is not reliable, the synthetic estimator will be given more weights. The equation of the estimator maybe written as

$$\hat{\bar{Y}}_{j,COMP} = \hat{\gamma}_j \hat{\bar{Y}}_{j,DRCT} + (1 - \hat{\gamma}_j)\hat{\bar{Y}}_{j,STH} \tag{2.2}$$

where $\hat{\gamma}_j$ is between 0 and 1 which controls the shrinkage of the two estimators. $\hat{\gamma}_j$ is chosen so that it minimizes the mean squared error(MSE) of (2.2) or the average MSE of all synthetic estimators by assuming $\text{cov}(\hat{\bar{Y}}_{j,DRCT}, \hat{\bar{Y}}_{j,STH}) \doteq 0$, Ghosh and Rao(1994). The optimal weight turns out to be

$$\gamma_{j,opt} = \frac{\text{MSE}(\hat{\bar{Y}}_{j,STH})}{\text{MSE}(\hat{\bar{Y}}_{j,STH}) + \text{V}(\hat{\bar{Y}}_{j,DRCT})}$$

Given a direct estimator $\hat{\bar{Y}}_{j,DRCT}$ for the jth small area, a James-Stein estimator is obtained by using a common weight for all the small areas, then minimizing the average estimated MSE with respect to this common weight by assuming the variances of the $\hat{\bar{Y}}_{j,DRCT}$'s are approximately equal. If the individual variances, the $V(\hat{\bar{Y}}_{j,DRCT})$'s, vary considerably then the James-Stein estimator can be less efficient than the director estimator for some individual areas.

To compare direct estimators and traditional indirect estimators, let us consider three simulated small areas. The covariate/auxiliary variable is simulated by taking values uniformly from an interval which is slightly different for each small area. For each individual, the variable of interest is computed as the sum of these three terms:

- **Fixed term** based on the covariate ($\beta = 0.21$)
- **Random term based on the area level effect**. These effects have been simulated from a normal distribution with mean 0 and variance 0.25.
- **Random term based on the individual variation**. It is the product of a

normal random variable with mean 0 and variance 1, multiplied by the square root of the value of the covariate.

We draw a random sample from each small area (sample sizes are 3, 7 and 10), and then calculate the sample mean as the direct estimator, the synthetic estimator based on formula (2.1) and get the composite estimator as their weighted average. The results are summarized in the following figure.

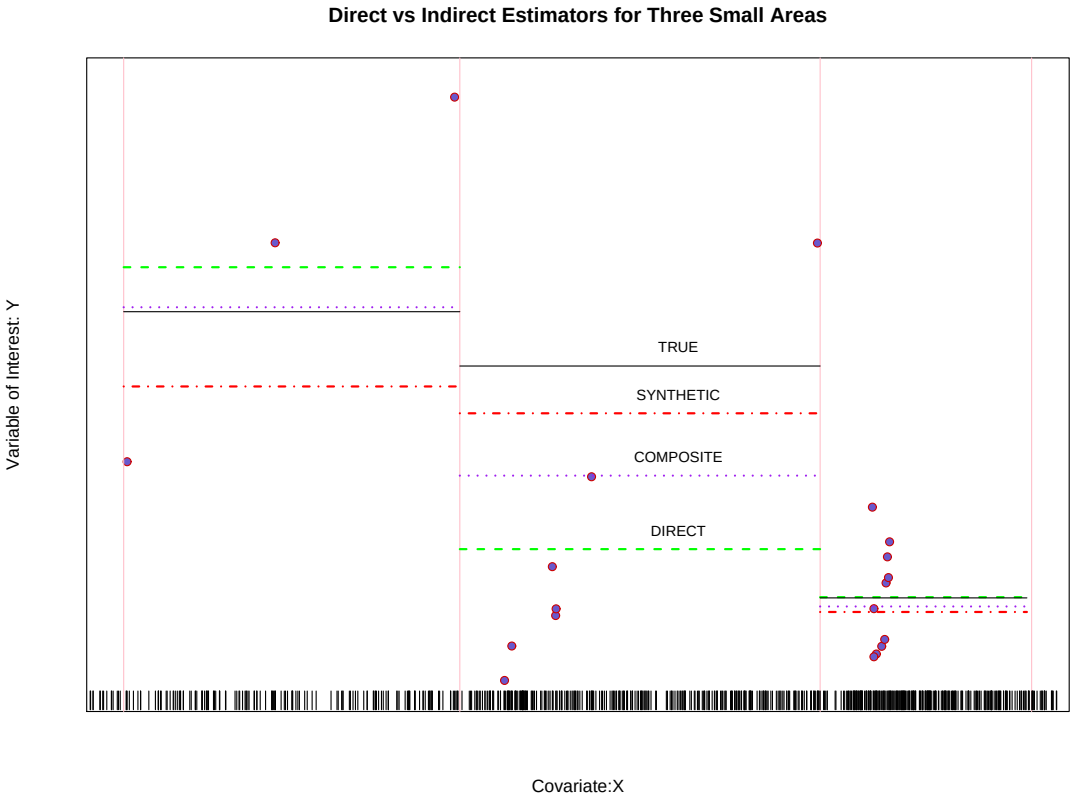

Figure 2.1: Direct and indirect estimators for simulated data

From Figure 2.1, we can see that the design variances (i.e., variances with respect to the probability distribution induced by the sampling design) of the indirect estimators

are usually smaller than the design variances of the direct estimators. Since the direct estimators are not reliable, all three synthetic estimators are smaller than their true values in this example, i.e. the bias in the direct estimators is passed on to the synthetic estimators. Nevertheless the shrinkage effect of the synthetic estimator against the direct estimator is significant after borrowing strength through the linking linear model. The composite estimator is worse than synthetic estimator for small area 2 after introducing the bad direct estimator, but the overall bias of composite estimators for all small areas are reduced. When the sample size is large enough (e.g. for small area 3), the performance of the three estimators are almost the same.

2.3 Basic Small Area Models

In this section, we will discuss explicit small area models that make specific allowance for between area variation. Two types of such models have been proposed in the literature: (i) Aggregate level(or area level) models that relate the small area means to area-specific auxiliary variables. Such models are essential if unit (element) level data are not available. (ii) Unit level models that relate the unit values of the study variable to unit-specific auxiliary variables.

In the basic area level model (Type I Model), we assume that θ_j, the small area mean of variable of y, is related to area-specific auxiliary data $\mathbf{X}_j = (X_{j1}, \cdots, X_{jm})$, by the equation

$$\theta_j = \beta \mathbf{X}_j + b_j \nu_j, \quad j = 1, \cdots, M, \tag{2.3}$$

where the b_j's are known positive constants, β is the $m \times 1$ vector of regression coefficients. The ν_j's are area-specific random effects assumed to be independent and identically distributed (iid) random variables with

$$E(\nu_j) = 0, \qquad V(\nu_j) = \sigma_\nu^2. \tag{2.4}$$

In addition, normality of the random effects ν_j is also often assumed. The parameter σ_ν^2 is a measure of homogeneity across areas after taking account of the covariates $\mathbf{X}_j$. To make inference about θ_j's, we assume that the direct estimator $\hat{\theta}_j$ is available and

$$\hat{\theta}_j = \theta_j + e_j, \quad j = 1, \cdots, M$$

where the sampling errors e_j are independent with $E(e_j|\theta_j) = 0$ and $V(e_j|\theta_j) = \varphi_j$. This actually leads to a special case of the generalized mixed linear model

$$\hat{\theta}_j = \beta \mathbf{X}_j + b_j \nu_j + e_j, \quad j = 1, \cdots, M \tag{2.5}$$

This model (Fay and Herriot(1979)) assumes that the direct estimators are design unbiased, which may not be valid since $\hat{\theta}_j$ may be seriously biased if θ_j is a nonlinear function of the population total or if the sample size is small.

In the basic unit level model (Type II Model), unit-specific auxiliary data are available for each population element l in each small area j. Thus the population means $\bar{X}_j$ are also known. The variable of interest, y_{jl}, is assumed to be related to $\mathbf{x}_{jl} = (x_{jl1}, \cdots, x_{jlm})$ through a nested error regression model:

$$y_{jl} = \beta \mathbf{x}_{jl} + \nu_j + e_{jl}, \quad l = 1, \cdots, N_j, \ j = 1, \cdots, M. \tag{2.6}$$

Here, the area-specific effects ν_j are assumed to be iid random variables satisfying (2.4), $e_{jl} = \tilde{e}_{jl} k_{jl}$ with known constants k_{jl}'sw and $\tilde{e}_{jl}$'s are iid random variables independent of the ν_j's, and

$$E(\tilde{e}_{jl}) = 0, \qquad V(\tilde{e}_{jl}) = \sigma_e^2,$$

In addition, normality of the ν_j's and $\tilde{e}_{jl}$'s is often assumed. This model is valid when we assume that sample values also obey the assumed model, which is satisfied under simple random sampling from each area or more generally for sampling designs where

the sample selection probabilities do not depend on y_{jl}'s, but may depend on auxiliary information(Smith(1983)).

We write the small area mean $\bar{Y}_j$ as

$$\bar{Y}_j = f_j \bar{y}_j + (1 - f_j)\bar{Y}_j^*$$

with $f_j = n_j/N_j$ and $\bar{y}_j$ and $\bar{Y}_j^*$ denoting the means of the sampled and nonsampled elements, respectively. Thus the estimation of a small area mean $\bar{Y}_j$ is equivalent to estimating the realization of the random variable $\bar{Y}_j^*$ given the sample data and auxiliary data.

Even if the data are collected with a random sampling procedure, model-based estimates do not involve the sampling weights. This follows from the fact that if the superpopulation model is true, then it holds true for all observations, regardless of how likely they are selected under a particular sampling design (Valliant et al., 2000). Thus a model-based estimator's statistical properties must be defined only in terms of the model. The use of explicit models has several advantages. Model diagnostics can be used to find suitable models that fit the data well such as residual analysis, which can detect departures from the assumed model. Area-specific measures of variability can be associated with each estimator, unlike the global measures (average over small areas) often used in traditional indirect estimators. Similar procedures can be extended to the cases where the variables of interest are multivariate as well as the case of binary response. On the other hand, the disadvantage of model-based estimators is that the assumed linking model may not be valid when the relationship between variables is complicated. Therefore any inference based on the model would not be convincing.

2.4 Model Based Estimates

In section 2.3, we introduced several small area models that may be regarded as special cases of a general mixed linear model involving fixed and random effects. Then the small

area parameters can be expressed as linear combinations of these effects. If we apply the general results of best linear unbiased prediction (BLUP) estimation, we are able to derive BLUP estimators of these parameters in the classical frequentist framework. These estimators minimize the MSE among the class of linear unbiased estimators and do not depend on normality of the random effects. But they depend on the variance parameters which are unknown. An empirical best linear unbiased prediction (EBLUP) estimator is obtained from the BLUP by substituting suitable estimators of variance parameters by an asymptotically consistent estimator (Harville(1991).

The BLUP estimator of $\theta_j = \beta \mathbf{X}_j + b_j \nu_j$ is just a weighted average of the direct estimator $\hat{\theta}_j$ and the regression-synthetic estimator $\hat{\beta}\mathbf{X}_j$ under the first type of model in section 2.3. The weights are determined by σ_ν^2 and φ_j, so the weights take proper account of between area variation relative to the precision of the direct estimator. It adjusts the synthetic estimator to account for model uncertainty. It is important to note that the BLUP estimator under the type I model is valid for general sampling designs since we are modeling only the θ_j's and not the individual elements in the population. The BLUP estimator under the type II model is a weighted average of the "survey regression" estimator (Prasad and Rao, 1990) and the regression synthetic estimator. Under each type of model, replacing the unknown variance parameters by their estimators, which can be estimated by using maximum likelihood (ML) or restricted maximum likelihood (REML) methods, leads to a two-stage estimator referred to as the EBLUP estimator. Ghosh and Rao (1994) and Rao (2003) presented the theoretical results about estimating the MSE of EBLUP estimators. In their theories, the normality assumption of random effects and errors is not needed for point estimation, but the normality assumption is often used for getting accurate MSE estimators.

2.5 Empirical Bayes Method

The EBLUP method is applicable to linear mixed models. However linear mixed models are designed for continuous variables. They are not suitable for handling binary or count data. Empirical Bayes (EB) and hierarchical Bayes (HB) methods are applicable more generally in handling binary and count data.

Morris (1983) gave an excellent account of the EB approach. Under the type I model, the EB approach may be summarized as follows:

Step 1. Obtain the posterior density, $f(\theta_j|\hat{\theta}_j, \beta, \sigma_\nu^2, \varphi_j)$, of the parameters of interest, given the data $\hat{\theta}_j$ (or $\mathbf{y}$). This can be accomplished by using the conditional density, $f(\hat{\theta}_j|\theta_j, \varphi_j)$ and the density $f(\theta_j|\beta, \sigma_\nu^2)$, where f denotes the density function of the normal distribution under the type I model.

Step 2. Estimate the model parameters $(\beta, \sigma_\nu^2, \varphi_j)$ from the marginal distribution of the data, $f(\hat{\theta}_j|\beta, \sigma_\nu^2, \varphi_j)$.

Step 3. Use the posterior density, $f(\theta_j|\hat{\theta}_j, \hat{\beta}, \hat{\sigma}_\nu^2, \hat{\varphi}_j)$, for making inferences about θ_j, where $\hat{\beta}, \hat{\sigma}_\nu^2, \hat{\varphi}_j$ are estimators of unknown parameters.

The EB and EBLUP estimators are identical under the type I model with normal errors and quadratic loss. When we handle discrete data and the normality assumption is not valid, we just need to replace f by other underlying density functions, for instance, f could be derived under the logistic regression model if binary data is of interest.

Datta, Fay and Ghosh (1991) applied the hierarchical Bayes (HB) approach to estimation of small area means, $\bar{Y}_j$'s, under general mixed linear models. In the HB approach, a prior distribution on the model parameters is specified, e.g. assuming β has a uniform distribution. Then the posterior distribution of the parameters of interest is obtained and a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance. The HB and BLUP estimators are identical under the type I model when a noninformative prior is assumed for β and σ_ν is

assumed to be known. The HB and BLUP approaches under the type II model lead to identical inferences when a noninformative prior is assumed for β and both σ_ν and σ^2 are assumed to be known. The HB approach can incorporate prior information on these parameters through informative priors, which allows more sensible estimators. Ghosh and Rao (1994) compared several estimators and through their examples they found that EBLUP and HB estimators are asymptotically identical when the number of small areas M goes to ∞. The standard error values for these two estimators are also similar.

Chapter 3

A Noninformative Bayesian Approach to Small Area Estimation

Much like the model-based perspective, Bayesian sampling theory emphasizes the prediction of unsampled units. A Bayesian approach assigns a prior distribution to the population parameter and applies Bayes' rule to estimate unknown parameters based on the predictive distribution of unobserved units in the population given the observed ones. The Bayesian perspective described here will make use of known auxiliary variables without assuming a linear relationship between the variables.

3.1 The Bayesian Approach in Finite Population Sampling

The Bayesian approach to statistical inference summarizes information concerning a parameter through its posterior distribution, which depends on a model and a prior distribution and is conditional on the observed data. In finite population sampling, the unknown parameter is just the entire population and the likelihood function for the

model comes from the sampling design. A Bayesian must specify a prior distribution over all possible values of the population. Once the sample is observed the posterior is just the conditional distribution of the unobserved units given the values of the observed units computed under the prior distribution for the population. Basu(1969) showed that after the sample has been observed, the sampling design plays no role in the posterior distribution for a Bayesian. (For this fact and more of Basu's thoughts on finite population sampling see Ghosh(1988).).

Assume that given the sample one can simulate values for all the unobserved units in the population from the posterior to generate a "complete copy" of the population. Then given the simulated and observed values any finite population quantity-means, totals, medians, may now be calculated from this simulated copy of the entire population. By simulating many such independent complete copies and in each case finding the quantity of interest for the simulated population the posterior for the desired population quantity is generated. One computes an estimate of the unknown population quantity by taking the average of these simulated values. This process computes approximately the Bayes estimate of the population quantity of interest under the squared error loss for the given prior. The problem then is to find a sensible Bayesian model which utilizes the type of prior information available for the problem at hand.

In the literature, the Polya posterior is a noninformative Bayesian approach to finite population sampling which uses little or no prior information about the population. It is a means of relating the unseen units to the seen ones. In the design based approaches this is often done through the sampling design and unbiasedness. Given the data, the Polya posterior is a predictive joint distribution for the unobserved units in the population conditioned on the values in the sample, which is based upon Polya sampling. To illustrate the basic mechanism of Polya sampling, we first consider a survey for which we believe the observed units are roughly exchangeable with the unseen units, but for which we have little prior information about the population. This is usually the case when simple random sampling is used to select the sample. For such surveys, Polya

posterior works as follows: Suppose that we have a sample of size n from a finite population of size N, and the values from n observed units are marked on n balls and placed in urn 1. The unseen $N - n$ units are represented by $N - n$ unmarked balls placed in urn 2. We begin by choosing one ball at random from each urn, and assigning the value of the ball from urn 1 to the ball from urn 2. Both balls are then returned to urn 1. Thus at the second stage of Polya sampling, urn 1 has $n + 1$ balls and urn 2 has $N - (n + 1)$ balls. Next another ball is chosen at random from each urn and we assign the value of the ball from urn 1 to the ball from urn 2. Again, both balls are then returned to urn 1. This process is continued until urn 2 is empty and all $N - n$ unobserved units are assigned a value, at which point the N balls in urn 1 constitute a complete copy of the population. Once this is done we have generated one realization of the complete population from the Polya posterior distribution. Hence by simple Polya sampling we have a predictive distribution for the unobserved given the observed.

The Polya posterior is appropriate when there is little known about the population and the sample is assumed to be representative of the population, i.e. the observed and unobserved units are roughly exchangeable. One can verify that under the Polya posterior distribution, the expected value of the population mean is just the sample mean and its posterior variance is approximately the frequentist variance of the sample mean under a simple random sampling design when $n \geq 25$. It has been shown for a variety of decision problems that procedures based on the Polya posterior are admissible because they are stepwise Bayes (Ghosh and Meeden(1997)). The Polya posterior is not a proper posterior since it does not come from a proper Bayesian model. It is the stepwise Bayes nature of the Polya posterior that explains its somewhat surprising properties. Given a sample it behaves just like a proper Bayesian posterior but the collection of all the possible posteriors does not come from a single prior but from a family of priors.

The interval estimate of the population mean and point and interval estimates for other population quantities under the Polya posterior usually cannot be found explicitly. One must use simulation to find these values approximately. This is done by simulating

many independent completed copies of the entire population and calculating the value of the parameter of interest for each copy. There is a second way to think about simulating complete copies of the population using the Polya posterior distribution. For simplicity assume that the sample values $\{y_1, \cdots, y_n\}$ are all distinct and that the sampling fraction $f = n/N$ is small, i.e. the population size N is large compared to the sample size n. For $i = 1, \cdots, n$, let λ_i be the proportion of units in a complete simulated copy of the entire population which take on the value y_i. Then under the Polya posterior $\lambda = (\lambda_1, \cdots, \lambda_n)$ has approximately the Dirichlet distribution with a parameter vector of all ones. i.e. it is uniform on the $n-1$ dimensional simplex where $\sum_{i=1}^{n} \lambda_i = 1$.

3.2 A Bayesian Model for Small Area Estimation

We may apply the Polya Posterior approach to each small area population to produce simulated complete copies of the population such that each copy consists of only the values appearing in the sample of the small area. Then an estimator of the population mean is obtained by taking the average of the means of these simulated copies. However, when the sample size is small, the accuracy of the predictive distribution for the unobserved given the observed may be poor under the Polya posterior framework. For example, suppose the sample size of a small area is 1, then Polya sampling procedure ensures that each simulated complete copy of the small area only consists of values appeared in the sample, i.e. a single value in this case. Thus the estimator of the small area mean is same as the only sampled value and the resulting estimator will have a large variance. For small area problems the central issue to be addressed is how to find a Bayesian way to borrow strength from related or similar areas and thus increase the "effective" sample size.

Regardless of the type of the approach, a common prerequisite of "borrowing strength" assumes that there exists a similarity across related small areas. In the model assisted

approach, this is done through the assumption that different areas may share common linear coefficients in the model. In a Bayesian perspective, this can be done by assuming that the parameters of small areas are related through similar prior distributions. However finding such prior distributions is often difficult since the prior information about parameters of small areas is not always available in practice. Thus additional assumptions are needed in order to utilize Bayesian ideas to borrow strength. Given a set of related small areas, an assumption about the similarity may be based on a belief that most elements belonging to a small area population may also belong to the other related small area populations. This assumption is reasonable in many situations. For example, it is satisfied when the underlying small areas are very much alike. In terms of the sample, this assumption may be interpreted in the manner that the observed units from different small areas are roughly exchangeable in addition to the assumption that the unobserved units are exchangeable with the observed ones within each area. To identify the sampling units, we will call the observed units from a specific small area local units, and call the observed units from other small areas foreign units. Then the assumption of exchangeability allows foreign units to be treated as importantly as local units from this specific small area when we estimate the parameter of this area.

3.2.1 The Parameter Space Is Finite

Suppose we have M small areas. The jth small area consists of N_j units labeled $1, 2, \cdots, N_j$ for $j = 1, \cdots, M$. The labels are assumed to be known and to contain no information. For each unit $l = 1, \cdots, N_j$ in small area j let y_{jl} be the unknown value of some characteristic of interest. Typically, y_{jl} will be a real number. Thus $\mathbf{y}^j = (y_{j1}, \cdots, y_{jN_j})$ is an unknown parameter of interest attached to the jth small area. $\mathbf{y}^j$, is assumed to belong to a parameter space $\mathcal{Y}^j$, a subset of N_j-dimensional Euclidean space. Sometimes one will assume that $\mathcal{Y}^j$ contains just finitely many elements

or vectors based on some previous knowledge. In such cases it is denoted by

$$\mathcal{Y}^j(\mathbf{b}^0) = \{\mathbf{y}^j : \text{ such that for } l = 1, \cdots, N_j, \ y_{jl} = b_i \text{ for some } i = 1, \cdots, K\} \quad (3.1)$$

where $\mathbf{b}^0 = (b_1, \cdots, b_K)$ denotes the K distinct known values that can be seen in all M small area populations. Notice that the b_i's are common values for all the small area populations reflecting the assumption that there exists a similarity among the small areas. One possible example is the case where the variable of interest y is binary. Then we have that $\mathbf{b}^0 = (0, 1)$ and the population unit y_{jl} is either 0 or 1. Thus the parameter space $\mathcal{Y}^j(\mathbf{b}^0)$ consists of 2^{N_j} elements each of length equal to N_j. Note that the actual form of $\mathcal{Y}^j(\mathbf{b}^0)$ is not important in our study here; what is important, however, is that $\mathcal{Y}^j$ contains just finitely many elements. As long as $\mathcal{Y}^j$ is finite for $j = 1, \cdots, M$, we can always find a vector $\mathbf{b}^0$ such that (3.1) holds.

A **sample** s^j of small area j is a subset of $\{1, 2, \cdots, N_j\}$ and S^j represents all possible samples. A sampling design is a probability measure p^j defined on S^j such that $p^j(s^j) \in [0, 1]$ for every nonempty $s^j \in S^j$ and $\sum_{s^j \in S^j} p^j(s^j) = 1$. For convenience, we will assume the sampling design p^j is the simple random sampling p for all j's. Let n_j denote the number of elements in s^j and $s^j = \{l_1, \cdots, l_{n_j}\}$.

Given a parameter space $\mathcal{Y}^j$ and design p, a **sample point** consists of the set of observed labels s^j along with the corresponding y values. Let z^j denote the observed y values in the jth small area sample s^j , $z^j = (z_{j1}, \cdots, z_{jn_j})$, where z_{jh} is the observation associated with label l_h. Then a sample point is denoted by

$$\mathbf{z}^j = (s^j, z^j) = (s^j, (z_{j1}, \cdots, z_{jn_j})^T)$$

The sample space for small area j is the collection of all possible sample points depending on both the parameter space and the design, and it will be written as

$$\mathcal{Z}(\mathcal{Y}^j, p) = \left\{(s^j, z^j) : p(s^j) > 0 \text{ and } z^j = \mathbf{y}^j(s^j) \text{ for some } \mathbf{y}^j \in \mathcal{Y}^j\right\}$$

where $\mathbf{y}^j(s^j) \triangleq (y_{jl_1}, \cdots, y_{jl_{n_j}})$.

For a fixed design and a fixed parameter point there are only finitely many points of the sample space which receive positive probability. The sample space contains finitely many points since the parameter space $\mathcal{Y}^j(\mathbf{b}^0)$ contains finitely many members. However if the parameter space is $\mathcal{R}^{N_j}$ then the sample space contains uncountably many data points. In the thesis the parameter space is always assumed to have the form of (3.1).

3.2.2 A Noninformative Stepwise Bayes Approach

A Bayesian will assign a prior distribution for $\mathbf{y}^j$ over the parameter space $\mathcal{Y}^j(\mathbf{b}^0)$, and make inference based on the posterior distribution of the parameter given the data. However, it is not easy to find a sensible prior distribution when we have little prior information about $\mathbf{y}^j$. Instead of searching for a single proper prior over $\mathbf{y}^j$ we will construct a sequence of prior distributions over subsets of $\mathcal{Y}^j(\mathbf{b}^0)$ such that the resulting posterior distributions are able to borrow strength across small areas.

Suppose we wish to estimate the jth small area population mean, $\mu^j = \sum_{l=1}^{N_j} y_{jl}/N_j$. We will show how to compute our estimator of μ^j given a sample point, and then a theoretical justification will follow. We start building the estimator of μ^j by making use of not only local observations z^j but also other foreign observations $z^{j'}$'s for $j' \neq j$. Let

$$Z = \{(s^1, z^1), (s^2, z^2), \cdots, (s^M, z^M)\}$$

denote a pooled small area sample point and $b = (b_1, \cdots, b_k)$ be the k distinct values that appear in this pooled set of observed values in all the small areas of interest, where $k \leq K$. For an example of our notation, let us consider three small areas and each area provides a sample of size 4. The values contained in these three samples are expressed by $z^1 = (2, 4, 5, 8)$, $z^2 = (1, 4, 3, 6)$ and $z^3 = (3, 5, 2, 1)$ respectively. Then b is taken to be $(2, 4, 5, 8, 1, 3, 6)$ and $k = 7$, where the order of the components of b does not matter.

Given a pooled sample point Z, we will assume that this sample is representative, i.e.

the "seen" and "unseen" units are exchangeable. In addition we will assume that the "seen" units between areas are also exchangeable. The first assumption is very similar to the Polya posterior approach, and the second assumption is also appropriate when the small areas are similar to each other. Under these assumptions, the unseen elements in the population can be represented by the seen ones to some extent regardless of the source of the seen elements. Hence we would assign a prior to our parameter $\mathbf{y}^j$ such that only those $\mathbf{y}^j$'s whose elements are consistent with the pooled observed values will have a chance to receive positive mass. More specifically, we will define a noninformative prior over the following space

$$\mathcal{Y}^j(b) = \{\mathbf{y}^j : \text{ such that for } l = 1, \cdots, N_j, \ y_{jl} = b_i \text{ for some } i = 1, \cdots, k\} \tag{3.2}$$

where $b = (b_1, \cdots, b_k)$ are the k distinct values that appear in the pooled sample Z. Obviously the length of b, k, depends on the sample Z, and thus $\mathcal{Y}^j(b)$ depends on what we have seen in the sample. $\mathcal{Y}^j(b)$ is a subset of the original parameter space $\mathcal{Y}^j(\mathbf{b}^0)$. Since $\mathcal{Y}^j(b)$ depends on the sample, so the prior that we will use later is not a single proper prior over $\mathcal{Y}^j(\mathbf{b}^0)$.

To count the number of y-values which take on particular b-values, we use the following notation. For a parameter point $\mathbf{y}^j \in \mathcal{Y}^j(b)$,

$$c_{\mathbf{y}^j}(i) = \text{number of } y_{jl}\text{'s in } \mathbf{y}^j, \text{ which equal } b_i.$$

It is clear that $c_{\mathbf{y}^j}(i) \geq 0$ and $\sum_{i=1}^{k} c_{\mathbf{y}^j}(i) = N_j$. For a sample point (s^j, z^j) of small area j,

$$n_{ji} = \text{the number of } z_{jl}\text{'s in } z^j, \text{ which equal } b_i, \text{ then } \sum_{i=1}^{k} n_{ji} = n_j.$$

(Note that $n_{ji} \geq 0$. It is positive only when b_i is observed in the jth small area sample.)

Finally we can propose a prior for $\mathbf{y}^j$ over the parameter space $\mathcal{Y}^j(b)$. In the following expression, note that the θ_i's in the Dirichlet integral must sum to 1. A prior for our

parameter $\mathbf{y}^j$ for small area j is defined by the following:

$$\begin{aligned}
\pi(\mathbf{y}^j) &\propto \int_0^1 \cdots \int_0^1 \prod_{i=1}^{k} \theta_i^{c_{\mathbf{y}^j}(i)-1+\epsilon} d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{k-1} \theta_i^{c_{\mathbf{y}^j}(i)-1+\epsilon} \left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{c_{\mathbf{y}^j}(k)-1+\epsilon} d\theta_1 \cdots d\theta_{k-1} \qquad (3.3) \\
&= \frac{\prod_{i=1}^{k} \Gamma(\, c_{\mathbf{y}^j}(i) + \epsilon)}{\Gamma(N_j + k\epsilon)}
\end{aligned}$$

where ϵ is a known positive number.

As we have mentioned before, the support of this prior depends on what we have seen in the sample, so what we have defined here is not a proper prior for the original parameter space $\mathcal{Y}^j(\mathbf{b}^0)$. However, once the sample point is fixed, this prior can be treated as a proper prior over the subspace, $\mathcal{Y}^j(b)$. On one hand, the pseudo prior mass of $\mathbf{y}^j$ depends on the number of its elements which take on particular b-values, and the order of elements here does not matter. On the other hand, the prior depends on the hyperparameter ϵ. The positiveness of ϵ ensures that no matter whether the unit b_i appears in the jth small area sample or not, it always contributes to the prior of $\mathbf{y}^j$ as long as it is observed in the other related areas. Such a property allows us to borrow the strength from related areas successfully through a Bayesian perspective. This idea is quite different from the traditional methodology in finite population sampling. When ϵ is large, the influence of $c_{\mathbf{y}^j}(i)$'s, the number of b_i's in $\mathbf{y}^j$, on the prior distribution is less important, which means more similarities are assumed to be shared across the different small areas.

Given the pseudo prior above, we next discuss finding its Bayes estimator for a small area mean.

3.2.3 Our Estimator

For a fixed design and a fixed parameter point $\mathbf{y}^j$, a probability function of a sample point (s^j, z^j) can be obtained based on the sampling design (Ghosh and Meeden 1997) as follows:

$$\pi(z^j|\mathbf{y}^j) = \begin{cases} p(s^j) & \text{if } z^j = \mathbf{y}^j(s^j) \\ 0 & \text{otherwise .} \end{cases}$$

When the sampling design p is simple random sampling and the sample size n_j is fixed, this probability function is just constant over the sample points which are consistent with the given parameter point and 0 elsewhere. With this probability function for the sample points and a prior distribution for the parameter, Bayesian inference is based upon the posterior distribution of the parameter.

Suppose we are interested in estimating the jth small area population mean, $\mu^j = \sum_{l=1}^{N_j} y_{jl}/N_j$. Our estimator of μ^j is the posterior expectation against the pseudo prior defined in (3.3):

$$\hat{\mu}^j\ (z^1, z^2, \cdots, z^M) = E\left(\sum_{l=1}^{N_j} y_{jl}|z^1, z^2, \cdots, z^M\right)/N_j, \tag{3.4}$$

where M is the number of small areas. The estimator for jth small area depends on the samples from the other areas since the pseudo prior relies on the pooled samples from all the small areas, when we are borrowing strength across areas.

Lemma 1. The posterior expectation of jth small area defined in (3.4) is given by

$$\hat{\mu}^j\ (z^1, z^2, \cdots, z^M) = \sum_{i=1}^{k} b_i \frac{n_{ji} + \epsilon}{n_j + k\epsilon}$$

where $j = 1, \cdots, M$.

Proof. To find $\hat{\mu}^j$, we need to compute the conditional probability $\pi(\mathbf{y}^j|z^1, z^2, \cdots, z^M)$, given our pooled sample point z^j's. For simplicity let us assume the sample s^j consists

of only the first n_j units in small area j. The unseen units in the jth small area are then denoted by $\{y_{jl} : l = n_j + 1, \cdots, N_j\}$. Among the unseen ones, there can be r_{ji} many units equal to b_i, $i = 1, 2, \cdots, k$, in some specified order, where $\sum_{i=1}^{k} r_{ji} = N_j - n_j$.

Let $\mathbf{y}^{j\,\prime} \in \mathcal{Y}^j(b)$ be such a point whose first n_j elements are consistent with the observed data, i.e., $\mathbf{y}^{j\,\prime}(s^j) = z^j$ or equivalently $(y'_{j1}, \cdots, y'_{jn_j}) = z^j$. Then under the prior defined in (3.3), the probability that we will have a parameter point whose first n_j elements are consistent with z^j is

$$
\begin{aligned}
\pi(\mathbf{y}^j(s^j) = z^j) &\propto \sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=z^j} \pi(\mathbf{y}^{j\,\prime}) \\
&= \sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=z^j} \int_0^1 \cdots \int_0^1 \prod_{i=1}^k \theta_i^{c_{\mathbf{y}^{j\,\prime}}(i)-1+\epsilon} d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 \sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=z^j} \prod_{i=1}^k \theta_i^{c_{\mathbf{y}^{j\,\prime}}(i)-1+\epsilon} d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 \sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=z^j} (\prod_{i=1}^k \theta_i^{n_{ji}-1+\epsilon})(\prod_{i=1}^k \theta_i^{r'_{ji}}) d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 (\prod_{i=1}^k \theta_i^{n_{ji}-1+\epsilon}) \sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=z^j} (\prod_{i=1}^k \theta_i^{r'_{ji}}) d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 (\prod_{i=1}^k \theta_i^{n_{ji}-1+\epsilon}) \sum_{\substack{r'_{ji}\geq 0 \\ \sum_{i=1}^k r'_{ji}=(N_j-n_j)}} (\prod_{i=1}^k \theta_i^{r'_{ji}}) d\theta_1 \cdots d\theta_k \\
&= \int_0^1 \cdots \int_0^1 (\prod_{i=1}^k \theta_i^{n_{ji}-1+\epsilon}) \cdot 1 \cdot d\theta_1 \cdots d\theta_k \\
&= \frac{\prod_{i=1}^k \Gamma(n_{ji}+\epsilon)}{\Gamma(n_j + k\epsilon)}
\end{aligned}
$$

Using this fact it is easy to calculate the posterior marginal distribution of any unobserved unit given the pooled observed sample point $(z^1, \cdots, z^M)$. For an example,

for $l > n_j$, we have that

$$\begin{aligned}
\pi(y_{jl} = b_1 | z^1, \cdots, z^M) &= \frac{\pi(y_{jl} = b_1, \text{ and } \mathbf{y}^j(s^j) = z^j, z^1, \cdots, z^{j-1}, z^{j+1}, \cdots, z^M)}{\pi(\mathbf{y}^j(s^j) = z^j, z^1, \cdots, z^{j-1}, z^{j+1}, \cdots, z^M)} \\
&= \frac{\pi(y_{jl} = b_1, \text{ and } \mathbf{y}^j(s^j) = z^j)}{\pi(\mathbf{y}^j(s^j) = z^j)} \\
&= \frac{\Gamma(n_{j1} + 1 + \epsilon) \prod_{i=2}^{k} \Gamma(n_{ji} + \epsilon)}{\frac{\Gamma(n_j + 1 + k\epsilon)}{\prod_{i=1}^{k} \Gamma(n_{ji} + \epsilon)}} \\
&= \frac{n_{j1} + \epsilon}{n_j + k\epsilon}
\end{aligned}$$

Similarly, the marginal probability of getting a value b_i on the lth unit for $l > n_j$ is

$$\pi(y_{jl} = b_i | z^1, \cdots, z^M) = \frac{n_{ji} + \epsilon}{n_j + k\epsilon}$$

So we can calculate our estimator as follows:

$$\begin{aligned}
\hat{\mu}^j \ (z^1, z^2, \cdots, z^M) &= E\left(\sum_{l=1}^{N_j} y_{jl} | z^1, z^2, \cdots, z^M \right) / N_j \\
&= \left(\sum_{i=1}^{k} n_{ji} b_i + (N_j - n_j) \sum_{i=1}^{k} \frac{n_{ji} + \epsilon}{n_j + k\epsilon} b_i \right) / N_j \\
&= \sum_{i=1}^{k} \frac{n_{ji} + \epsilon}{n_j + k\epsilon} b_i + \sum_{i=1}^{k} \frac{\epsilon(k n_{ji} - n_j)}{N_j (n_j + k\epsilon)} b_i
\end{aligned}$$

Clearly when N_j is sufficiently large and n_j / N_j is small, the estimator can be obtained as the following

$$\hat{\mu}^j \ (z^1, z^2, \cdots, z^M) = \sum_{i=1}^{k} b_i \frac{n_{ji} + \epsilon}{n_j + k\epsilon}$$

and the proof is complete. □

Let us take a further look at the posterior distribution of $\mathbf{y}^j$. The conditional probability of $\mathbf{y}^j$ given the sample point is

$$\begin{aligned}
\pi(\mathbf{y}^j|z^1, z^2, \cdots, z^M) &= \frac{\pi(z^j|\mathbf{y}^j)\pi(\mathbf{y}^j)}{\sum_{\mathbf{y}^{j\,\prime}:\, \mathbf{y}^{j\,\prime}(s^j)=\mathbf{y}^j(s^j)} \pi(z^j|\mathbf{y}^{j\,\prime})\pi(\mathbf{y}^{j\,\prime})} \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^k \theta_i^{n_{ji}+r_{ji}-1+\epsilon} d\theta_1 \cdots \theta_k}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^k \theta_i^{n_{ji}-1+\epsilon} d\theta_1 \cdots \theta_k} \qquad (3.5) \\
&= \frac{\prod_{i=1}^k \frac{\Gamma(n_{ji}+r_{ji}+\epsilon)}{\Gamma(n_{ji}+\epsilon)}}{\frac{\Gamma(N_j+k\epsilon)}{\Gamma(n_j+k\epsilon)}}
\end{aligned}$$

We can see that the mass in the posterior distribution of $\mathbf{y}^j$ only depends on the frequencies of b_i's, and it is not related to their order.

3.2.4 Generate Samples from The Posterior

Given a sample we now show how to generate a set of possible values for the unobserved units from this posterior distribution given the observed ones. For definiteness assume that we are interested in estimating the mean of the first small area. The other small areas are handled in exactly the same way.

Following the previous notation, let $b_1, \cdots, b_k$ be the k distinct units appearing in the pooled samples from all small areas and n_{1i} be the number of units which equal b_i for $i = 1, \cdots, k$ in the sample from the first small area. Without loss of generality, assume that only the first k^* values $b_1, \cdots, b_{k^*}$ for some $k^* \leq k$ appeared in the sample

from the first small area. If the sample size of the first small area is n_1, then $\sum_{i=1}^{k} n_{1i} = \sum_{i=1}^{k^*} n_{1i} = n_1$, and $n_{1i} = 0$ for $k^* < i \leq k$.

Consider three urns where the first contains the n_1 observed units (or balls) from the the sample of the first small area, the second contains the $N_1 - n_1$ unsampled units from the first small area and the third contains $n_{1i} + \epsilon$ many units of value b_i for $1 \leq i \leq k^*$ as well as containing ϵ many units of value b_i for $k^* < i \leq k$, where ϵ is the hyperparameter of the prior defined in formula (3.3). Thus the overall number of units in the third urn is $n_1 + k^*\epsilon + (k - k^*)\epsilon = n_1 + k\epsilon$.

We pick a ball at random from the second urn and the third urn, assign the value of the ball from the third urn to the ball from the second urn. We then place the ball drawn from the second urn into the first urn and return the ball from the third urn and another copy of it back into urn three. We keep doing this until the first urn contains N_1 balls and the second urn is empty.

Suppose after R random draws from the third urn, the first urn contains $(n_{1i} + r_{1i})$ balls with the value of b_i respectively for $i = 1, \cdots, k$, where $\sum_{i=1}^{k} r_{1i} = R$. When $R = N_1 - n_1$ many random drawings have been made, we have generated a complete realization of the first small area population based on the pooled observed units from all small area samples. It can be shown that the resulting probability of getting r_{1i} units of value b_i in some specific order is $\left(\prod_{i=1}^{k} \frac{\Gamma(n_{1i} + r_{1i} + \epsilon)}{\Gamma(n_{1i} + \epsilon)}\right) / \left(\frac{\Gamma(N_1 + k\epsilon)}{\Gamma(n_1 + k\epsilon)}\right)$. This is consistent with what we obtained in formula (3.5). It can be easily shown that this probability does not depend on the order for any possible sequence with same frequencies of b_i's. Thus we have successfully generated one complete realization of the first small area population with the proposed posterior distribution. This simulated, completed copy contains the n_1 observed values along with the $N_1 - n_1$ simulated values for the unobserved members of the small area. Thus, the posterior yields a predictive distribution for the unobserved given the observed.

For the other small areas, if we replace the original members of three urns cor-

respondingly and repeat the aforementioned procedure, then we will be able to get simulated completed copies of the other small area populations. The posterior proposed here is a way of relating the unseen units to the seen ones. The posterior here is a predictive distribution for the unobserved units in the population given the observed values in the sample. The inference based on it does not depend on the sampling design. Its usage is sensible when one believes that the observed values and the unobserved values of the characteristic of interest are roughly exchangeable across the small areas.

3.3 A Useful Approximation

The interval estimate of the population mean and point and interval estimates for other population quantities under the proposed posterior usually cannot be found explicitly. In this section, we will give a way to find these estimators approximately. Let λ_{ji} be the proportion of units in a complete copy of an entire small area which take on the value b_i for i in $\{1 \cdots, k\}$. We have $\sum_{i=1}^{k} \lambda_{ji} = 1$. Let $\lambda_j = (\lambda_{j1}, \cdots, \lambda_{jk})$ and suppose the sampling fraction f_j, n_j/N_j, is small. Since we are interested in small area problems this is a reasonable assumption and as we will see is also mathematically convenient.

In this case if the sample plan is simple random sampling without replacement we can assume that

$$y_{jl}|\lambda_j \underset{appr}{\overset{iid}{\sim}} \text{Multinomial}(1, \lambda_j), \qquad l = 1, 2, \cdots, N_j.$$

i.e $\Pr(y_{jl} = b_i|\lambda_j) = \lambda_{ji},\ i = 1, \cdots, k.$

In this case a convenient form for a prior distribution for λ_j is the Dirichlet. In particular we will assume that it is Dirichlet $(\epsilon, \cdots, \epsilon)$ and so

$$\pi(\lambda_j) \propto \lambda_{j1}^{\epsilon-1}\lambda_{j2}^{\epsilon-1} \cdots \lambda_{jk}^{\epsilon-1},$$

where ϵ is a known positive number. In the future we will see that as ϵ increases the

prior distribution assumes that more similarity is shared across different small areas. On the other hand as ϵ approaches zero the resulting posterior will share little information between small areas.

Thus given the sample $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \cdots, \lambda_{jk})$ has a Dirichlet posterior distribution given by

$$\begin{aligned} \pi(\lambda_j|z^1, z^2, \cdots, z^M) &\propto \pi(z^j|\lambda^j) \cdot \pi(\lambda_j) \\ &\propto \prod_{i=1}^{k} \lambda_{ji}^{n_{ji}} \prod_{i=1}^{k} \lambda_{ji}^{\epsilon-1} \\ &\propto \prod_{i=1}^{k} \lambda_{ji}^{n_{ji}+\epsilon-1} \\ &\sim \text{Dirichlet}(n_{j1}+\epsilon, n_{j2}+\epsilon, \cdots, n_{jk}+\epsilon) \end{aligned} \tag{3.6}$$

where n_{ji} is the number of z_{jl}'s equal to b_i and $\sum_{i=1}^{k} n_{ji} = n_j$. Note that n_{ji} could be 0 when b_i is not observed in the jth small area. If no observation is drawn from some small area, then all corresponding n_{ji}'s are 0.

It is easy to see that the posterior expectations of proportions, $E(\lambda_{ji}|Z)$'s are

$$\begin{aligned} \hat{\lambda}_{ji} &= E(\lambda_{ji}|z^1, z^2, \cdots, z^M) \\ &= \tfrac{n_{ji}+\epsilon}{n_j+k\epsilon} \end{aligned} \tag{3.7}$$

A natural estimator of the population mean for jth small area is then $\sum_{i=1}^{k} \frac{n_{ji}+\epsilon}{n_j+k\epsilon} b_i$, which is same as our estimator derived in the previous section. Thus we have an alternative way to calculate the original posterior estimator. It is clear that our estimator for the jth population mean is a weighted sum of the observed y values from all related small areas, where the weights are determined by the prior and the observed data. From the formulas above, we can see that the positiveness of ϵ guarantees that our estimator

of a small area mean will depend on all values that have appeared in the samples even if a b_i was not observed in the sample of the small area due to the deficiency of small sample size, in which $n_{ji} = 0$ but $\hat{\lambda}_{ji} = \epsilon/(n_j + k\epsilon) > 0$. For a fixed ϵ, if the unit b_i did not appear in both samples of two small areas, then its contribution to the estimator of the area mean is less for the small area with larger sample size than it is for the small area with smaller sample size. This adjustment makes sense since a sample of larger size usually gives us more accurate information about a population than a sample of smaller size.

When the hyperparameter ϵ in the prior is very large, these weights are close to the constant, $1/k$, which does not depend on j very much. As a result all the observed distinct y values in the pooled sample contribute to the estimator equally no matter which area the y values came from, which means that we assume more similarity across small areas for larger values of ϵ. When $\epsilon \longrightarrow 0$, $\hat{\lambda}_{ji} \longrightarrow n_{ji}/n_j$ for jth small area, so the estimator of the small area mean does not depend on the samples from other small areas at all. In practice, it would be appropriate to divide the small areas into a couple of groups to ensure that all the small areas within each group are similar in nature and may share the same ϵ.

From the formula above we can see that making an inference about the small area parameter μ^j is equivalent to making inference about the new parameters λ_{ji}'s as long as k and b_i's are known. Thus in what follows we will concentrate on the parameters λ_{ji}'s, which will be very helpful when we begin making use of auxiliary information.

3.4 Incorporating Auxiliary Information

There are usually many characteristics of interest in a typical survey. Information available at the estimation stage beyond that in the sample is called *auxiliary information*. Such information can be placed into two categories: (1) knowledge of population totals, or means, of characteristics that are observed on the elements of the sample but not

on all elements of the population, and (2) knowledge of characteristics for every element in the population. In terms of small area estimation, the knowledge of first type is generally referred to as area-specific auxiliary information while the second type of knowledge is generally referred to as element-specific auxiliary information.

As an example of the first situation, the age distribution of the population of Minnesota may be treated as known on the basis of a recent census, but the age of people in a sample of households is not known until the households are contacted, and the age of nonsampled persons are unknown. An example of a characteristic known for all households in the population is the geographic location of the households on an address list.

Most existing methods fit "global" linear models relating the variable of interest to auxiliary variables, which maybe perform poorly if the model is incorrectly specified. In the rest of this chapter we will propose a stepwise Bayesian approach to small area estimation, which simultaneously takes into account both area-specific and element-specific auxiliary information. In addition, our approach does not need to assume a linear relationship between variables. The resulting estimator has a stepwise Bayesian justification, which will allow us to prove the admissibility of the estimator.

We assume that in addition to the characteristic of interest, y, the population has a set of auxiliary variables $X^1, \cdots, X^m$. For unit l in small area j, let

$$(y_{jl}, X_{jl}) \;=\; (y_{jl}, x^1_{jl}, \cdots, x^m_{jl}), \quad l = 1 \cdots, N_j.$$

For now let us consider the situation where all the auxiliary variables can only take on a finite number of values. Let $(z_{jl}, x^1_{jl}, \cdots, x^m_{jl}),\ l = 1, \cdots, n_j$ be the observed y value along with auxiliary values in the sample from the jth small area for $j = 1, \cdots, M$. We pool the M samples together and let $\{b_i, i = 1, \cdots, k\}$ be all the distinct vectors which appeared in the pooled sample where $b_i = (b_{i0}, b_{i1}, \cdots, b_{im})$ for $i = 1, \cdots, k$. Notice here that the notation b_i is recycled since it plays the same role in what follows as it does

in the previous sections. Let λ_{ji} denote the proportion of (y_{jl}, X_{jl})'s in jth small area population which are equal to b_i, then $\sum_{i=1}^{k} \lambda_{ji} = 1$.

Let $\lambda_j = (\lambda_{ji}, \cdots, \lambda_{jk})$. Just as before we assume that

$$y_{jl}|\lambda_j \underset{appr}{\overset{iid}{\sim}} \text{Multinomial}(1, \lambda_j), \qquad l = 1, 2, \cdots, N_j.$$

That means $\Pr\big((y_{jl}, X_{jl}) = b_i|\lambda_j\big) = \lambda_{ji}$ where $b_i = (b_{i0}, b_{i1}, \cdots, b_{im})$ for $i = 1, \cdots, k$.

If we assign a prior of Dirichlet$(\epsilon, \cdots, \epsilon)$ to λ_j, then the posterior distribution of λ_j is Dirichlet$(n_{j1} + \epsilon, \ \cdots, \ n_{jk} + \epsilon)$ based on the similar argument in the previous section. Thus the estimator of jth small area population mean is an explicit function of the posterior expectation of λ_j, i.e.

$$\begin{aligned} \hat{\mu}^j(z^1, z^2, \cdots, z^M) &= E(\textstyle\sum_{i=1}^{k} b_{i0}\lambda_{ji}|z^1, z^2, \cdots, z^M)/N_j \\ &= \textstyle\sum_{i=1}^{k} b_{i0}\dfrac{n_{ji} + \epsilon}{n_j + k\epsilon} \end{aligned} .$$

where n_{ji} is the number of (z_{jl}, X_{jl})'s which are equal to b_i and $\sum_{i=1}^{k} n_{ji} = n_j$.

In the case with auxiliary variables, both k and n_{ji} are determined by not only y but also the x's. Even though the posterior estimator is still the weighted sum of the observed y values, the weights are different from the ones for the case without auxiliary variables.

We have seen that the posterior estimator of the population mean can be calculated explicitly through the formula above. However this is the only estimator of interest, based on the posterior which can be found in closed form. When estimating the median or when interval estimates of the median or mean are desired, one needs to simulate the posterior to find the estimators approximately. If we mark the observed vector b_i's on some balls and put all the balls in urns, then a sampling method similar to that described in subsection 3.2.4 will still be able to simulate a complete copy of each small area population corresponding to the posterior distribution of λ_j defined in this section. For each copy we are able to derive the statistics of interest, and the

average of the statistics from a variety of copies would be an approximate estimator of the corresponding parameter. Note however that this does not incorporate population information about the auxiliary variables.

3.4.1 Auxiliary Information within Small Areas

In many problems we have some extra information about auxiliary variables for each small area. A very common case is when the population mean of an auxiliary variable is known. In this situation either the ratio or the regression estimator is often used when estimating the population mean. However the regression approach provides less accuracy due to small sample size of each small area.

A Bayesian estimate for the population mean or point and interval estimates of other population quantities under our Dirichlet posterior can be approximated by simulating independent completed copies of the entire population. However, given extra auxiliary information, the direct sampling method described in subsection 3.2.4 will fail. For example, let us consider two small areas whose elements y_{jl}, for l in $\{1, \cdots, 10\}$ and j in $\{1, 2\}$ are associated with the auxiliary elements x_{jl} for l in $\{1, \cdots, 10\}$. Assume that the population mean of the x_{1l}'s is known to be 1.75 for small area 1. Suppose that we take a simple random sample of size 1 from each small area and units l_1 and l_2 are chosen respectively. Suppose that we observe $(y_{1l_1}, x_{1l_1}) = (0, 1)$ and $(y_{2l_2}, x_{2l_2}) = (1, 2)$. If we use the sampling scheme in subsection 3.2.4 to generate the copies of the entire population for small area 1, then the possible proportions λ_{11}, λ_{12} of the two pairs $(0, 1)$, $(1, 2)$ in the simulated completed copy of small area 1 must belong to the set $\{(0, 1), (.1, .9), (.2, .8), (.3, .7), \cdots, (.9, .1), (1, 0)\}$. Given any pair of λ_{11} and λ_{12}, the simulated population mean of variable x of small area 1 is determined to be $\lambda_{11}x_{1l_1} + \lambda_{12}x_{2l_2} = \lambda_{11} \cdot 1 + \lambda_{12} \cdot 2$. Since we know that the population mean of the auxiliary variable x of small area 1 is 1.75, it would make sense to require that the λ_{11} and λ_{12} satisfy the constraint: $\lambda_{11} + 2\lambda_{12} = 1.75$ which is not possible in this simple case. However when the sample size is small compared to the population size we will

see next such prior auxiliary information can be used in small area problems by making some adjustments to the posterior distribution.

The key of estimating the population mean in our Bayesian framework lies in the posterior expectation of the proportion parameters. This expectation $E(\lambda_{ji}|z^1, \cdots, z^M)$ is with respect to the Dirichlet distribution when no additional auxiliary information is present and it has an explicit expression $(n_{ji} + \epsilon)/(n_j + k\epsilon)$. Given further auxiliary information, rather than considering all λ_{ji}'s under the Dirichlet posterior and take expectations over them, we prefer to employ only those λ_{ji}'s that are consistent with the auxiliary information and take expectations over these restricted λ_{ji}'s only. For example, if the population median of the first auxiliary variable for each small area is known and we denote it by c_j for $j = 1, \cdots, M$. When we calculate the posterior expectations of λ_{ji}'s, it would be appropriate to take the expectation restricted to those λ_{ji}'s satisfying the following constraint

$$\sum_{i=1}^{k} \lambda_{ji}\, I(b_{i1}) = 0.5 \tag{3.8}$$

where

$$I(x) = \begin{cases} 1 & \text{if } x \leq c_j \\ 0 & \text{if } x > c_j \end{cases}$$

This is the constrained Dirichlet posterior based for the λ_j's for $j = 1, \cdots, M$ and its domain is restricted to a subset of the original simplex

$$\Lambda_j = \{(\lambda_{j1}, \cdots, \lambda_{jk}) | \sum_{i=1}^{k} \lambda_{ji} = 1,\ \lambda_{ji} \geq 0,\ \forall i \in \{1, \cdots, k\}\}.$$

More generally, if the qth quantile is known instead of the median, then we just need to replace 0.5 by $q/100$ in formula (3.8).

Sometime other auxiliary information might be available, say the population mean of the rth auxiliary variable in small area j belongs to a known interval $[\nu_{j1}, \nu_{j2}]$ for

$r \in \{1, \cdots, m\}$, thus the underlying constraint on λ_j becomes

$$\nu_{j1} \leq \sum_{i=1}^{k} \lambda_{ji} b_{ir} \leq \nu_{j2} \tag{3.9}$$

Another case could be that both the population mean μ_j^r and variance σ_j^2 are known for the rth auxiliary variable. Given a simulated complete copy, there are $N_j \times \lambda_{ji}$ many units taking value $(b_{i0}, b_{i1}, \cdots, b_{im})$ for $i = 1, 2, \cdots, k$. Therefore a good guess for the population variance of X^r is $\frac{1}{N_j} \sum_{i=1}^{k} N_j \lambda_{ji} (b_{ir} - \mu_j^r)^2$. After rewriting this formula and combining with the auxiliary information on the population mean X^r, our approach works as a Dirichlet distribution of λ_j restricted to the following constraints

$$\begin{cases} \sum_{i=1}^{k} \lambda_{ji}\, b_{ir} & = \mu_j^r \\ \sum_{i=1}^{k} \lambda_{ji}\, (b_{ir} - \mu_j^r)^2 & = \sigma_j^2 \end{cases} \tag{3.10}$$

More generally, given a family of population constraints based on prior auxiliary information and a sample we will be able to represent the corresponding constraints by two systems of equations

$$\begin{cases} A_{j1}\, \lambda_j & = & c_{j1} \\ A_{j2}\, \lambda_j & \leq & c_{j2} \end{cases} \tag{3.11}$$

where A_{j1}, and A_{j2} are $m_1 \times k$ and $m_2 \times k$ matrices and c_{j1} and c_{j1} are vectors of the appropriate dimensions. These linear constraints determine a convex subset of the simplex Λ_j, on which the posterior distribution of λ_j is a restricted Dirichlet distribution with a parameter equal to $(n_{j1} + \epsilon, \cdots, n_{jk} + \epsilon)$.

So far we have shown that the prior auxiliary information about the population for each small area can be expressed through a set of weighted linear equality and inequality constraints on the proportions of a simulated population taking individual distinct values from the pooled samples. We can simulate completed copies of the population consistent with such constraints and the average of the statistics of interest

based over these simulated completed copies is an estimate of the area population mean.

3.4.2 Auxiliary Information across Small Areas

As we have discussed, the following linear equality and inequality constraints on λ_j can be introduced by the auxiliary information present in the jth small area:

$$\begin{cases} A_{j1}\,\lambda_j &= c_{j1} \\ A_{j2}\,\lambda_j &\leq c_{j2} \end{cases} \tag{3.12}$$

where A_{j1}, A_{j2} are matrices that have k columns, c_{ji}'s are vectors whose dimensions are the same as the number of rows of A_{ji}, for $i = 1, 2$, and the sum of the numbers of rows of A_{j1} and A_{j2} is less than k for all small areas, i.e. for all j in $\{1, \cdots, M\}$.

However, when combined auxiliary information in the high-level sampling unit are present, inference only based on the parameter λ_j seems to be limited. For example, sometimes we might know the mean value of the rth auxiliary variable for the entire population is μ^r, but have no idea about the mean of the rth auxiliary variable with respect to each individual small area. In such a case, none of constraints above can utilize such information for a fixed j. This stimulates us to consider all small areas simultaneously as needed. That means, we are interested in all the λ_j's simultaneously instead of considering restrictions for each λ_j separately.

Let

$$\lambda = (\lambda_1, \cdots, \lambda_M) = (\lambda_{11}, \cdots, \lambda_{1k}, \cdots, \lambda_{M1}, \cdots, \lambda_{Mk})$$

be the joint parameter vector. Thus all the linear constraints regarding the λ_j's expressed through formula (3.12) can be replaced by the following linear constraints for the joint parameter vector λ:

$$\begin{cases} A_1\lambda = c_1 \\ A_2\lambda \leq c_2 \end{cases}$$

where A_1 and A_2 are block partitioned matrices with the following format

$$A_1 = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{21} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{M1} \end{pmatrix}$$

and

$$A_2 = \begin{pmatrix} A_{12} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{M2} \end{pmatrix},$$

in which A_{ji} represents the constraints introduced by the jth individual small area auxiliary information and defined in the formula (3.12). In addition, $c_1 = (c_{11}^T, \cdots, c_{M1}^T)^T$ and $c_2 = (c_{12}^T, \cdots, c_{M2}^T)^T$.

The benefit of considering the joint parameter λ is that besides the individual small area auxiliary information, some joint auxiliary information in the high-level sampling unit can be also reflected by linear constraints in terms of λ. For example, suppose we know the mean value of the rth auxiliary variable for all M small area populations is μ^r. The mean value of the rth auxiliary variable for jth small area can be evaluated by $\sum_{i=1}^k b_{ir}\lambda_{ji}$, so the population total of the rth auxiliary variable for jth small area would be $(\sum_{i=1}^k b_{ir}\lambda_{ji}) \cdot N_j$. Therefore the mean value of the rth auxiliary variable for all M small area populations can be evaluated by $(\sum_{j=1}^M (\sum_{i=1}^k b_{ir}\lambda_{ji}) \cdot N_j)/N$, where $N = N_1 + \cdots + N_M$ is the sum of M many small area population sizes. The joint auxiliary information given in the example can be expressed by the following linear constraint regarding λ

$$\left(b_{1r}\frac{N_1}{N}, \cdots, b_{kr}\frac{N_1}{N}, \cdots, \cdots, b_{1r}\frac{N_M}{N}, \cdots, b_{kr}\frac{N_M}{N} \right) \lambda = \mu^r. \tag{3.13}$$

Sometimes besides the population mean μ^r, the population variance σ_r^2 of the rth variable for high-level sampling unit is also known. The estimated variance for the overall population based on λ is $\sum_{j=1}^{M}\sum_{i=1}^{k} N_j\lambda_{ji}(b_{ir}-\mu_r)^2/(N-1)$. A good estimator should concur with the prior auxiliary information on variance as well. Thus in addition to satisfying the constraint in (3.18), λ has to satisfy the following linear constraint

$$\left(\frac{(b_{1r}-\mu_r)^2N_1}{N-1},\cdots,\frac{(b_{kr}-\mu_r)^2N_1}{N-1},\cdots,\cdots,\frac{(b_{1r}-\mu_r)^2N_M}{N-1},\cdots,\frac{(b_{kr}-\mu_r)^2N_M}{N-1}\right)\lambda=\sigma_r^2. \tag{3.14}$$

More generally speaking, joint auxiliary information in the high-level sampling unit can give rise to linear equalities and inequalities involving the proportion vector λ of the form

$$\begin{cases} D_1\lambda = d_1 \\ D_2\lambda \leq d_2 \end{cases}$$

where D_1 and D_2 are matrices with $M\times k$ columns and d_1 and d_2 are of the appropriate dimensions. All the elements of D_1, D_2, d_1 and d_2 are determined by the prior joint auxiliary information and the observed data. Here D_1 and D_2 are not block partitioned matrices anymore compared with A_1 and A_2. Inference based on λ make it possible to utilize both within area auxiliary information and across area auxiliary information simultaneously.

3.5 Simultaneous Small Area Estimation Using Auxiliary Variables

3.5.1 A Sampling Plan

Suppose we want to estimate $\mu^j = \mu(\lambda_j) = \sum_{i=1}^{k} b_{i0}\lambda_{ji}$, the population mean of the jth small area when auxiliary information is available. In terms of λ it is equivalent to

estimate

$$\mu_j(\lambda) \stackrel{def.}{=} \mu(\lambda_j) = \sum_{i=1}^{k} b_{i0}\lambda_{ji} \tag{3.15}$$

for $j = 1, \cdots, M$, when it is known that

$$\begin{cases} A_1\lambda = c_1 \\ A_2\lambda \leq c_2 \\ D_1\lambda = d_1 \\ D_2\lambda \leq d_2 \end{cases} \tag{3.16}$$

Note that $\lambda = (\lambda_1, \cdots, \lambda_M) = (\lambda_{11}, \cdots, \lambda_{1k}, \cdots, \lambda_{M1}, \cdots, \lambda_{Mk}) \in \mathbb{R}^{Mk}$, where $\sum_{i=1}^{k} \lambda_{ji} = 1$ and $\lambda_{ji} \geq 0$ for i in $\{1 \cdots, k\}$ and j in $\{1 \cdots, M\}$. Let Λ denote the parameter space restricted to the constraints above. To estimate $\mu_j(\lambda)$'s, we just need to compute the posterior expectation of λ restricted to the constraints on λ given the data. Since λ is the collection of proportion parameters of all small areas, the idea here is to find estimates for related small areas simultaneously.

In practice samples collected will almost always satisfy the constraints. When running simulations however there can be a slight chance that one can see such a "bad" sample where it is not possible to simulate complete copies which satisfy all the constraints. In order to avoid this problem we consider the following modification of simple random sampling as our sampling plan when doing simulation studies.

Sampling plan

We first take a simple random sample from each small area, such that the sample size is n_j for the jth small area and observe all x_{jl}^r's in each sample for all j in $1, \cdots, M$. Denote all distinct vectors in the pooled sample by b_i's, where $b_i = (b_{i0}, \cdots, b_{im})$. With the auxiliary information and b_i's, we are able to determine A_1, A_2, D_1 and D_2 as well as c_1, c_2, d_1 and d_2 as our constraint coefficients.

Let $\lambda^s = (\lambda_1^s, \cdots, \lambda_M^s)$, where $\lambda_j^s = (\lambda_{j1}^s, \cdots, \lambda_{jk}^s)$. If the set

$$\Lambda^s := \left\{ \lambda^s \;\middle|\; A_1\lambda^s = c_1, A_2\lambda^s \leq c_2, D_1\lambda^s = d_1, D_2\lambda^s \leq d_2, \sum_{i=1}^{k} \lambda_{ji}^s = 1, \lambda_{ji}^s \geq 0 \right\} \quad (3.17)$$

is empty then we discard the pooled sample, otherwise we keep the pooled sample. In this way we keep only samples that are consistent with the constraints including individual area-specific constraints and combined high level sampling unit constraints. Once we get a sample for which the constraints are satisfied we observe the corresponding y_{jl}'s values and compute our estimates.

Let Z_{ji} be the number of (y_{jl}, X_{jl})'s in the sample that equal b_i for i in $1, \cdots, k$ and j in $1, \cdots, M$, then $Z_j = (Z_{j1}, \cdots, Z_{jk})$ follows a $Multinomial$ $(n_j, \lambda_{j1}, \cdots, \lambda_{jk})$ distribution with density $f_j(z_j|\lambda_j) \propto \lambda_{j1}^{z_{j1}} \cdots \lambda_{jk}^{z_{jk}}$, where the notation z_{ji} has been released from its previous usage. Let $\mathcal{Z}$ be the sample space of the joint counts $(Z_{11}, \cdots, Z_{1k}, \cdots, Z_{M1}, \cdots, Z_{Mk})$'s. Then under the assumptions of independencies across small areas the distribution of the joint counts, $f_\Lambda(z|\lambda)$, is

$$f_\Lambda(z|\lambda) = \prod_{j=1}^{M} f_j(z_j|\lambda_j)$$

over Λ for $z = (z_1, \cdots, z_M) \in \mathcal{Z}$.

This assumes that our small area sample sizes are small compared to the population sizes or the sampling was done with replacement.

3.5.2 A Stepwise Bayesian Approach

In this section we will demonstrate the admissibility of a stepwise Bayes estimator that incorporates prior information about the population through linear constraints involving auxiliary variables. Dealing with constraints, when proving admissibility for finite populations, introduces some technical issues which are difficult to handle. For this reason, we will assume that we are sampling from an infinite population when proving

admissibility. We will also be assuming an infinite population when computing our procedures approximately through simulations. We need to find a sequence of disjoint prior distributions such that our estimator is the Bayes estimator against each prior. Then the stepwise Bayes theory (Ghosh and Meeden (1997) and Lazar et al. (2008)) ensures that this estimator is admissible. To obtain this sequence of priors we begin by constructing an appropriate partition of the parameter space Λ.

If y_{jl} is the value of interest for the lth unit in the jth small area population and we assume that each y_{jl} is associated with a set of auxiliary values $X_{jl} = (x_{jl}^1, \cdots, x_{jl}^m)$. Suppose that (y_{jl}, X_{jl}) can take only a finite number of values, say b_i, for $i \in \{1, \cdots, k\}$, where $b_i = (b_{i0}, \cdots, b_{im})$. Given some auxiliary information, the parameter space Λ is a convex polytope which is the intersection of the $kM - M$ dimensional simplex $\mathcal{F} := \left\{\lambda \mid \sum_{i=1}^k \lambda_{ji} = 1, \lambda_{ji} \geq 0, \forall i \in \{1, \cdots, k\}, \forall j \in \{1, \cdots, M\}\right\}$ with the space $\mathcal{G} := \left\{\lambda \mid A_1\lambda = c_1, A_2\lambda \leq c_2, D_1\lambda = d_1, D_2\lambda \leq d_2\right\}$.

Let $\mathcal{U} := \{1, \cdots, k\}$. Thus $\mathcal{U}$ has 2^k subsets. Denote these subsets by $\mathcal{U}_0, \mathcal{U}_1, \cdots, \mathcal{U}_{2^k-1}$, where $\mathcal{U}_0 = \emptyset$ and $\mathcal{U}_{2^k-1} = \mathcal{U}$. For $g = 0, 1, \cdots, 2^k - 1$, let $|\mathcal{U}_g|$ be the number of elements in the subset $\mathcal{U}_g$, then $0 \leq |\mathcal{U}_g| \leq k$. Let

$$\mathcal{F}_{\mathcal{U}_g} := \left\{ \lambda \in \mathcal{F} \middle| \begin{array}{r} \text{such that for } i \notin \mathcal{U}_g,\ \lambda_{ji} = 0 \text{ for any } j = 1, \cdots, M; \\ \text{for } i \in \mathcal{U}_g,\ \lambda_{ji} > 0 \text{ for some } j = 1, \cdots, M \end{array} \right\}$$

$\mathcal{F}_{\mathcal{U}_g}$ is the collection of all parameters for which if $i \notin \mathcal{U}_g$ then b_i can not appear in any of the small areas while if $i \in \mathcal{U}_g$ then b_i can appear in at least one small area. If $\lambda \in \mathcal{F}_{\mathcal{U}_g}$, then $\sum_{i \in \mathcal{U}_g} \lambda_{ji} = 1$ for any j in $\{1, \cdots, M\}$.

If $g \neq g^*$, then $\mathcal{F}_{\mathcal{U}_g} \cap \mathcal{F}_{\mathcal{U}_{g^*}} = \emptyset$. For instance, $\mathcal{F}_{\mathcal{U}_0} = \mathcal{F}_\emptyset = \left\{\lambda \in \mathcal{F} | \lambda_{ji} = 0, \forall j = 1, \cdots, M,\ \forall i \in \mathcal{U} = \{1, \cdots, k\}\right\} = \emptyset$; $\mathcal{F}_{\mathcal{U}_{2^k-1}} = \mathcal{F}_\mathcal{U} = \left\{\lambda \in \mathcal{F} | \text{ for any } i \in \{1, \cdots, k\}, \exists j_i \text{ s.t. } \lambda_{j_i i} > 0\right\}$.

It is obvious that $\mathcal{F} = \cup_{g=0}^{2^k-1} \mathcal{F}_{\mathcal{U}_g}$. Since the $\mathcal{F}_{\mathcal{U}_g}$'s are disjoint, $\left\{\mathcal{F}_{\mathcal{U}_g}, g \in \{0, \cdots, 2^k - 1\}\right\}$ is a partition of $\mathcal{F}$. Let $\mathcal{G}_{\mathcal{U}_g} = \mathcal{F}_{\mathcal{U}_g} \cap \mathcal{G}$ for $g = 0, \cdots, 2^k - 1$, then $\{\mathcal{G}_{\mathcal{U}_g}, g \in$

$\{1, \cdots, 2^k - 1\}\}$ is a partition of the parameter space Λ. Note that some of $\mathcal{G}_{\mathcal{U}_g}$'s might be empty.

A coarser partition of the parameter space Λ can be obtained by considering the unions of $\mathcal{G}_{\mathcal{U}_g}$'s as follows. Let $\mathcal{F}_u = \cup_{g \in \{0,\cdots,2^k-1\} \big| |\mathcal{U}_g| = u} \mathcal{F}_{\mathcal{U}_g}$ for $u = 0, \cdots, k$. Note that if $u = 0$, then $\mathcal{F}_0 = \mathcal{F}_\emptyset = \emptyset$; if $u = k$ then $\mathcal{F}_k = \mathcal{F}_\mathcal{U}$. Therefore $\mathcal{F} = \cup_{u=1}^k \mathcal{F}_u$ and $\{\mathcal{F}_u, u \in \{1, \cdots, k\}\}$ is a coarser partition of $\mathcal{F}$ based on the modules of subsets $\mathcal{U}_g$'s. Let $\mathcal{G}_u = \mathcal{F}_u \cap \mathcal{G} = \cup_{g \in \{0,\cdots,2^k-1\} \big| |\mathcal{U}_g| = u} \mathcal{G}_{\mathcal{U}_g}$ for $u = 1, \cdots, k$, then $\{\mathcal{G}_u, u \in \{1, \cdots, k\}\}$ is a coarser partition of the parameter space Λ. Note that some of $\mathcal{G}_u$'s might be empty and each $\mathcal{G}_u$ consists of disjoint subsets with a common module value u.

Using the partition $\{\mathcal{G}_u, u \in \{1, \cdots, k\}\}$ of the parameter space Λ, we will next show it leads to a corresponding partition of the sample space $\mathcal{Z}$.

Let $1 \leq k^* \leq k$ be the smallest number of b_i's that can appear in a sample for which we will have all the constraints satisfied for some value of λ. Now the stepwise Bayes argument starts the first stage at k^* and continues next to where $k^* + 1$ values appear and so on. Suppose we are at the stage where we are considering a g^* with $|\mathcal{U}_{g^*}| = u^* > k^*$. Let $\bar{\mathcal{Z}}(\mathcal{G}_{\mathcal{U}_{g^*}})$ be the restriction of the sample space that corresponds to $\mathcal{G}_{\mathcal{U}_{g^*}}$. Then let $\mathcal{Z}(\mathcal{G}_{\mathcal{U}_{g^*}})$ be $\bar{\mathcal{Z}}(\mathcal{G}_{\mathcal{U}_{g^*}})$ with all the sample points that have been taken care of at an early stage removed. Note there can be as many as $\binom{k}{u^*}$ different cases that can be considered at this stage in any order before moving on to the next stage.

Let us first consider a particular type of subset of the parameter space, say $\mathcal{G}_{\mathcal{U}_g}$. Let $\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$ be the predefined restriction of the sample space that relates to $\mathcal{G}_{\mathcal{U}_g}$. That is $\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$ contains vectors whose components correspond to the b_i's, for only $i \in \mathcal{U}_g$, all of which appear at least once. That is, if $z = (z_{11}, \cdots, z_{Mk}) \in \mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$, then z_{ji} is the count of b_i in the jth small area sample, where $z_{ji} = 0$ for any $i \notin \mathcal{U}_g$ and for any $j \in \{1, \cdots, M\}$; at the same time for any $i \in \mathcal{U}_g$, there exists at least one $j_i \in \{1, \cdots, M\}$ such that $z_{j_i i} > 0$. In this sense, $\{\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}},\ g \in \{0, \cdots, 2^k - 1\}\}$ is a partition of the sample space $\mathcal{Z}$.

For $\lambda \in \mathcal{G}_{\mathcal{U}_g}$ let

$$q(\lambda) = \sum_{\{z \mid z \in \mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}\}} f_\Lambda(z|\lambda)$$

The probability function for the restricted space is

$$f_{\mathcal{G}_{\mathcal{U}_g}}(z|\lambda) = \frac{f_\Lambda(z|\lambda)}{q(\lambda)}$$

The prior we take on $\mathcal{G}_{\mathcal{U}_g}$ is

$$\pi_{\mathcal{G}_{\mathcal{U}_g}}(\lambda) \propto \frac{q(\lambda)}{\prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{1-\epsilon}} \tag{3.18}$$

which can be made into a proper prior by using the right normalizing constant. Such a constant exists since

$$\begin{aligned}
\frac{q(\lambda)}{\prod\limits_{j=1}^{M} \prod\limits_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{1-\epsilon}} &= \frac{\sum\limits_{\{z|z\in\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}\}} f_\Lambda(z|\lambda)}{\prod\limits_{j=1}^{M} \prod\limits_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{1-\epsilon}} \\
&= \frac{\sum\limits_{\{z|z\in\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}\}} \prod\limits_{j=1}^{M} \prod\limits_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}}}{\prod\limits_{j=1}^{M} \prod\limits_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{1-\epsilon}} \\
&= \sum_{\{z|z\in\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}\}} \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}+\epsilon-1}
\end{aligned}$$

and the integral of each term in the sum on the right hand side can be controlled as

follows

$$\begin{aligned}
\int_{\mathcal{G}_{\mathcal{U}_g}} \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}+\epsilon-1} \mathrm{d}\lambda &\leq \int_{\mathcal{F}_{\mathcal{U}_g}} \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}+\epsilon-1} \mathrm{d}\lambda \\
&= \int_{\substack{\lambda_{ji}>0 \\ \forall i\in\mathcal{U}_g,\forall j}} \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}+\epsilon-1} \mathrm{d}\lambda \\
&= \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \frac{\Gamma(z_{ji}+\epsilon)}{\Gamma(n_j+|\mathcal{U}_g|\epsilon)}
\end{aligned}$$

This means $\int_{\mathcal{G}_{\mathcal{U}_g}} \frac{q(\lambda)}{\prod_{j=1}^{M}\prod_{\{i|i\in\mathcal{U}_g\}}\lambda_{ji}^{1-\epsilon}} \mathrm{d}\lambda \leq \sum_{\{z|z\in\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}\}} \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \frac{\Gamma(z_{ji}+\epsilon)}{\Gamma(n_j+|\mathcal{U}_g|\epsilon)} = \text{Constant}$, where $\sum_{i\in\mathcal{U}_g} z_{ji} = n_j$ for $j = 1, \cdots, M$ and $z_{ji} = 0$ for $\forall i \notin \mathcal{U}_g$, so $\pi_{\mathcal{G}_{\mathcal{U}_g}}(\lambda)$ is a proper prior.

Thus, by choosing the prior in this way the posterior becomes

$$\begin{aligned}
f_{\mathcal{G}_{\mathcal{U}_g}}(\lambda|z) &\propto f_{\mathcal{G}_{\mathcal{U}_g}}(z|\lambda)\pi_{\mathcal{G}_{\mathcal{U}_g}}(\lambda) \\
&\propto \prod_{j=1}^{M} \prod_{\{i|i\in\mathcal{U}_g\}} \lambda_{ji}^{z_{ji}+\epsilon-1}
\end{aligned} \tag{3.19}$$

which represents the product of multiple Dirichlet density kernels restricted to $\mathcal{G}_{\mathcal{U}_g}$. We call this posterior distribution a generalized constrained Dirichlet posterior (GCDP) given that $\sum_j \sum_i \lambda_{ji} = M$, $\sum_i \lambda_{ji} = 1$ for $\forall j = 1, \cdots, M$ and λ_{ji}'s have to satisfy the constraints in (3.16). Therefore the Bayes estimator of $\mu_j(\lambda)$, where λ belongs to $\mathcal{G}_{\mathcal{U}_g}$ against $\pi_{\mathcal{G}_{\mathcal{U}_g}}$ is $\delta_{\pi_{\mathcal{G}_{\mathcal{U}_g}}}(z) = E(\mu_j(\lambda) \mid z)$ for all z in $\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$.

So far we have considered the sample points belonging to $\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$ and the resulting Bayes estimator restricted on this subset is found. Similarly we can consider other sample points for any other selection of $g \in \{0, \cdots, 2^k - 1\}$. Since $\mathcal{Z}_{\mathcal{G}_{\mathcal{U}_g}}$'s are disjoint to each other according to our previous definition, the estimator is uniquely determined for any given sample point in terms of the way that we have described.

To generalize the whole procedure, next we will take care of all the subsets of the

parameter space Λ, such that these subsets belong to $\mathcal{G}_u$ for a given $u \in \{1, \cdots, k\}$. Let F be such a subset and Λ^F be the restriction of the parameter space Λ to F. Let $\mathcal{Z}_{\Lambda^F}$ be the restriction Λ of the sample space $\mathcal{Z}$ determined by Λ^F.

For $u = 1$, if there is at least a nonempty set F in $\mathcal{G}_1$ then F includes only one kM dimensional vector, say

$$F = (\underbrace{0, \cdots, \underbrace{1}_{\text{ith position}}, \cdots, 0}_{\text{length of k}}, \underbrace{0, \cdots, \underbrace{1}_{\text{ith position}}, \cdots, 0}_{\text{length of k}}, \cdots\cdots, \underbrace{0, \cdots, \underbrace{1}_{\text{ith position}}, \cdots, 0}_{\text{length of k}})$$

for some i in $\{1, \cdots, k\}$ and $\mathcal{Z}_{\Lambda^F}$ is just one vector in the sample space $\mathcal{Z}$. In this case we consider the prior π_{Λ^F} which assigns mass one to this parameter vector F. Therefore the posterior puts mass one on F as well. If $F = (\lambda^*_{11}, \cdots, \lambda^*_{kM})$, then the Bayes estimator is $\delta_{\pi_{\Lambda^F}}(z) = E(\mu_j(\lambda) \mid z) = \mu_j(\lambda^*)$.

If there is at least one non-empty set F in $\mathcal{G}_u$ for $u \geq 1$ then we have the following two cases:

Case 1. If the dimension of F is zero, i.e. F consists of one vector, say λ^0, then we take the prior that puts mass one on the vector. The posterior puts mass one on it as well and $\delta_{\pi_{\Lambda^F}}(z) = E(\mu_j(\lambda) \mid z) = \mu_j(\lambda^0)$.

Case 2. If the dimension of F is greater than zero then the distribution of $(Z_{11}, \cdots, Z_{Mk})$ restricted to $\mathcal{Z}_{\Lambda^F}$ is

$$f_{\Lambda^F}(z|\lambda) = \frac{f_\Lambda(z|\lambda)}{\sum_{z \in \mathcal{Z}_{\Lambda^F}} f_\Lambda(z|\lambda)}$$

Since $F \in \mathcal{G}_u$, by the previous definition there exists $\mathcal{U}_g$ for some $g \in \{0, 1, \cdots, 2^k - 1\}$, where $\mathcal{U}_g$ is a subset of $\mathcal{U}$ such that $|\mathcal{U}_g| = u$ and if $\lambda \in \Lambda^F$, then $\lambda \in \mathcal{G}_{\mathcal{U}_g}$. The prior we consider on Λ^F is

$$\pi_{\Lambda^F}(\lambda) \propto \frac{\sum_{z \in \mathcal{Z}_{\Lambda^F}} f_\Lambda(z|\lambda)}{\prod_{j=1}^{M} \prod_{\{i | i \in \mathcal{U}_g\}} \lambda_{ji}^{1-\epsilon}}$$

which can be normalized so that it can be a proper prior. Thus, by choosing the prior in this way the posterior becomes

$$\begin{aligned} f_{\Lambda^F}(\lambda|z) &\propto f_{\Lambda^F}(z|\lambda)\pi_{\Lambda^F}(\lambda) \\ &\propto \prod_{j=1}^{M}\prod_{\{i|i\in\mathcal{U}_g\}}\lambda_{ji}^{z_{ji}+\epsilon-1} \end{aligned} \tag{3.20}$$

which represents the generalized constrained Dirichlet posterior(GCDP) restricted to Λ^F. The posterior distribution is the GCDP restricted to Λ^F. The Bayes estimator of $\mu_j(\lambda)$, where λ belongs to Λ^F against π_{Λ^F} is $\delta_{\pi_{\Lambda^F}}(z) = E(\mu_j(\lambda) \mid z)$ for all z in $\mathcal{Z}_{\Lambda^F}$. Hence if we use the sequence priors, ignoring the empty sets at each step,

$$\left\{\left\{\pi_{\Lambda^F}\Big|_{F\in\,\mathcal{G}_1}\right\},\ \left\{\pi_{\Lambda^F}\Big|_{F\in\,\mathcal{G}_2}\right\},\ \cdots,\left\{\pi_{\Lambda^F}\Big|_{F\in\,\mathcal{G}_\gamma}\right\}\right\},$$

then an estimator of $\mu_j(\lambda)$ denoted by $\delta(z)$ can be defined by

$$\delta(z) = \begin{cases} \delta_{\pi_{\Lambda^F}}(z) & \text{if } z \in \mathcal{Z}_{\Lambda^F} \text{ for all } F \in \mathcal{G}_1 \\ \delta_{\pi_{\Lambda^F}}(z) & \text{if } z \in \mathcal{Z}_{\Lambda^F} \text{ for all } F \in \mathcal{G}_2 \\ \quad\vdots & \\ \delta_{\pi_{\Lambda^F}}(z) & \text{if } z \in \mathcal{Z}_{\Lambda^F} \text{ for all } F \in \mathcal{G}_\gamma \end{cases}$$

where

$$\gamma = \begin{cases} k, & \text{if } k < n = n_1 + \cdots + n_M \\ n, & \text{if } k \geq n = n_1 + \cdots + n_M \end{cases} \tag{3.21}$$

We summarize our result in the following theorem:

Theorem 1. For estimating the small area population means, under squared error loss, the estimator $\delta(z)$, is admissible, under any sampling design.

Proof. This follows just as in Lazar et al. (2008) since we have successfully constructed a sequence of priors such that $\delta(z)$ restricted on the support of each prior is a Bayes estimator of $\mu_j(\lambda)$ against this prior, i.e. $\delta(z)$ is a stepwise Bayes. □

Note that when there is no additional auxiliary information available, our estimator

is just based on the posterior expectation of λ with respect to the generalized Dirichlet distribution with the density $f_{\mathcal{F}}(\lambda|z) \propto \prod_{j=1}^{M}\prod_{i=1}^{k}\lambda_{ji}^{z_{ji}+\epsilon-1}$. In this case, we just need to replace $\mathcal{G}$ in the previous argument by $\mathbb{R}^{kM}$, then the admissibility of the estimator without constraints can be proved.

3.5.3 An Example of Step-wise Bayes Estimators

To better understand how the step-wise Bayesian estimator works, let us consider an example. Suppose we have three small areas, i.e. $M = 3$, of which the elements can take on only four values say $\{b_1, b_2, b_3, b_4\}$, where b_i could be a vector for $i = 1, 2, 3, 4$ when auxiliary variables are involved. If λ_{ji} denotes the proportion of elements equal to b_i in the jth small area population, then $\lambda \in \mathbb{R}^{12}$, where $\lambda = (\lambda_{11}, \lambda_{12}, \cdots, \lambda_{34})$. Let

$$\{\{b_1, b_4, b_4\}, \{b_1, b_1, b_1, b_2\}, \{b_2, b_2, b_4, b_4, b_4, b_4, b_4\}\}$$

be a pooled random sample obtained by our sampling plan. Then

$$z = (1, 0, 0, 2,\ 3, 1, 0, 0,\ 0, 2, 0, 5)$$

is the observed joint counts indicating the frequencies of b_i's for $i = 1, \cdots, 4$. Intuitively speaking, it is appropriate to make an inference about the posterior distribtuion that the parameter λ should have at least 6 positive components at the positions that positive counts are observed, i.e. $\lambda_{11}, \lambda_{14}, \lambda_{21}, \lambda_{22}, \lambda_{32},$ and λ_{34} should be all greater than 0. Further, we found that b_2 did not show up in the sample of small area 1, and the corresponding count is $z_{12} = 0$. Such a case happens frequently when the sample size is small. If we only look at the sample from small area 1 independently, it seems to be reasonable to put $\lambda_{12} = 0$ in our inference. However if we believe there exists similarity across small areas, we assume that $\lambda_{12} > 0$ since b_2 has been observed in the other small areas and it could contribute to small area 1 as well. Such a contribution is masked by

the small sample size. By accounting for the contribution of the foreign element b_2 to small area 1, say assigning a little positive mass on λ_{12} in the prior distribution, we are able to borrow strength across small areas. In other words, as long as b_i is observed at least once in the pooled sample, the exchangeability guarantees that the corresponding λ_{ji} should be greater than zero in all the priors for $j = 1, 2, 3$ even if b_i might not be observed for some j. In what follows we will see how our proposed method will accomplish this goal. Let $\mathcal{U}_g$ be the index set $\{1, 2, 4\}$. b_i appeared at least once in the pooled sample for $i \in \mathcal{U}_g$. $|\mathcal{U}_g| = 3$, therefore $\mathcal{G}_{\mathcal{U}_g} \in \mathcal{G}_3$ by the definition. The proposed method assigns z into $\mathcal{Z}_{\Lambda^F}$ where $F \in \mathcal{G}_3$ and $F = \mathcal{G}_{\mathcal{U}_g}$. Thus the posterior distribution of λ is

$$\begin{aligned} f_{\Lambda^F}(\lambda|z) &\propto f_{\Lambda^F}(z|\lambda)\pi_{\Lambda^F}(\lambda) \\ &\propto \prod_{j=1}^{3} \lambda_{j1}^{z_{j1}+\epsilon-1}\lambda_{j2}^{z_{j2}+\epsilon-1}\lambda_{j4}^{z_{j4}+\epsilon-1} \end{aligned}$$

restricted to Λ^F, which is the restriction of the parameter space Λ to F. The posterior distribution assumes $\lambda_{j3} = 0$ for $j = 1, 2, 3$. It can be used to generate a copy of the whole population which is completely based on the observed values in the pooled sample. Such a property is similar to the nature of the polya sampling method. It also told us that the parameter $\lambda \in \mathbb{R}^{12}$ degenerated to be

$$\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{14}, \cdots, \lambda_{31}, \lambda_{32}, \lambda_{34}) \in \mathbb{R}^9$$

for the given z, i.e k has actually decreased to 3.

In conclusion, in practice once we have obtained our sample, k is determined by the number of distinct units $(y_{jl}, x_{jl}^1, \cdots, x_{jl}^m)$, for $l = 1, \cdots, n_j$ and $j = 1, \cdots, M$, appearing in the pooled samples of all small areas. Denote these distinct units by $\{b_1, \cdots, b_k\}$ and $\lambda = (\lambda_{11}, \cdots, \lambda_{Mk})$, where λ_{ji} is the proportion of units equal to b_i for jth small area population. Let n_{ji} be the count of b_i in the sample of small area j. Then the joint

constrained posterior distribution is

$$h^*(\lambda) \propto \prod_{j=1}^{M} h_j(\lambda_j)$$

restricted to Λ where $h_j(\lambda_j) \propto \lambda_{j1}^{n_{j1}+\epsilon-1} \cdots \lambda_{jk}^{n_{jk}+\epsilon-1}$ for j in $\{1, \cdots, M\}$. Our purpose here is to find the expectation of functions of λ with respect to this joint constrained posterior distribution when joint auxiliary information is present.

3.6 Generalized Weighted Dirichlet Posterior

As we have discussed in this chapter, our final point estimator is just the weighted sum of observed y values, where the weights are the constrained posterior expectations of proportion parameters λ_{ji}'s. Strief and Meeden (2007) proposed a weighted Dirichlet posterior (WDP), which provides us an alternative perspective to analyze the problem. Let w_{ji}'s be a set of weights defined by $w_{ji} = N_j E(\lambda_{ji}) = N_j \mu_{ji}$ for i in $\{1, \cdots, k\}$, where the expectation is taken with respect to GCDP. Consider the generalized Dirichlet distribution over the simplex $\mathcal{F}$ defined by the vector $(k\mu_{11}, \cdots, k\mu_{1k}, \cdots, k\mu_{M1}, \cdots, k\mu_{Mk})$ as an alternative posterior distribution for $\lambda = (\lambda_{1,1}, \cdots, \lambda_{1k}, \lambda_{M1}, \cdots, \lambda_{Mk})$ when using the observed joint sample to generate complete simulated copies of the population. This is a generalized version of WDP, and we will call this posterior GWDP. Under the GCDP every complete simulated copy of the population will satisfy the constraints of (3.16); however, under the GWDP, only the average of all the simulated populations will satisfy the constraints. We define the GWDP estimator of jth small area population mean as

$$\hat{\mu}^j_{\text{GWDP}} \stackrel{\Delta}{=} E_{\text{GWDP}}(\sum_{i=1}^{k} b_{i0}\lambda_{ji}) = \sum_{i=1}^{k} b_{i0}\mu_{ji} = \sum_{i=1}^{k} b_{i0} E_{\text{GCDP}}(\lambda_{ji}) = \hat{\mu}^j_{\text{GCDP}} \qquad (3.22)$$

where $b_i = (b_{i0}, \cdots, b_{im})$'s are distinct sampling units (vectors if auxiliary variables exist). The right hand side of formula (3.22) is $E(\mu_j|z)$.

Thus when estimating the population mean, if we plug the weights, $k\mu_{ji}$, estimated from GCDP with constraints into GWDP without constraints, then we get the same point estimate. Another way of constructing 95% interval for the jth small area population mean is simulating from the Dirichlet$(k\mu_{j1}, \cdots, k\mu_{jk})$ model and observing the 2.5% and 97.5% quantiles, where μ_{ji} is estimated from GCDP approach. Since the closed-form variance of the Dirichlet without constraints is known, 95% intervals could alternatively be constructed through a Normal approximation:

$$\sum_{i=1}^{k} b_{i0}\mu_{ji} \pm 1.96 \sqrt{V\Big(\sum_{i=1}^{k} b_{i0}\lambda_{ji}\Big)}, \text{ and} \tag{3.23}$$

$$\begin{aligned}
V\Big(\sum_{i=1}^{k} b_{i0}\lambda_{ji}\Big) &= \sum_{i=1}^{k} b_{i0}^2 \mathrm{V}(\lambda_{ji}) + 2\sum\sum_{i<l} b_{i0}b_{l0}\mathrm{Cov}(\lambda_{ji}, \lambda_{jl}) \\
&= \sum_{i=1}^{k} b_{i0}^2 \frac{k\mu_{ji}(k - k\mu_{ji})}{k^2(k+1)} + 2\sum\sum_{i<l} b_{i0}b_{l0}\frac{-k\mu_{ji}k\mu_{jl}}{k^2(k+1)} \\
&= \sum_{i=1}^{k} b_{i0}^2 \frac{\mu_{ji}(1 - \mu_{ji})}{(k+1)} + 2\sum\sum_{i<l} b_{i0}b_{l0}\frac{-\mu_{ji}\mu_{jl}}{(k+1)} \\
&= \frac{1}{(k+1)}\Big(\sum_{i=1}^{k} \mu_{ji} b_{i0}^2 - \sum_{i=1}^{k}\sum_{l=1}^{k} b_{i0}b_{l0}\mu_{ji}\mu_{jl}\Big) \\
&= \frac{1}{(k+1)}\sum_{i=1}^{k} \mu_{ji}\left(b_{i0} - \sum_{i=1}^{k} b_{i0}\mu_{ji}\right)^2 \\
&= \frac{1}{(k+1)}\sum_{i=1}^{k} \mu_{ji}\left(b_{i0} - \hat{\mu}^j_{\mathrm{GCDP}}\right)^2 \\
&= \frac{1}{(k+1)}\,\hat{\sigma}^2_{\mathrm{GCDP}}
\end{aligned} \tag{3.24}$$

The interval estimates under the weighted Dirichlet need not to be equal to those from

the ones under GCDP. In addition, Strief and Meeden(2007) showed that the credible interval based on WDP has better coverage rate than constrained Polya posterior when the sample size is small. We will use both the GWDP and the GCDP to construct 95% credible intervals and compare the results.

Chapter 4

Computation

In chapter 3, we have discussed that given the prior information present in auxiliary variables, the proportion vector λ_j from the Dirichlet posterior distribution needs to satisfy some linear constraints. The support of the normal Dirichlet distribution is a $k-1$ dimensional bounded convex polytope. A linear constraint for λ_j corresponds to either a hyperplane or a halfspace of k-dimensional Euclidean space $\mathbb{R}^k$. Therefore the support of the constrained posterior,

$$\Lambda_j := \big\{(\lambda_{j1}, \cdots, \lambda_{jk}) \;\big|\; A_{j1}\lambda_j = c_{j1}, A_{j2}\lambda_j \leq c_{j2}, \sum_{i=1}^{k} \lambda_{ji} = 1, \lambda_{ji} \geq 0, \forall i \in \{1, \cdots, k\}\big\},$$

is also a bounded convex polytope, which is a subset of $\mathbb{R}^k$. Λ_j is not of full rank and it will have an empty interior in $\mathbb{R}^k$. If we were able to generate a sequence $\mathbf{x_1}, \cdots, \mathbf{x_T}$, of independent random variates from the Dirichlet posterior distribution over Λ_j then the Monte Carlo estimator $\frac{\sum_{i=1}^{T} \mu(\mathbf{x_i})}{T}$ would give a good approximation of the expected value of $\mu(\lambda_j)$ for T sufficiently large. However in this thesis we need to be able to find expectations of functions defined on convex polytopes that have empty interiors. It is not easy to generate independent proportion vectors λ_j's from the Dirichlet posterior distribution directly under the linear constrains, which is equivalent to generating high

dimensional points independently from a complicated polytope. When such a target density h can be evaluated but not easily sampled, Markov chain Monte Carlo (MCMC) methods can be used to generate a draw from a distribution that approximates h, thus the expectations of functions of λ_j with respect to the constrained posterior distribution can reliably be estimated. An alternative way to approximate the posterior expectation of $\mu(\lambda_j)$ is the importance sampling trick in Monte Carlo techniques. In this chapter we use the Metroplis-Hastings algorithm to construct a suitable Markov chain.

Markov chain Monte Carlo (MCMC) is a powerful simulation technique for exploring high-dimensional probability distributions. It is particularly useful for exploring posterior probability distributions that arise in Bayesian statistics. Although there are some generic tools (such as WinBugs and JAGS) for doing MCMC, for non-standard problems it is often desirable to code up MCMC algorithms from scratch.

4.1 Metroplis-Hastings Algorithm

As we have discussed in chapter 3, our proposed estimator of $\mu(\lambda_j)$ is the expected value of $\mu(\lambda_j)$ with respect to $Dirichlet(n_{j1}+\epsilon, \cdots, n_{jk}+\epsilon)$ restricted to the parameter space Λ_j for the jth small area. Let $h(\cdot)$ denote the unnormalized density function of this constrained posterior distribution from which we want to draw samples. We know that $h(\lambda_j) \propto \lambda_{j1}^{n_{j1}+\epsilon-1} \cdots \lambda_{jk}^{n_{jk}+\epsilon-1}$. Here we need to keep in mind that the kernel of $h(\cdot)$ is just the usual Dirichlet kernel, however the support of $h(\cdot)$ is not over all the λ_j's, but over some subset defined by the constraints. To calculate such a posterior expectation of $\mu(\lambda_j)$, we will resort to the Metropolis-Hastings algorithm which generates a Markov chain that converges to any target distribution given by an unnormalized density with respect to a new measure. Let $\mathbf{X}$ denote a random variable on whose state space we will simulate a Markov chain. In this thesis, it is important to remember that the random variables $\mathbf{X}^{(t)}$ simulated in a MCMC procedure are typically parameter vectors whose posterior distribution is the constrained Dirichlet distribution over Λ_j. The Metroplis-

Hastings method begins at $t = 0$ with the selection of $\mathbf{X}^{(0)} = \mathbf{x}^{(0)}$ drawn at random from some *proposal distribution* g, with the requirement that $h(\mathbf{x}^{(0)}) > 0$. Given $\mathbf{X}^{(t)} = \mathbf{x}^{(t)}$, the algorithm generates $\mathbf{X}^{(t+1)}$ as follows:

Step 1. Sample a candidate value $\mathbf{X}^*$ from a proposal distribution $g(\cdot|\mathbf{x}^{(t)})$.

Step 2. Compute the Metroplis-Hastings ratio $R(\mathbf{x}^{(t)}, \mathbf{X}^*)$, where

$$R(u, v) = \frac{h(v)g(u|v)}{h(u)g(v|u)}.$$

Note that $R(\mathbf{x}^{(t)}, \mathbf{X}^*)$ is always defined, because the proposal $\mathbf{X}^* = \mathbf{x}^*$ can only occur if $h(\mathbf{x}^{(t)}) > 0$ and $g(\mathbf{x}^*|\mathbf{x}^{(t)}) > 0$.

Step 3. Sample a value for $\mathbf{X}^{(t+1)}$ according to the following:

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\} \\ \mathbf{x}^{(t)} & \text{otherwise} \end{cases}$$

Step 4. Increment t and return to step 1.

We will call the tth iteration the process that generates $\mathbf{X}^{(t)} = \mathbf{x}^{(t)}$. It can be shown that the marginal distribution of $\mathbf{X}^{(t+1)}$ is h, and h must be the stationary distribution of the chain. This means that we would be able to generate dependent sample points from the parameter space Λ_j with respect to the constrained posterior distribution if we can find the appropriate proposal distribution g.

4.2 Hit-and-Run Algorithm

As given in the previous subsection, the Metroplis-Hastings algorithm is time-homogeneous in the sense that the proposal distribution g does not change as t increases. It is still problematic to find a proposal distribution when our parameter space is a complicated

polytope of high dimensions. Chen (1993) constructed an MCMC approach that relys on time-varying proposal distributions, $g^{(t)}(\cdot|\mathbf{x}^{(t)})$. Such a strategy that resembles a random walk chain is known as the *hit-and-run* algorithm. In this approach, the proposed move away from $\mathbf{x}^{(t)}$ is generated in two stages: by choosing a direction to move, and then a distance to move in the chosen direction. After initialization at $\mathbf{x}^{(0)}$, the chain proceeds from $t = 0$ with the following steps.

Step 1. Draw a random direction $\mathbf{d}^{(t)} \sim p(\mathbf{d})$, where p is a density defined over the surface of the unit k-sphere.

Step 2. Find the set of all real numbers a for which $\mathbf{x}^{(t)} + a\ \mathbf{d}^{(t)}$ is in the state space of $\mathbf{X}$. Denote this set of signed lengths as $A^{(t)}$.

Step 3. Draw a random signed length $a^{(t)}|(\mathbf{x}^{(t)}, \mathbf{d}^{(t)}) \sim g_a^{(t)}(a|\mathbf{x}^{(t)}, \mathbf{d}^{(t)})$, where the density $g_a^{(t)}(a|\mathbf{x}^{(t)}, \mathbf{d}^{(t)}) = g^{(t)}(\mathbf{x}^{(t)} + a\ \mathbf{d}^{(t)})$ is defined over $A^{(t)}$. The proposal distribution may differ from one iteration to the next only through a dependence on $A^{(t)}$.

Step 4. For the proposal $\mathbf{X}^* = \mathbf{x}^{(t)} + a^{(t)}\ \mathbf{d}^{(t)}$, compute the Metroplis-Hastings ratio

$$R(\mathbf{x}^{(t)}, \mathbf{X}^*) = \frac{h(\mathbf{X}^*)g^{(t)}(\mathbf{x}^{(t)})}{h(\mathbf{x}^{(t)})g^{(t)}(\mathbf{X}^*)}.$$

Step 5. Set

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\} \\ \mathbf{x}^{(t)} & \text{otherwise.} \end{cases}$$

Step 6. Increment t and go to step 1.

The direction distribution p is frequently taken to be uniform over the surface of the unit sphere. In k dimensions, a random variable may be drawn from this distribution by sampling a $k-$ dimensional standard normal variable $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I_k})$ and making the transformation $\mathbf{d} = \mathbf{Z}/\sqrt{\mathbf{Z}^T\mathbf{Z}}$.

The performance of this approach has been compared with that of other simple MCMC mehtods (Chen 1993). It has been noted that the hit-and-run algorithm can offer particular advantage when the state space of $\mathbf{X}$ is sharply constrained (Chen 1993), thereby making it difficult to explore all regions of the space effectively with other methods. The choice of p has a strong effect on the performance and convergence rate of the algorithm, with the best choice often depending on the shape of h and the geometry of the state space (including constraints and the chosen units for the coordinates of $\mathbf{X}$).

4.3 Algorithm for Non-full Dimensional Convex Polytopes

In this thesis, we are interested in finding the expectation of functions of the proportion vector λ_j over a convex polytope Λ_j for jth small area, where

$$\Lambda_j := \{(\lambda_{j1}, \cdots, \lambda_{jk}) | A_{j1}\lambda_j = c_{j1}, A_{j2}\lambda_j \leq c_{j2}, \sum_{i=1}^{k} \lambda_{ji} = 1, \lambda_{ji} \geq 0, \forall i \in \{1, \cdots, k\}\} \tag{4.1}$$

The constraint $\sum_{i=1}^{k} \lambda_{ji} = 1$ can be treated as a type of linear constraint. Therefore more generally speaking, we are interested in finding the expectation of functions over a convex polytope Λ_j that have empty interiors in the space they lie, i.e. a convex polytope represented as

$$A_1\mathbf{x} = c_{j1} \tag{4.2}$$

$$A_2\mathbf{x} \leq c_{j2} \tag{4.3}$$

$$\mathbf{x} \geq 0 \tag{4.4}$$

Let A_1 be a $m_1 \times k$ matrix, A_2 a $m_2 \times k$ matrix and c_{j1}, c_{j2} and $\mathbf{x}$ vectors of appropriate dimensions with $m_1 + m_2 \leq k$.

If h is the unnormalized kernel density function of the constrained Dirichlet posterior distribution of λ_j, the following procedure (Lazar 2005) generates a Markov chain within

a convex polytope represented by the system of equalities and inequalities in 4.2, 4.3 and 4.4 using the rejection-acceptance update of the hit-and-run algorithm.

Step 1. Choose an initial point $\mathbf{x}^{(0)}$ in Λ_j, such that at least one in 4.3 and 4.4 holds strictly, and set $t = 0$.

Step 2. Draw a random direction $\tilde{\mathbf{d}}^{(t)}$ uniformly distributed over the unit k-sphere.

Step 3. Project $\tilde{\mathbf{d}}^{(t)}$ onto the null space of A_1. Let $\mathbf{d}^{(t)}$ be the normalized projection.

Step 4. Find the set of all real numbers a for which $\mathbf{x}^{(t)} + a\ \mathbf{d}^{(t)}$ is in Λ_j. Denote this set of signed lengths as $A^{(t)}$.

Step 5. Draw a random signed length $a^{(t)}$ uniformly over $A^{(t)}$.

Step 6. For the proposal $\mathbf{X}^* = \mathbf{x}^{(t)} + a^{(t)}\ \mathbf{d}^{(t)}$, compute the Metroplis-Hastings ratio

$$R(\mathbf{x}^{(t)}, \mathbf{X}^*) = \frac{h(\mathbf{X}^*) \cdot 1}{h(\mathbf{x}^{(t)}) \cdot 1} = \frac{h(\mathbf{X}^*)}{h(\mathbf{x}^{(t)})}.$$

Step 7. Set

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\} \\ \mathbf{x}^{(t)} & \text{otherwise.} \end{cases}$$

Step 8. Increment t and go to step 2.

The initial point $\mathbf{x}^{(0)}$ that is in the interior of $\Lambda_j = \{\mathbf{x} \in \mathbb{R}^k | A_1\mathbf{x} = c_1,\ A_2\mathbf{x} \leq c_2;\ \mathbf{x} \geq 0\}$ can be found by finding the vertices of the convex polytop and then taking a convex combination of the vertices. (For details see Geyer 2008). Another way of obtaining an initial point that does not lie on the boundary is by solving a linear program problem, which involves a slack variable $\mathbf{y}$ as follows:

$$A_1\mathbf{x} = c_{j1} \tag{4.5}$$

$$A_2\mathbf{x} + \mathbf{y} = c_{j2} \tag{4.6}$$

$$\mathbf{x} \geq 0,\ \mathbf{y} \geq 0 \tag{4.7}$$

To avoid getting solutions on the boundary of the polytopes, we replace 0 in 4.7 by a positive number η. This new problem can be solved in the same way by making a change of variables.

To obtain the normalized direction $\mathbf{d}^{(t)}$ in the null space of A_1 in step 3, we firstly apply the singular value decomposition (SVD) theorem. According to the SVD theorem there exist an $m_1 \times m_1$ orthogonal matrix U and an $k \times k$ orthogonal matrix V such that

$$A_1 = U \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T$$

where matrix Σ is a $r \times r$ diagonal matrix with real numbers $\sigma_1, \cdots, \sigma_r$ on the diagonal and r is the rank of A_1. Denote $U = (U_1,\ U_2)$ and $V = (V_1,\ V_2)$, where U_1 is the first r columns of U and V_1 is the first r columns of V. The theorem shows that $A_1 = U_1 \Sigma V_1^T$. Given any random direction $\tilde{\mathbf{d}}^{(t)}$ from step 2 of the algorithm, $V_2 V_2^T \tilde{\mathbf{d}}^{(t)}$ belongs to $N(A_1)$, the null space of A_1, since $A_1 V_2 V_2^T \tilde{\mathbf{d}}^{(t)} = U_1 \Sigma V_1^T V_2 V_2^T \tilde{\mathbf{d}}^{(t)} = 0$. Therefore the normalized $\mathbf{d}^{(t)}$ in step 3 is calculated to be $V_2 V_2^T \tilde{\mathbf{d}}^{(t)}$.

When auxiliary information is available, we are interested in finding the constrained posterior expectation of functions of the joint proportion vector $\lambda = (\lambda_1,\ \cdots, \lambda_M) = (\lambda_{11}, \cdots, \lambda_{1k}, \cdots, \lambda_{M1}, \cdots, \lambda_{Mk})$ over a parameter space Λ, where

$$\Lambda = \{\lambda | A_1 \lambda = c_1;\ A_2 \lambda \leq c_2;\ D_1 \lambda = d_1;\ D_2 \lambda \leq d_2\}$$

with A_i's and D_i's defined in section 3.4. Again Λ is a convex polytope in $\mathbb{R}^{Mk}$. In this case, the corresponding unnormalized density function for the joint constrained posterior distribution is

$$h^*(\lambda) \propto \prod_{j=1}^{M} h_j(\lambda_j)$$

where $h_j(\lambda_j) \propto \lambda_{j1}^{n_{j1}+\epsilon-1} \cdots \lambda_{jk}^{n_{jk}+\epsilon-1}$ for j in $\{1, \cdots, M\}$. A Markov chain can be generated by replacing Λ_j by Λ in step 1 and replacing h by h^* in step 6 of the proposed

algorithm, and then the constrained posterior expectation of functions can be approximated by averaging the functions over the chain.

Chapter 5

Simulation Studies

In this chapter we compare the generalized constrained Dirichlet posterior(GCDP) approach and some other traditional approaches to small area estimation through simulations in R.

5.1 Populations with A Linear Relationship

In this simulation, we consider four small areas with one auxiliary variable. N_j denotes the population size for each small area, where

$$N_j \sim \text{Poisson}(\lambda = 100),\ j = 1, 2, 3, 4.$$

(y_{jl}, x_{jl}) is the lth unit of small area j. μ_j is the population mean of auxiliary variable, $x_{jl}, l = 1, \cdots, N_j$, for each small area. We assumed that

$$\mu_j \sim N(\mu = 150, \sigma = 2),\ j = 1, 2, 3, 4.$$

Let σ_j denote the population standard deviation of the auxiliary variable , i.e. $x_{jl} \sim N(\mu_j, \sigma_j)$. For each small area, the units of interest, y_{jl}'s are generated through the following model:

Model 1: $y_{jl} = 2 * x_{jl} + \epsilon_{jl}, \quad$ where $\sigma_j = 1,\ \epsilon_{jl} \sim N(0, \sigma = 1)$

We generated 500 random samples each of size 5 ($n_j = 5$) from each population according to our sampling plan explained in section 3.5, that is we only kept samples that satisfied the constraints. In all cases we used the same random numbers to generate the populations. For each sample we computed the sample mean (direct estimator), the regression estimator, EBLUP estimator and generalized constrained Dirichlet posterior (GCDP) estimator as well as their corresponding 95% confidence intervals/credible intervals given that the population mean of the auxiliary variable is known. The GCDP estimator is calculated with a hyperparameter $\epsilon = 1$ in the prior distribution.

20 pairs of (y_{jl}, x_{jl})'s are observed in each joint sample from 4 small areas under the settings above. If we denote distinct sampling vectors by $b_i = (b_{i0}, b_{i1})$, $i = 1, \cdots, k$, then the number of distinct sampling pairs is 20 due to the continuous nature of the populations. Therefore the dimension of proportion parameter for each small area is 20 for the GCDP approach, i.e. $k = 20$ as we have discussed in the section 3.5. Thus the parameter space of GCDP's posterior distribution is $\Lambda \subseteq \mathcal{R}^{80}$. Let z_{ji} be the number of (y_{jl}, X_{jl})'s in the jth small area sample that equal b_i. Note that each z_{ji} equals either 0 or 1.

$$\lambda = (\lambda_{1,1}, \cdots, \lambda_{1,20}, \ldots, \lambda_{4,1}, \cdots, \lambda_{4,20})$$

Given a sample the posterior for λ is proportional to $f(\lambda | z_{1,1}, \cdots, z_{4,20})$, where $f(\cdot)$ is defined in formula (3.20).

Markov chains of length 8,000,000 were generated and the estimate were computed based on 7,200,000 points after discarding the first 800,000 sampled values as "burn in". 95% credible intervals for the GCDP method was obtained in the following way. We picked every 2000th point in the Markov chain after the burn-in period, and so 3600 points were selected. For each such point or estimated weight vector $\hat{\lambda}$, we calculated the estimated population means based on formula (3.15). Then the 2.5% quantile and

Small Area	N_j	Corr (y, x)	$\mu_Y(\text{Var}(Y))$	$\mu_X(\text{Var}(X))$
1	105	0.889	309.346(4.399)	154.647(0.758)
2	107	0.888	295.961(4.909)	148.017(0.830)
3	98	0.857	287.690(4.316)	143.842(0.789)
4	95	0.880	299.846(4.735)	149.860(0.964)
Total	405	0.992	298.341(65.582)	149.158(15.972)

Table 5.1: Population statistics of simulated small areas for Model 1

97.5% quantile of these 3600 estimates were chosen to be the endpoints of a 95% credible interval.

In chapter 3, we introduced the weighted Dirichlet posterior (WDP) proposed by Strief(2007), which is a looser version of the constrained Polya posterior (CPP). The interval estimates under the weighted Dirichlet will be longer than those from the CPP even though the point estimates are the same. In addition, Strief(2007) showed that credible intervals based on the WDP can have a better coverage rate than the CPP when the sample size is small. We will use the WDP version of our GCDP approach to construct a second 95% credible interval. (For details, see chapter 3).

The simulated population information of model 1 is summarized in the Table 5.1. The first column represents the label of the small area; the second column represents the population size of each small area; the third column corresponds to the correlation between the simulated variables in each small area; the fourth column represents the population mean of the variable of interest for each small area with its population variance and the fifth column represents the population mean of the auxiliary variable for each small area with its population variance.

The results based on the different approaches are presented in table 5.2. The first column of the table represents the label of small area. The second column of the table represents the method used to estimate the small area population mean. The third column is the average of the 500 estimates. The fourth column is the average of the absolute differences between the estimate and the true value of the population mean.

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
1	Direct	309.396	0.730	307.681	3.430	0.886
2		295.935	0.718	294.135	3.599	0.878
3		287.620	0.681	285.903	3.434	0.924
4		299.823	0.770	298.047	3.550	0.882
1	Regression	309.333	0.369	308.541	1.585	0.858
2		295.865	0.423	295.014	1.701	0.812
3		287.665	0.406	286.810	1.709	0.840
4		299.915	0.446	299.117	1.597	0.780
1	EBLUP	309.453	0.664	307.381	4.143	0.960
2		296.018	0.472	294.384	3.268	0.956
3		287.597	0.605	285.569	4.055	0.982
4		299.793	0.480	298.173	3.240	0.950
1	GCDP	309.436	0.417	309.152	0.547	0.392
2	(2000th)	296.012	0.197	295.641	0.745	0.844
3	($\epsilon = 1$)	287.640	0.485	287.351	0.565	0.356
4		299.770	0.214	299.405	0.724	0.810
1	GCDP	309.435	0.416	307.461	3.948	0.986
2	-WDP	296.011	0.195	293.003	6.016	1.000
3	($\epsilon = 1$)	287.641	0.489	285.682	3.916	0.968
4		299.769	0.215	296.680	6.178	1.000

Table 5.2: Comparison of small area estimates from different approaches for the population of Model 1

The fifth column is the average of the lower bounds of the 95% confidence/credible intervals. The sixth column represents the average length of the 95% confidence/credible intervals and the seventh column represents the average of number of times in which the true value of the population was contained in the 95% confidence/credible intervals.

In order to assess the quality of each estimator, we have considered the Average Empirical Mean Square Error (AEMSE), which is obtained by taking the mean value of all the EMSEs obtained for the 4 areas:

$$AEMSE = \frac{1}{4}\sum_{j=1}^{4}(\hat{\mu}^j - \mu^j)^2$$

where $\hat{\mu}^j$ is an estimate of the true area mean μ^j. For each sample we calculated AEMSE, then took the average of them to assess the overall performance of each method. The average of AEMSE for the direct estimator is 0.8238. The average of AEMSE for the regression estimator is 0.2835. The average of AEMSE for the EBLUP estimator is 0.5056. The average of AEMSE for the GCDP-WDP estimator is 0.2035. The average of AEMSE for the GCDP estimator based on 3600 sample points at every 2000th position in the Markov chain is 0.2028. Since the true model is the regression model, it is not a surprise that the EBLUP estimator is worse than the regression estimator, but the EBLUP estimator does beat the regression estimator as an interval estimator. The GCDP-WDP estimator is also better than the regression estimator with respect to AEMSE because the regression estimator is usually not good when the sample size is as small as 5. The correlation between the variables is strong, so the direct estimator without taking advantage of the auxiliary variable and borrowing strength performs the worst for both point and interval estimation. The GCDP-WDP performs best among these approaches regarding to AEMSE.

The GCDP(2000th) interval estimator based on the 2.5% and 97.5% quantiles of 3600 points selected at every 2000th position of the Markov chain has lower coverage rate than the GCDP-WDP interval estimator described in the formula (3.23).

For the GCDP-WDP estimators the average of absolute error of small area 3 is the largest and the average of absolute error of small area 1 is the second largest among the four areas. This is not a surprise at all since the true population mean of small area 1 is the largest and the true population mean of small area 3 is the smallest among all areas. When we use the GCDP-WDP approach to estimate the population mean of small area 1, we presumed that the local sampling units from small area 1 and the foreign sampling units from the other three small areas are exchangeable by assigning positive ($\epsilon > 0$) weight to each unit in the prior. However the y values of most foreign sampling units are relatively small compared with the ones directly from small area 1 under our simulated model 1. Even though the constraints based on the auxiliary information tended to give

more weight on foreign sampling units with larger x values as well as larger y values, the existence of smaller foreign y's introduced more bias to the estimator when we were estimating the population mean for small area 1. On the other hand, the y values of local sampling units for small areas 2 and 4 were not systematically greater than or less than the foreign sampling units under our simulated model 1, which provided more similarities between these two areas and the other areas and borrowed strength more successfully in the estimating procedure. Therefore the absolute error of the estimator for small area 1 was larger than it was for small area 2 and 4. Similar arguments apply to small area 3. Notice that the EBLUP estimator has a similar performance in terms of the absolute errors. Such a property is very common in small area estimation and it is hard to eliminate when people want to borrow strength across small areas. For the local estimators, say direct and regression estimators, such a property of point estimates does not exist.

Keeping the same parameter settings, let us consider a second example with simulated populations through the following model.

Model 2: $y_{jl} = 2 * x_{jl} + \epsilon_{jl}, \quad$ where $\sigma_j = 2,\ \epsilon_{jl} \sim N(0, \sigma = \sqrt{x_{jl}})$

The simulated population information of model 2 is summarized in the Table 5.3. Note that the correlations between the two variables in model 2 are much weaker than they are in model 1. The population variances of y and x are relatively large compared with the simulated small areas in model 1.

Small Area	N_j	Corr (y, x)	$\mu_Y(\text{Var}(Y))$	$\mu_X(\text{Var}(X))$
1	104	0.387	302.538(181.901)	151.130(3.309)
2	103	0.281	297.253(223.010)	149.431(4.988)
3	87	0.185	301.970(180.662)	151.163(3.061)
4	97	0.304	296.779(149.390)	147.554(4.700)
Total	391	0.334	299.591(189.966)	149.803(6.187)

Table 5.3: Population statistics of simulated small areas for Model 2

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
1	Direct	302.224	4.990	291.426	21.598	0.888
2		296.140	5.605	284.131	24.019	0.827
3		301.870	4.670	291.159	21.422	0.897
4		296.838	4.412	286.770	20.137	0.888
1	Regression	302.132	5.088	292.320	19.623	0.822
2		296.565	5.995	285.347	22.437	0.762
3		302.248	5.213	292.220	20.056	0.822
4		297.166	5.069	287.541	19.250	0.822
1	EBLUP	301.675	3.898	290.584	22.183	0.942
2		297.605	3.584	287.808	19.594	0.918
3		301.415	3.673	290.002	22.827	0.953
4		295.857	4.434	282.771	26.173	0.951
1	GCDP	301.812	3.203	296.911	9.814	0.764
2	(2000th)	298.245	2.672	293.561	9.470	0.838
3	($\epsilon = 1$)	301.711	3.126	296.822	9.855	0.752
4		295.537	3.608	291.388	8.382	0.648
1	GCDP	301.812	3.197	296.137	11.351	0.830
2	-WDP	298.241	2.675	292.448	11.585	0.924
3	($\epsilon = 1$)	301.703	3.138	296.033	11.341	0.832
4		295.548	3.603	290.259	10.578	0.752

Table 5.4: Results from simulations for Model 2

The performances of the different approaches for model 2 are summarized in table 5.4. The average of AEMSE of the direct estimator is 37.618. The average of AEMSE of the regression estimator is 46.752. The average of AEMSE of the EBLUP estimator is 24.343. The average of AEMSE of the GCDP estimator based on 3600 sample points at every 2000th position in Markove chain is 16.403. The average of AEMSE of the GCDP-WDP estimator for $\epsilon = 1$ is 16.386. The average coverage rate of the 95% confidence/credible interval of EBLUP is the highest, however the average length of its intervals is much longer than those of the GCDP-WDP approach. The GCDP-WDP estimator for $\epsilon = 1$ provides credible intervals of the shortest length and the average coverage rate is about the same as direct estimator and regression estimator. The

average coverage rate for small area 4 is the smallest and the average coverage rate for small area 1 is the second smallest since the true population mean of small area 4 is the smallest and the true population mean of small area 1 is the largest. Hence they were shrunken the most to the center (the true overall mean of four small area is 299.59) when borrowing strength across small areas. For the local estimators, say the direct and regression estimators, such a pattern of interval coverage rate does not exist. We have explained this clearly in the previous example of model l. In summary, the GCDP-WDP estimator for $\epsilon = 1$ performs better than the direct and regression estimators and beats EBLUP regarding to AEMSE and average absolute errors.

To explore the effect of the hyperparameter ϵ in our estimating procedure, we consider the GCDP-WDP method for various choices of ϵ. The simulation results are presented in table 5.5.

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
1	GCDP-WDP	300.744	9.006	296.863	7.761	0.258
2	$\epsilon = 0.1$	297.692	7.816	293.408	8.568	0.280
3		300.801	8.970	296.936	7.730	0.244
4		294.322	8.068	290.397	7.849	0.314
1	GCDP-WDP	300.749	8.980	296.828	7.842	0.258
2	$\epsilon = 0.5$	297.695	7.790	293.371	8.646	0.286
3		300.806	8.945	296.900	7.811	0.248
4		294.328	8.044	290.365	7.925	0.316
1	GCDP-WDP	301.812	3.197	296.137	11.351	0.830
2	$\epsilon = 1$	298.241	2.675	292.448	11.585	0.924
3		301.703	3.138	296.033	11.341	0.832
4		295.548	3.603	290.259	10.578	0.752
1	GCDP-WDP	301.672	3.239	295.950	11.445	0.826
2	$\epsilon = 5$	298.538	2.548	292.785	11.505	0.924
3		301.683	3.198	295.970	11.427	0.842
4		295.217	3.895	289.864	10.705	0.700
1	GCDP-WDP	301.638	3.279	295.913	11.451	0.826
2	$\epsilon = 15$	298.596	2.547	292.857	11.480	0.918
3		301.677	3.225	295.959	11.436	0.834
4		295.149	3.979	289.788	10.723	0.692
1	GCDP-WDP	301.624	3.296	295.899	11.451	0.818
2	$\epsilon = 50$	298.615	2.553	292.881	11.467	0.918
3		301.673	3.240	295.953	11.438	0.830
4		295.120	4.011	289.756	10.728	0.692

Table 5.5: Comparison of small area estimates from the GCDP-WDP approach with different hyperparameters for the population of Model 2

The average AEMSE of the GCDP-WDP estimator when $\epsilon = 0.1$ is 105.463. The average AEMSE of the GCDP-WDP estimator when $\epsilon = 0.5$ is 104.832. The average AEMSE of the GCDP-WDP estimator when $\epsilon = 1$ is 16.386. The average AEMSE of the GCDP-WDP estimator when $\epsilon = 5$ is 17.261. The average AEMSE of the GCDP-WDP estimator when $\epsilon = 15$ is 17.685. The average AEMSE of the GCDP-WDP estimator when $\epsilon = 50$ is 17.868. The GCDP-WDP estimator for $\epsilon = 1$ is the best. Noticed that the performances of GCDP-WDP estimators are very similar for the larger values of ϵ.

This can be explained through the formula (3.20) in the following manner. Imagine that we did not have constraints for the auxiliary variables, the approximate posterior distribution of the proportion parameters would be

$$f(\lambda|z) \propto \prod_{j=1}^{4} \prod_{i=1}^{20} \lambda_{ji}^{z_{ji}+\epsilon-1} \tag{5.1}$$

where $z_{ji} = 1$ and $\sum_{i=1}^{20} z_{ji} = 5$ in our simulated model 2. The marginal posterior distribution of $\lambda_j = (\lambda_{j1}, \cdots, \lambda_{j,20})$ is approximately a Dirichlet distribution with parameters $(z_{j,1} + \epsilon, \cdots, z_{j,20} + \epsilon)$ when we assume the small areas are independent of each other. Thus the marginal posterior expectation of each proportion λ_{ji}, $E(\lambda_{ji}|z)$, is approximately equal to $(z_{ji} + \epsilon)/(5 + 20\epsilon)$ if we ignore the constraints. For large ϵ, this expectation is close to $1/20$ leading to a uniform posterior distribution of λ_j, and it does not rely on j very much. A large ϵ presumes that the foreign sampling units should be treated as importantly as the local sampling units, which means that we believe that there exists strong similarities between small areas. Therefore when ϵ is sufficiently large and when we ignore the constraints, the GCDP-WDP estimators for all small areas should be roughly the same given that marginal posterior distributions of the λ_j's are almost the same. However, on the other hand the actual posterior distributions of the λ_j's depend on their constraints. These constraints vary across small areas and they adjust the GCDP-WDP estimators so that the marginal constrained posterior distributions of the λ_j's are different across small areas no matter how large ϵ is. In other words, ϵ controls how much foreign strength should be borrowed and the constraints determine how this strength is allocated. The influence of the constraints seem to be dominant in our approach when ϵ is sufficiently large.

5.2 Populations with a Nonlinear Relationship

In this section we consider two examples where the relationship between the quantity of interest and the auxiliary variable is nonlinear. In the first example the small areas will be relatively similar and we shall see that our methods still works. In the second example the small areas are not so alike and we shall see that our methods work poorly.

5.2.1 A Good Example

As before let N_j denote the population size for each small area, where

$$N_j \sim \text{Poisson}(\lambda = 1000),\ j = 1, 2, 3, 4.$$

(y_{jl}, x_{jl}) is the lth unit of small area j. Let

$$x_{jl} \sim \text{Gamma}(\alpha_j, \beta_j), \quad l = 1, \cdots, N_j$$

where α_j is the shape parameter and β_j is the scale parameter of the Gamma distribution for the auxiliary variable for the jth small area. We assumed that

$$\alpha_j \sim N(\mu = 4, \sigma = 0.5),\ j = 1, 2, 3, 4.$$

Let $\beta_j = 1$ for each small area. For each small area, the units of interest are generated through the following model:

Model 3: $\quad y_{jl} = 20 + 12 * x_{jl} - 1.5 * (x_{jl} - 5)^2 + \epsilon_{jl}, \quad \text{where } \epsilon_{jl} \sim N(0, \sigma = 0.5)$

We assume the population mean of the auxiliary variable is known for each small area. 500 random samples each of size 5 ($n_j = 5$) were generated from each small area such that the samples satisfy the constraints described in formula (3.16).

For each sample we computed the sample mean (direct estimator), two regression es-

timators and the EBLUP estimator and the GCDP-WDP estimator as well as their corresponding 95% confidence intervals/credible intervals given that the population mean of the auxiliary variable is known. For the regression estimators, we considered two cases: in the first case we just did simple linear regression and in the second we used both a linear and quadratic term. For the later we assumed that for each small area the population second moment of the auxiliary moment was known as well. The GCDP estimator was calculated with a hyperparameter of $\epsilon = 1$ in the prior distribution. The simulated population information of model 3 is summarized in table 5.6.

Small Area	N_j	Corr (y, x)	$\mu_Y(\text{Var}(Y))$	$\mu_X(\text{Var}(X))$
1	1035	0.887	63.545 (643.380)	4.264 (4.55)
2	985	0.911	61.813 (640.772)	4.076 (3.86)
3	963	0.953	54.177 (724.074)	3.559 (3.61)
4	990	0.924	71.920 (566.131)	4.881 (4.43)
Total	3973	0.919	62.932 (681.682)	4.200 (4.34)

Table 5.6: Population statistics of simulated small areas for Model 3

The comparison of the estimation methods for model 3 are summarized in table 5.7. The average AEMSE for the direct estimator is 121.746. The average AEMSE for the linear regression estimator is 22.819. The average AEMSE for the quadratic regression estimator is 4.997. This is based on knowing that the degree of polynomial is a quadratic and that the first and the second moments of the x_{jl}'s are known for each small area. The average AEMSE of the EBLUP estimator is 88.129. The average AEMSE of the GCDP-WDP estimator is 5.257. Our GCDP-WDP approach performs almost as well as the quadratic regression estimator in terms of AEMSE. However the GCDP-WDP estimator does not need to know the form of the model and the area-level second moments of the auxiliary variable.

The average of the four small area point estimators (denoted by $\hat{\mu}^j_{Reg1}$) for the simple linear regression method is systematically larger than the true population mean for all of the four small areas. This result matches with the theoretical finding proposed by

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
1	Direct	63.841	8.928	42.903	41.875	0.894
2		61.998	8.979	40.835	42.328	0.852
3		54.379	9.527	31.553	45.650	0.874
4		72.356	7.850	52.484	39.746	0.882
1	Regression1	66.515	4.464	61.550	9.929	0.500
2		64.205	3.793	59.594	9.221	0.524
3		56.273	3.274	52.282	7.982	0.550
4		74.311	4.048	69.313	9.995	0.576
1	Regression2	63.530	0.746	53.337	20.386	0.992
2		62.057	0.789	52.752	18.610	0.992
3		54.236	0.694	46.364	15.744	0.996
4		71.823	0.652	62.245	19.156	0.994
1	EBLUP	64.756	6.037	45.105	39.302	0.918
2		62.442	6.081	42.409	40.067	0.940
3		54.926	9.052	30.458	48.935	0.934
4		73.245	7.922	48.291	49.907	0.960
1	GCDP	64.231	1.909	53.518	21.425	0.998
2	-WDP	61.917	1.581	51.163	21.509	1.000
3	($\epsilon = 1$)	54.543	1.346	43.591	21.904	1.000
4		71.424	2.193	60.880	21.088	0.994

Table 5.7: Results from the simulations for Model 3

Deng and Chhikara (1990). In their paper, they showed that the sign of the quadratic term in the quadratic regression fit would be indicative of an over- or under-estimation of a simple linear regression fit. The sign of the approximate bias of a linear fit depends only on the sign of the quadratic term. If it is greater than 0, then we expect $\hat{\mu}^j_{Reg1}$ to underestimate μ^j, whereas a negative quadratic term indicates overestimation of μ^j. In our simulated example, the true model has a negative quadratic term, thus the simple linear regression estimator is expected to be larger than the true mean for all small areas.

Table 5.7 showed that the average absolute error of the GCDP-WDP estimator is the smallest among all the estimators except the quadratic regression estimator, which

was obtained under stronger assumptions. The GCDP-WDP approach produced less absolute error, larger coverage rate and shorter length of credible intervals than the EBLUP approach in this nonlinear situation.

5.2.2 A Bad Example

As we have discussed in chapter 3, our GCDP approach assumed that the sampling units are to some extent exchangeable across small areas which presumes small areas are similar to each other. What will happen if the small areas are not similar? Is GCDP still applicable? Let us consider another nonlinear example. Suppose we have four small areas with one auxiliary variable. $N_j \sim \text{Poisson}(\lambda = 100),\ j = 1, 2, 3, 4.$ (y_{jl}, x_{jl}) is the lth unit of small area j. α_j is the shape parameter of the Gamma distribution of the auxiliary variable, $x_{jl}, l = 1, \cdots, N_j$, for each small area. We assumed that

$$\alpha_j \sim N(\mu = 4, \sigma = .25),\ j = 1, 2, 3, 4.$$

β_j is the scale parameter population of the Gamma distribution of auxiliary variable, i.e.

$$x_{jl} \sim \text{Gamma}(\alpha_j, \beta_j).$$

We set $\beta_j = 1$ in the following simulation. For each small area, the unit of interest, y_{jl} is generated through the following model:

$$\textbf{Model 4:}\quad y_{jl} = 60 - 8 * (x_{jl} - 4)^2 + \epsilon_{jl}, \quad \text{where } \epsilon_{jl} \sim N(0, \sigma = x_{jl})$$

We assumed that the population mean of the auxiliary variable is known for each small area. 500 random samples each of size 5 ($n_j = 5$) were generated from each small area such that the samples satisfy the constraints described in formula (3.16). For each sample we computed the sample mean (direct estimator), two regression estimators and the EBLUP estimator and the GCDP-WDP estimator as well as their corresponding

95% confidence intervals/credible intervals given that the small area population means of the auxiliary variable are known. For the regression estimators, we considered two cases: in the first case we just did simple linear regression and in the second we used both a linear and quadratic term. For the later we assumed that for each small area the population second moment of the auxiliary moment was known as well. The GCDP estimator was calculated with a hyperparameter of $\epsilon = 1$ in the prior distribution. The simulated population information of model 4 is summarized in table 5.8. The population means and variances of y for all small areas are not very close. The large values of the population variances for the auxiliary variable allow less homogeneity among small areas than the ones of the previous example.

Small Area	N_j	Corr (y, x)	$\mu_Y(\text{Var}(Y))$	$\mu_X(\text{Var}(X))$
1	103	-0.712	21.758 (4132.352)	4.507 (4.50)
2	97	-0.816	3.151 (19268.120)	4.377 (7.05)
3	119	-0.618	31.435 (3027.268)	3.859 (3.61)
4	106	-0.708	12.854 (28263.258)	4.531 (5.58)
Total	425	-0.692	18.000 (13309.501)	4.302 (5.14)

Table 5.8: Population statistics of simulated small areas for Model 4

The simulation results for model 4 are summarized in table 5.9. The average AEMSE of the direct estimator is 2469.449. The average AEMSE of the simple linear regression estimator is 1160.044. The average AEMSE of the quadratic regression estimator is 573.0162. The average AEMSE of the EBLUP estimator is 824.110. The average AEMSE the GCDP-WDP estimator ($\epsilon = 1$) is 327.911. Our GCDP-WDP approach performs best in terms of AEMSE.

The quadratic regression estimator has larger AEMSE than GCDP-WDP estimator since our simulated population did not assume constant variance in the superpopulation model, which introduced more variability into the quadratic fit. The average absolute error of the quadratic regression estimator is minimum given that we know the small

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
1	Direct	20.787	21.627	-20.570	82.715	0.676
2		2.670	47.979	-75.669	156.678	0.512
3		30.523	18.179	-2.900	66.846	0.716
4		14.272	35.606	-43.399	115.343	0.534
1	Regression1	34.459	20.351	8.699	51.519	0.548
2		33.246	40.761	-1.054	68.599	0.396
3		45.029	19.150	24.296	41.467	0.504
4		37.597	30.680	9.062	57.071	0.392
1	Regression2	21.380	5.207	-30.898	104.555	1.000
2		4.618	12.076	-85.427	180.091	0.992
3		32.612	6.394	-13.442	92.108	0.996
4		10.400	10.185	-64.328	149.457	0.992
1	EBLUP	38.275	19.239	17.223	42.104	0.456
2		38.164	35.719	19.301	37.727	0.196
3		34.253	17.510	-1.376	71.257	0.804
4		38.732	27.106	17.157	43.150	0.326
1	GCDP	18.785	11.568	-14.256	66.082	0.898
2	-WDP	19.673	20.561	-14.364	68.074	0.588
3	($\epsilon = 1$)	30.908	8.208	5.955	49.907	0.900
4		18.149	14.244	-15.797	67.890	0.786

Table 5.9: Results from the simulations for Model 4

area second moments of the auxiliary variable as well as the degree of the relationship. The average absolute error of the GCDP-WDP estimator is the second smallest and does not need to know the second moments of the auxiliary variable. It is clear that even though the linear correlations between variables are moderate and close for all small areas, the linear regression estimator and EBLUP perform poorly.

5.3 Categorical Populations

The overall sample size in some national surveys are usually determined to provide specific accuracy at nation-level or state-level of aggregation. The sample size is typically small for a county-level study, which identifies counties as small areas. In the survey,

face-to-face interviews provide a great deal of individual auxiliary information. In addition, a previous census sometimes provides useful county-level auxiliary information. Most survey questions are yes or no questions, thus binary variables are frequently used in such data sets. For example, we are interested in county-level estimates of the percentage of women age 40 or over who have had a mammogram in the past 2 years. We let the variable of interest $y = 1$, if a women in the county have had a mammogram, otherwise $y = 0$. It is quite possible to find counties where either all the sampled women have had a mammogram or none of them have had one. This is likely to happen when the sample size in a county is small or when the true percentage of women having had mammogram is close to either 1 or 0. In this case, the direct estimator of the true proportion of women who have had a mammogram in the county is either 0 or 1 based only on the local data, which is not realistic. If the direct estimator is not reliable, then the other indirect estimators relying on the direct estimator in their model become problematic. This stimulates us to find an indirect estimator which does not rely on the direct estimator and which uses some information from similar counties to construct reasonable estimators. Our approach can construct such estimators.

In this section we assume that we have 4 small areas and that the elements of each small area are associated with the elements of m categorical auxiliary variables. For simplicity, we consider $m = 2$, but the same theory applies to more than two categorical variables. The population size for each small area is N_j, $N_j \sim \text{Poisson}(\lambda = 250)$, $j = 1, 2, 3, 4$.

In small area j, the lth unit is either 0 or 1. The units of small areas are generated with the following probability model

$$\textbf{Model 5:} \quad P(y_{jl} = 1 | x_{jl1}, x_{jl2}) = \frac{\exp\left(0.5 + 2x_{jl1} - 4x_{jl2} + \nu_j\right)}{1 + \exp\left(0.5 + 2x_{jl1} - 4x_{jl2} + \nu_j\right)}$$

where $x_{jl1} \sim$ Bernoulli$(1/3)$, $x_{jl2} \sim$ Bernoulli$(1/5)$ and $\nu_j \sim N(0, \sigma = 0.3)$ for j in $\{1, \cdots, 4\}$ and l in $\{1, \cdots, N_j\}$.

Small Area	N_j	μ_Y	μ_X^1	μ_X^2
1	257	0.576	0.237	0.331
2	256	0.535	0.184	0.312
3	230	0.670	0.217	0.348
4	245	0.490	0.220	0.359
Total	988	0.566	0.337	0.215

Table 5.10: Population statistics of simulated small areas for Model 5

We are interested in estimating the population mean of y for each small area, i.e. the true percentage of $y = 1$. For each small area we take 500 random samples according to the sampling plan that the samples need to satisfy the prior constraints. The sample size has three choices $n_j = 3$, $n_j = 10$ and $n_j = 30$ for $j = 1, \cdots, 4$. For each sample we compute the sample mean and the GCDP-WDP estimator given that the means of the auxiliary variables are known. The hyper parameter ϵ is chosen to be 1. We should note that the situation for a categorical population is different than that of a population with continuous variables. For a continuous population, say for the simulated populations in section 5.1, the dimension of the parameter space Λ of the GCDP approach is 80 when $n_j = 5$ and $M = 4$ since the number of distinct units in the sample will always be 20, i.e. $k = 20$. When n_j gets larger and M is fixed, then k gets larger as well. This means that the dimension of Λ gets larger too. However, for a categorical population, the dimension of parameter space Λ behaves differently. For a fixed M, say $M = 4$, the number of distinct vectors $(y_{jl}, x_{jl}^1, x_{jl}^1)$ in the sample is at most 8 if all three variables are binary, so the dimension of Λ is at most 32 no matter how large the sample sizes are. It could be much less than 32 if n_j is very small, but it will not be greater than 32. In this categorical situation, we are able to consider more small areas and larger sample sizes without increasing the dimension of parameter space dramatically.

Table 5.10 summarizes the statistics of the simulated population of model 5.

The GCDP-WDP beats the direct estimator in terms of AEMSE for all three choices of the sample size. When the sample size is as small as 3, the sample mean (direct

Small Area	Method	Point estimate		95% Confidence or credible intervals		
		Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
			$n_j = 3$			
1	Direct	0.557	0.234	0.064	0.987	0.760
2		0.512	0.255	0.038	0.948	0.730
3		0.645	0.191	0.188	0.914	0.704
4		0.499	0.257	0.027	0.945	0.728
1	GCDP	0.541	0.092	0.140	0.801	0.990
2	-WDP	0.526	0.102	0.126	0.798	0.994
3	($\epsilon = 1$)	0.573	0.116	0.174	0.797	0.984
4		0.521	0.099	0.121	0.801	0.990
			$n_j = 10$			
1	Direct	0.568	0.116	0.266	0.605	0.966
2		0.532	0.125	0.229	0.605	0.874
3		0.667	0.117	0.385	0.565	0.898
4		0.493	0.122	0.190	0.605	0.888
1	GCDP	0.546	0.080	0.177	0.739	1.000
2	-WDP	0.513	0.084	0.142	0.740	1.000
3	($\epsilon = 1$)	0.605	0.087	0.242	0.727	1.000
4		0.495	0.080	0.124	0.741	1.000
			$n_j = 30$			
1	Direct	0.575	0.066	0.409	0.333	0.934
2		0.531	0.064	0.362	0.337	0.944
3		0.666	0.061	0.508	0.315	0.954
4		0.497	0.068	0.329	0.336	0.950
1	GCDP	0.564	0.050	0.220	0.689	1.000
2	-WDP	0.523	0.049	0.176	0.694	1.000
3	($\epsilon = 1$)	0.634	0.053	0.300	0.669	1.000
4		0.493	0.052	0.146	0.694	1.000

Table 5.11: Results from simulations for Model 5

Sample size n_j	Ave. AEMSE of Direct	Ave. AEMSE of GCDP-WDP
$n_j = 3$	0.082	0.017
$n_j = 10$	0.023	0.010
$n_j = 30$	0.007	0.004

Table 5.12: Average of AEMSE for Model 5 when $n_j = 3$, $n_j = 10$ and $n_j = 30$.

estimator) produced very large absolute errors compared with GCDP-WDP estimator. The coverage rate of its confidence interval is very low. With the increasing of sample size the performance of two point estimators are getting closer while the direct estimator has better interval estimates for the larger sample size.

5.4 Constraints Across Small Areas

To examine the performance of simultaneous estimating procedure, we consider three small areas in this section. Each unit in the jth small area is generated by the following model:

$$\textbf{Model 6:} \quad y_{jl} = 100 + x_{1jl} + 10 * x_{2jl} + \epsilon_{jl}$$

where $x_{1jl} \sim$ Gamma(α_j) for $\alpha_1 = 5,\ \alpha_2 = 7,\ \alpha_3 = 9,\ x_{2jl} \sim$ Bernoulli(p_j) for $p_1 = 0.6,\ p_2 = 0.72,\ p_3 = 0.75$ and $\epsilon_{jl} \sim N(0, \sigma = 7)$.

Small Area	N_j	$\rho(y, x_1)$	$\rho(y, x_2)$	μ_y	var_y	μ_{x_1}	var_{x_1}	μ_{x_2}	var_{x_2}
1	300	0.214	0.574	111.384	80.322	4.798	4.609	0.637	0.232
2	300	0.225	0.558	114.326	81.327	6.940	6.223	0.713	0.205
3	300	0.464	0.484	116.374	86.631	8.936	8.566	0.703	0.209
Total	900	0.377	0.537	114.028	86.776	6.891	9.310	0.684	0.216

Table 5.13: Population statistics of simulated small areas for Model 6

The simulated population statistics are summarized in table 5.13.

500 random samples are drawn from each small area. All the approaches assume that the value of x_2 is not observed in the sample. Each small area population mean of the first auxiliary variable, x_1, is assumed to be known. The simulation results are summarized in table 5.14. The average of AEMSE over 550 samples is 16.411 for direct estimator; it is 22.389 for regression estimator; it is 13.200 for EBLUP estimator; it is 9.223 for our Bayesian GCDP-WDP esimator, where the Markov Chain is of length 8,000,000 and the first 800,000 points were disregarded. For our own Bayesian approach, we considered one more case where the overall population mean of x_2 is also known in addition to the auxiliary information about x_1. Such additional across area auxiliary information is not able to be considered by the other methods, however our approach can utilize it. This requires the simultaneous small area estimation discussed in section 3.5. The average of AEMSE for GCDP-WDP with the additional auxiliary information

Small Area	Point estimate		95% Confidence or credible intervals		
	Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
Direct					
1	111.621	3.130	104.327	14.588	0.887
2	114.559	3.267	107.271	14.575	0.856
3	116.644	3.325	108.707	15.873	0.889
Regression					
1	111.313	3.718	104.310	14.007	0.800
2	114.119	3.707	107.275	13.687	0.802
3	116.637	3.434	109.660	13.953	0.818
EBLUP					
1	111.701	3.121	101.707	19.988	0.965
2	114.435	2.250	106.187	16.495	0.958
3	116.744	3.185	106.766	19.956	0.960
GCDP-WDP ($\epsilon = 1$)					
1	112.190	2.735	108.189	8.001	0.758
2	114.288	1.888	110.004	8.569	0.911
3	116.679	2.346	112.417	8.525	0.822
GCDP-WDP($\epsilon = 1$) with joint constraints for x_2					
1	112.345	2.564	108.357	7.977	0.778
2	114.310	1.797	110.038	8.544	0.929
3	116.673	2.248	112.418	8.510	0.847

Table 5.14: Results from simulations for Model 6

about x2 is 8.461, where the Markov Chain is of length 8,000,000, and the first 800,000 points were disregarded.

In summary, our Baysian approach provide best point estimates in terms of AEMSE and the absolute error. Further more additional joint auxiliary information has significant improvement on the accuracy of both point estimates and interval estimates.

Chapter 6

Application: County-Level Small Area Estimation Using BRFSS

To further evaluate the performance of the new method introduced in this thesis, we apply both frequentist and the new Bayesian methods to a real example in this chapter. To best assess the performance of these estimation methods, it would be ideal to find an example for which the population is fully known. However, it is very hard to obtain such a complete population in practice. In order to study the new Bayesian methods, it was necessary to sacrifice knowledge of the full population.

The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project of the Centers for Disease Control and Prevention (CDC) and U.S. states and territories. The BRFSS, administered and supported by CDC's Behavioral Surveillance Branch, is an ongoing data collection program designed to measure behavioral risk factors for the adult population (18 years of age or older) living in households. The BRFSS objective is to collect uniform, state specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. Factors assessed by the BRFSS include tobacco use, health care coverage, HIV/AIDS knowledge and prevention, physical activity, and fruit and

vegetable consumption. Data are collected from a random sample of adults (one per household) through a telephone survey, which covers approximately 94% of U.S. households. In this chapter, we will assume the full population is a set of 2336 Minneapolis residents from the BRFSS (2009) data source. These residents come from the following four counties respectively: "Anoka County", "Dakota County", "Hennepin County" and "Ramsey County". Each county will be treated as a small area in our study.

6.1 Population Information of County Levels

BRFSS (2009) provides a rich array of numerical variables related to the health characteristics of Minneapolis residents. Most of the variables are categorical ones. The primary variables of our interest are listed below:

1. *county.* A factor identifying the county or the small area which the resident belongs to. The levels 1 to 4 represent "Anoka County", "Dakota County", "Hennepin County" and "Ramsey County" respectively.

2. *health.* A binary variable indicating the general health status of the resident. If the resident's general health status is good, *health* takes the value of '1'. If the resident's health status is not good, *health* takes the value of '0'.

3. *exercise.* A binary variable indicating whether the resident participates in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise other than his or her regular job during the past month. If the resident did exercises, *exercise* takes the value of '1'. Otherwise, it takes the value of '0'.

4. *age.* A numerical variable which reported the age of the resident being interviewed in the survey.

In this chapter, we are interested in county-level estimates of the percentage of Minneapolis residents age 18 and over whose general health status were good in 2009.

We conducted simulations which examine the proposed Bayesian approaches in chapter 3 under different settings of constraints by using the four variables listed above. Let us denote the variable *health* by y in what follows. The variable *exercise* is denoted by x_1 and the variable *age* is denoted by x_2.

The true population information is summarized in Table 6.1. Plots of the data are given in Figure 6.1 and Figure 6.2. The table and figures show that there exists similarity between counties, therefore it would be appropriate to borrow strength across these four counties when we analyze the data.

County		Health		Exercise		Age		Correlations	
j	N_j	μ_y	Var_y	μ_{x_1}	Var_{x_1}	μ_{x_2}	Var_{x_2}	$\rho(y, x_1)$	$\rho(y, x_2)$
1	291	0.567	0.246	0.811	0.154	52.405	248.973	0.180	-0.177
2	381	0.651	0.228	0.840	0.135	53.738	248.699	0.251	-0.119
3	1135	0.620	0.236	0.841	0.134	56.226	280.034	0.200	-0.157
4	529	0.590	0.242	0.854	0.125	56.841	263.380	0.233	-0.246
Total	2336	0.612	0.238	0.840	0.135	55.483	269.314	0.213	-0.173

Table 6.1: Population statistics of county level for BRFSS Study

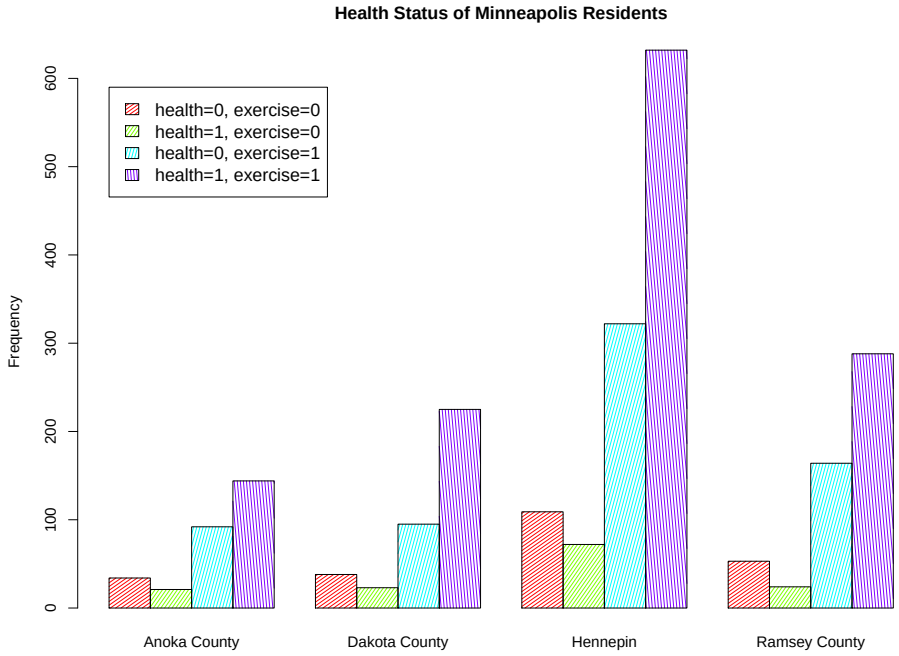

Figure 6.1: Health and exercise distribution of county level for BRFSS study

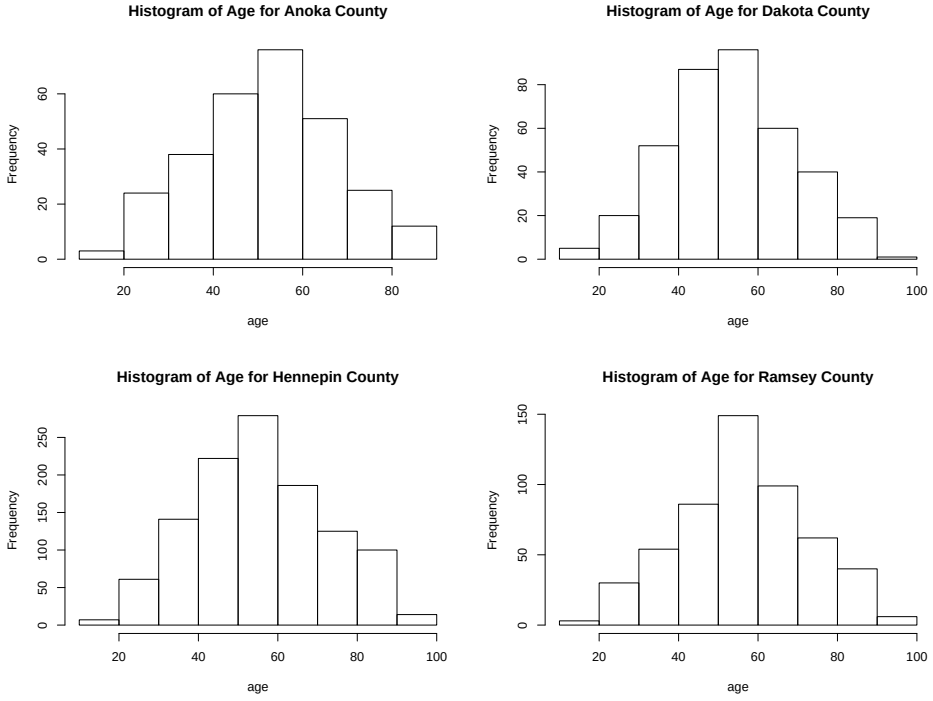

Figure 6.2: Age distribution of county level for BRFSS study

6.2 Estimating the Percentage of Health Status

Suppose now that the jth county population size N_j for $j = 1, \cdots, 4$ is always known to us. Let us use μ_y to denote the percentage of residents age 18 and over whose general health status is good. We are interested in estimating μ_y for each county.

To study the point and interval estimates associated with the various methods discussed in this thesis, 500 samples of either size 5 or 10 were taken from each county population respectively. The Generalized Constrained Weighted Dirichlet Posterior(GCDP-WDP) estimates of μ_y^j were computed for each county j and each random sample. For GCDP-WDP estimates, we considered the following three cases:

I. We assume that the population mean of the auxiliary variable "x_1" for each county is known and then we further assume that the overall population mean of the aux-

iliary variable "x_2" for the whole Minneapolis area (the union of four counties) is known. This second assumption introduces the auxiliary information across small areas to our Bayesian approach, so it becomes necessary to consider simultaneous small area estimation which was discussed in section 3.5. In this case we need to estimate μ_y^j for $j = 1, \cdots, 4$ simultaneously in order to utilize the auxiliary information across small areas.

II. We assume that the population mean of the auxiliary variable "x_1" for each county is known and the population mean of the auxiliary variable "x_2" is also known for each county. In terms of our Bayesian methodology proposed in chapter 3, case II focuses on the auxiliary information within small areas. In other words, it is possible to estimate the parameter μ_y^j for each county separately.

III. Besides the assumptions in case II, we consider one more constraint for our GCDP-WDP approach in case III. Recall that our main idea in this thesis is to simulate a complete copy of each county given what we have observed in all the counties. To this end it would make sense to force that the overall simulated population mean of y, to be consistent with the sample mean of y for the purpose of quality control. Therefore in case III we used this extra across area constraint when simulating complete copies of the counties.

For the purpose of comparison, we also compute the sample mean as the direct estimator of μ_y^j, and the EBLUP point estimates as indirect estimators of μ_y^j for any given random sample. EBLUP estimators are based on the Fay-Herriot model proposed in section 2.3 of chapter 2. Their corresponding interval estimates were also calculated. For EBLUP point estimates, three cases were under investigation:

1. Only the auxiliary variable *exercise*, x_1, was utilized in the linear model.

2. Only the auxiliary variable *age*, x_2, was utilized in the linear model.

3. Both auxiliary variables *exercise* and *age* were utilized in the linear model.

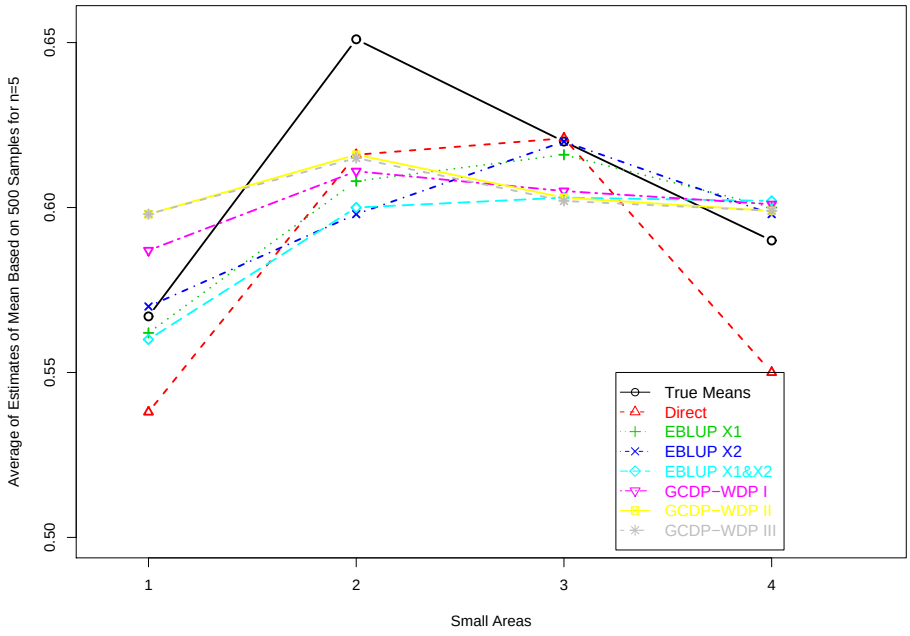

Figure 6.3: Average of point estimates based on 500 samples for BRFSS study: n_j=5

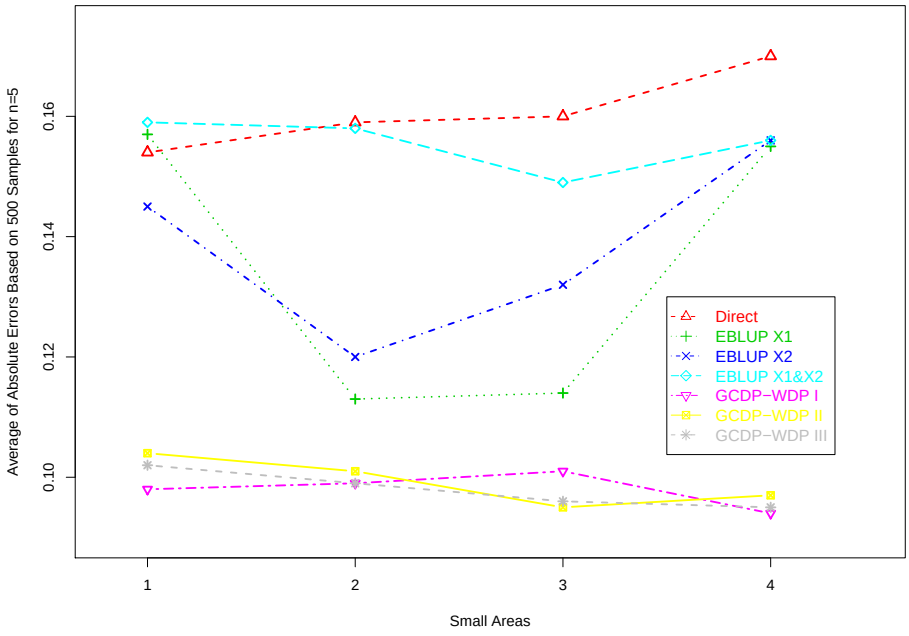

Figure 6.4: Average of absolute errors based on 500 samples for BRFSS study: n_j=5

The results of the simulations with sample size of $n_j = 5$ are given in Table 6.2, Table 6.3, Figure 6.3 and Figure 6.4. The numbers in the table cells are the averages of statistics based on 500 random samples. From the table, we can see that for the point estimation problem the direct estimator performs the worst. This is not surprising since the other methods are using additional information. The GCDP-WDP point estimator of case III which utilized the most auxiliary information among all the approaches, is the best. The performance of the EBLUP estimators is worse than the GCDP-WDP estimators and better than the direct estimator. This makes sense since the EBLUP approach partially relies on the direct estimator, which is usually not reliable due to the nature of the small sample size. In addition, the Fay-Herriot model assumes that there is a linear relationship between the direct estimator and the auxiliary variables, which may not be satisfied in our data set. Figure 6.3 shows that both the EBLUP approach and the GCDP-WDP approach seem to have a shrinkage effect on the point estimates. That is, on the average, the estimator for small area 2, which has the largest mean among four areas, tends to be smaller than the truth; while the estimator for small area 1, which has the smallest mean, tends to be larger than the truth.

The GCDP-WDP interval estimates are much shorter than the EBLUP and their coverage rate falls below the nominal 0.95 level although the intervals in case III are nearly in the acceptable range.

The Average Empirical Mean Square Errors (AEMSE) of the various approaches are summarized in Table 6.3 for $n_j = 5$. Clearly, the overall performance of the GCDP-WDP estimator of case III is the best. This estimator utilized not only the common auxiliary information available to other approaches, but also required that the average of the four small area estimators agree with the overall sample mean. This is information which is usually not considered by other methods. As a result it was able to improve on their performance.

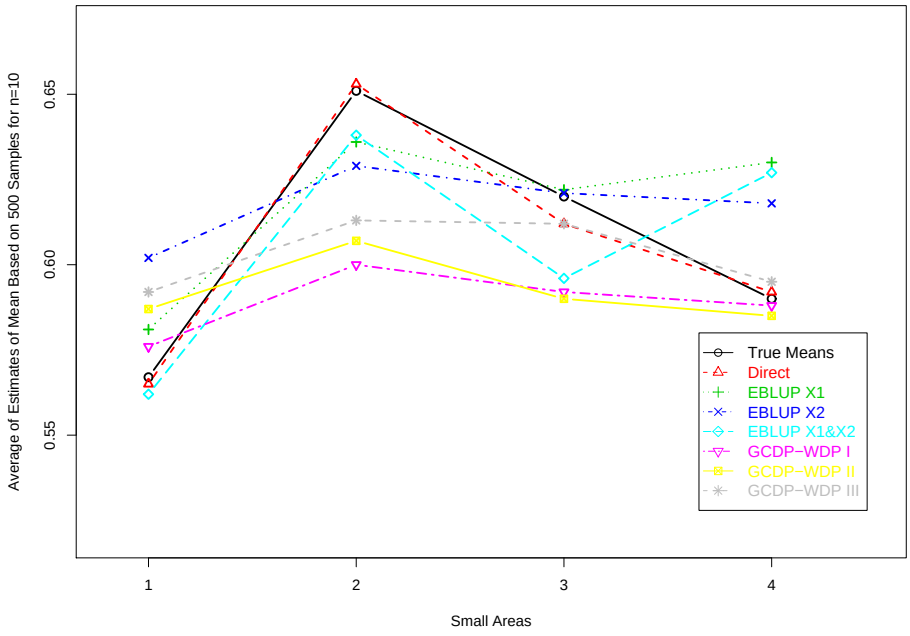

Figure 6.5: Average of point estimates based on 500 samples for BRFSS study: n_j=10

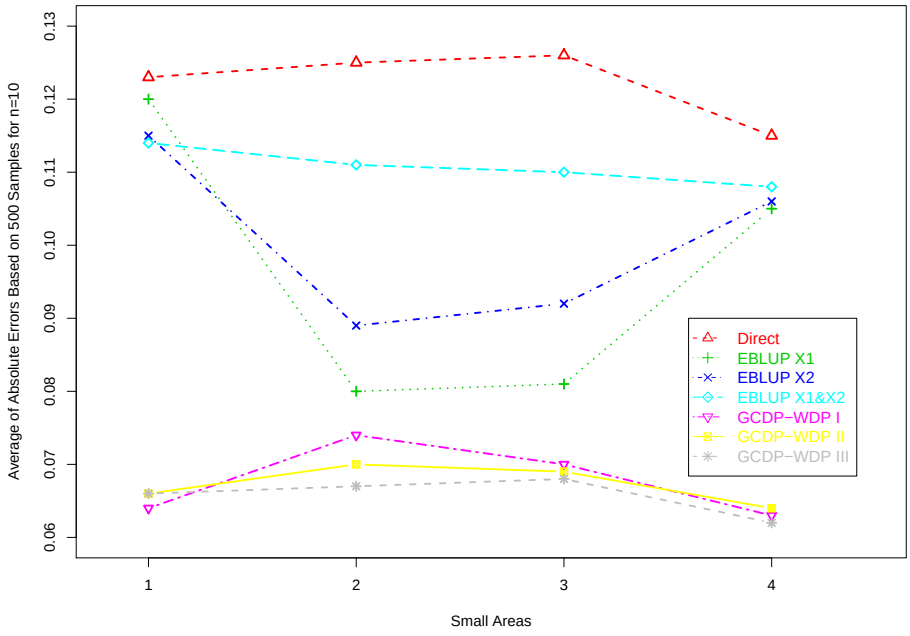

Figure 6.6: Average of absolute errors based on 500 samples for BRFSS study: n_j=10

The results of the simulations with sample size of $n_j = 10$ are given in Table 6.4, Table 6.5, Figure 6.5 and Figure 6.6. All the point estimates with sample size of 10 have smaller values of the average of absolute errors and AEMSE's compared with their corresponding simulation results with sample size of 5. All the interval estimates for $n_j = 10$ are shorter and have higher coverage rate than the corresponding ones for $n_j = 5$. For the sample size of 10, the GCDP-WDP estimators still outperform the direct estimator and EBLUP estimators in terms of the average of absolute errors of point estimates. The GCDP-WDP interval estimates are much shorter than the EBLUP and the GCDP-WDP's coverage rates of all three cases fall around the nominal 0.95 level. The overall performance of the GCDP-WDP estimator of case III with a sample size of 10 is the best regarding the AEMSE.

Small Area	Point estimate		95% Confidence or credible intervals		
	Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
Direct					
1	0.538	0.154	0.094	0.887	0.992
2	0.616	0.159	0.189	0.853	0.890
3	0.621	0.160	0.210	0.823	0.856
4	0.550	0.170	0.128	0.843	0.946
EBLUP (with *exercise*)					
1	0.562	0.157	-0.023	1.169	0.974
2	0.608	0.113	0.136	0.944	0.948
3	0.616	0.114	0.159	0.915	0.910
4	0.600	0.155	0.086	1.027	0.922
EBLUP (with *age*)					
1	0.570	0.145	0.021	1.097	0.980
2	0.598	0.120	0.110	0.975	0.964
3	0.620	0.132	0.144	0.951	0.906
4	0.598	0.156	0.095	1.007	0.908
EBLUP (with both auxiliary variables)					
1	0.560	0.159	-0.043	1.206	0.978
2	0.600	0.158	0.005	1.190	0.970
3	0.603	0.149	0.081	1.045	0.898
4	0.602	0.156	0.075	1.054	0.918
GCDP-WDP($\epsilon = 1$) Case I					
1	0.587	0.098	0.377	0.420	0.885
2	0.611	0.099	0.403	0.416	0.881
3	0.605	0.101	0.398	0.415	0.883
4	0.601	0.094	0.392	0.418	0.901
GCDP-WDP($\epsilon = 1$) Case II					
1	0.598	0.104	0.390	0.417	0.845
2	0.616	0.101	0.409	0.414	0.881
3	0.603	0.095	0.394	0.418	0.914
4	0.599	0.097	0.390	0.418	0.885
GCDP-WDP($\epsilon = 1$) Case III					
1	0.598	0.102	0.389	0.417	0.866
2	0.615	0.099	0.408	0.415	0.889
3	0.602	0.096	0.393	0.417	0.904
4	0.599	0.095	0.390	0.418	0.895

Table 6.2: Simulation results of BRFSS study for $n_j = 5$

Method	AEMSE
Direct	0.04231
EBLUP (with *exercise*)	0.03135
EBLUP (with *age*)	0.03194
EBLUP (with *both*)	0.03904
GCDP-WDP (Case I)	0.01506
GCDP-WDP (Case II)	0.01529
GCDP-WDP (Case III)	0.01501

Table 6.3: Average of AEMSE for BFRSS study for $n_j = 5$

Small Area	Point estimate		95% Confidence or credible intervals		
	Ave. of estimate	Ave. of abs. error	Ave. of lower bound	Ave. of length	Freq. of coverage
Direct					
1	0.565	0.123	0.262	0.605	0.958
2	0.653	0.125	0.367	0.572	0.898
3	0.612	0.126	0.315	0.593	0.894
4	0.592	0.115	0.289	0.605	0.968
EBLUP (with *exercise*)					
1	0.581	0.120	0.169	0.825	0.992
2	0.636	0.080	0.303	0.666	0.968
3	0.622	0.081	0.288	0.669	0.976
4	0.630	0.105	0.254	0.751	0.980
EBLUP (with *age*)					
1	0.602	0.115	0.215	0.775	0.982
2	0.629	0.089	0.288	0.682	0.972
3	0.621	0.092	0.269	0.704	0.982
4	0.618	0.106	0.245	0.746	0.974
EBLUP (with both auxiliary variables)					
1	0.562	0.114	0.139	0.846	1.000
2	0.638	0.111	0.216	0.844	0.990
3	0.596	0.110	0.211	0.769	0.982
4	0.627	0.108	0.241	0.772	0.974
GCDP-WDP($\epsilon = 1$) Case I					
1	0.576	0.064	0.416	0.319	0.947
2	0.600	0.074	0.442	0.316	0.915
3	0.592	0.070	0.434	0.316	0.935
4	0.588	0.063	0.429	0.318	0.955
GCDP-WDP($\epsilon = 1$) Case II					
1	0.587	0.066	0.428	0.318	0.942
2	0.607	0.070	0.449	0.315	0.924
3	0.590	0.069	0.432	0.317	0.933
4	0.585	0.064	0.426	0.318	0.947
GCDP-WDP($\epsilon = 1$) Case III					
1	0.592	0.066	0.433	0.317	0.951
2	0.613	0.067	0.456	0.314	0.927
3	0.612	0.068	0.455	0.314	0.931
4	0.595	0.062	0.437	0.317	0.951

Table 6.4: Simulation results of BRFSS study for $n_j = 10$

Method	AEMSE
Direct	0.02271
EBLUP (with *exercise*)	0.01551
EBLUP (with *age*)	0.01655
EBLUP (with *both*)	0.01891
GCDP-WDP (Case I)	0.00724
GCDP-WDP (Case II)	0.00706
GCDP-WDP (Case III)	0.00679

Table 6.5: Average of AEMSE for BFRSS study for $n_j = 10$

Chapter 7

Summary

Sample surveys provide a cost effective way of obtaining estimates for characteristics of interest at both population and subpopulation, i.e. domain, level. In most practical applications, however, domain sample sizes are not large enough to allow direct estimation. When direct estimation is not possible, one has to rely upon alternative methods that depend on the availability of population-level auxiliary information and are commonly referred to as indirect or model-based methods. Model-based methods can be classified into two categories, namely methods based on fixed effects models, i.e. models that explain between-area variation in the target variable using only the auxiliary information, and methods based on mixed random effects models that include area-specific random-effects to account for between-area variation beyond that explained by the auxiliary information. However, mixed effects models depend on parametric and distributional assumptions as well as requiring specification of the random part of the model, which may not be satisfied by the real data.

For making inferences, both indirect and model-based methods often assume that direct estimators are available. In this dissertation, a simple simulated example was used to illustrate that these traditional estimators will become problematic when the direct estimator is either not available or not reliable.

In this dissertation, we proposed a Bayesian noninformative approach to joint small area estimation, which is able to utilize various kinds of auxiliary information (including area-specific, element-specific and across-area information) without assuming a linear relationship between variables. The step-wise Bayesian characteristic ensures that our estimator is admissible. In addition, our estimator does not depend on direct estimators, and the simulations showed that this estimator outperforms common alternatives when sample size is small. An application to a BRFSS study showed that, comparing to traditional small area estimation strategies, the proposed Bayesian approach is a better choice in terms of the absolute error and mean square error of point estimation.

We utilized the Metroplis-Hastings algorithm to compute our estimators approximately. The gain of our Bayesian methodology is significant, although it becomes computionally intensive for continuous populations. However, the heavy computational load is reduced when all variables are categorical.

Bibliography

[1] Basu, D. (1969). "Role of Sufficiency and Likelihood Principles in Sample Survey Theory". *Sankya B*, 31, 411-454.

[2] Berger, J.O. and Chen, M.-H. (1993). "Predicting Retirement Patterns: Prediction for A Multinomial Distribution with Constrained Parameter Space". *The Statistician*, 42(4), 427-443.

[3] Chen, J. and Sitter, S.S. (1999). "A Pseudo Empirical Likelihood Approach to The Effective Use of Auxiliary Information in Complex Surveys". *Biometrika*, 80, 107-116.

[4] Chen, M.-H. and Schmeiser, B.M. (1993). "Performance of the Gibbs, Hit-and-run and Metroplis Samplers". *Journal of Computational and Graphical Statistics*, 2, 251-272.

[5] Chen, M.-H. and Schmeiser, B.M. (1996). "General Hit-and-Run Monte Carlo Sampling for Evaluating Multidimensional Integrals". *Operations Research Letters*, 19, 161-169.

[6] Datta, G.S., Fay, R.E. and Ghosh, M. (1991). "Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation", in *Proceedings of Bureau of the Census 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, pp. 63-79.

[7] Deng, L. and Chhikara, R.S. (1990). "On The Ratio and Regression Estimation in Finite Population Sampling". *The American Statistician*, 44(4), 282-284.

[8] Devroye, L. (1986). *Non-uniform Random Variate Generation*. Spring-Verlag.

[9] Fay, R.E. and Herriot, R.A. (1979). " Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data". *Journal of the American Statistical Association*, 74, 269-277.

[10] Gamerman, D., and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. 2nd Edition. Chapman and Hall.

[11] Geyer, C.J. (2008). Homepage for Locating the R Library "rcdd" for Vertex Enumeration. *http://www.stat.umn.edu/geyer/rcdd*.

[12] Ghosh, J.K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays*. By (Ed. D. Basu). New York: Springer-Verlag.

[13] Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London.

[14] Ghosh, M., and Rao, J.N.K. (1994). "Small Area Estimation: An Appraisal". *Statistical Science*, 9, 55-93.

[15] Givens, G.H. and Hoeting, J.A. (2005). *Computational Statistics*. Wiley-Interscience.

[16] Gonzalez, M.E. (1973). "Use and Evaluation of Synthetic Estimates", Proceedings of the Scocial Statistics Section, *American Statistical Association*, pp. 33-36

[17] Harville, D.A. (1991). Comment. *Statistical Science*, 6, 35-39.

[18] Kostanich, D.L. and Dippo, C.S. (2002). "Design and methodology: 63rv". Technical report, The U.S. Census Bureau and The Department of Labor Statistics.

[19] Lazar, R. (2005). "Computations for Bayesian Procedures under Linear Constraints in Finite Population Sampling and Imprecise Probability". Ph.D. dissertation, School of Statistics, University of Minnesota.

[20] Lazar, R., Meeden, G. and Nelson, D. (2005). "A Noninformative Bayesian Approach to Domain Estimation". *Journal of Statistical Planning and Inference*, 34, 51-64.

[21] Lazar, R., Meeden, G. and Nelson, D. (2008). "A Noninformative Bayesian Approach to Finite Population Sampling Using Auxiliary Variables". *Survey Methodology*, 34, 51-64.

[22] Lohr, S.L., and Rao, J.N.K. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury.

[23] Meeden, G. (1999). "A Noninformative Bayesian Approach for Two Stage Cluster Sampling". *Sankhyā*, Series A, 61, 133-134.

[24] Morris, C.A. (1983). Parametric Empirical Bayes Inference: Theory and Applications, *Journal of the American Statistical Association*, 78, 47-54.

[25] Prasad, N.G.N., and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimation, *Journal of the American Statistical Association*, 85, 163-171.

[26] Rao, J.N.K. (2003). *Small Area Estimation*. Wiley-IEEE.

[27] Rubio V. and Salvati N. (2007). "Introduction to Small Area Estimation". *http://www.bias-project.org.uk/software/* for R library "SAE".

[28] Smith, T.M.F. (1983). "On the Validity of Inference from Non-Random Samples". *Journal of the Royal Statistical Society*. Series A, 146, 394-403.

[29] Strief, J. (2007). "Bayesian Sampling Weights: Toward a Practical Implementation of the Polya Posterior". Ph.D. dissertation, School of Statistics, University of Minnesota.

[30] Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.