ABSTRACT

Title of Dissertation: Advancements in Small Area Estimation Using Hierarchical Bayesian Methods and Complex Survey Data

Soumojit Das
Doctor of Philosophy in Applied Statistics, 2024
*A*MSC, University of Maryland, College Park

Dissertation Directed by: Prof. Partha Lahiri
Department of Mathematics & JPSM
*U*niversity of Maryland, College Park

This dissertation addresses critical gaps in the estimation of multidimensional poverty measures for small areas and proposes innovative hierarchical Bayesian estimation techniques for finite population means in small areas. It also explores specialized applications of these methods for survey response variables with multiple categories. The dissertation presents a comprehensive review of relevant literature and methodologies, highlighting the importance of accurate estimation for evidence-based policymaking.

In Chapter 2, the focus is on the estimation of multidimensional poverty measures for small areas, filling an essential research gap. Using Bayesian methods, the dissertation demonstrates how multidimensional poverty rates and the relative contributions of different dimensions can be estimated for small areas. The proposed approach can be extended to various definitions of multidimensional poverty, including counting or fuzzy set methods.

Chapter 3 introduces a novel hierarchical Bayesian estimation procedure for finite population means in small areas, integrating primary survey data with diverse sources, including social media data. The approach incorporates sample weights and factors influencing the outcome variable to reduce sampling informativeness. It demonstrates reduced sensitivity to model misspecifications and diminishes reliance on assumed models, making it versatile for various estimation challenges.

In Chapter 4, the dissertation explores specialized applications for survey response variables with multiple categories, addressing the impact of biased or informative sampling on assumed models. It proposes methods for accommodating survey weights seamlessly within the modeling and estimation processes, conducting a comparative analysis with Multilevel Regression with Poststratification (MRP).

The dissertation concludes by summarizing key findings and contributions from each chapter, emphasizing implications for evidence-based policymaking and outlining future research directions.

Advancements in Small Area Estimation Using Hierarchical Bayesian Methods and Complex Survey Data

by

Soumojit Das

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Prof. Partha Lahiri, Advisor
Prof. Ansu Chatterjee
Prof. Cinzia Cirillo, Dean's representative
Dr. Jean Opsomer
Prof. Jörg Drechsrter
Prof. Yan Li

Dedication

To my beloved late grandparents. Your doting care, boundless benevolence, and fervent faith in the good of people continue to be my guiding light.

Acknowledgments

One does not simply begin this section without thanking their academic parent. To my academic father, Prof. Partha Lahiri, I sincerely owe you a debt of peerless gratitude. Thank you for taking me under your wings and teaching me to think critically and independently. Always accommodating my persistent requests to review my clumsy algebra, with a smile on your face, and instantly spotting mistakes just by glancing at them. Your sheer experience and algebraic erudition spared me hours of futile derivations. I am also immensely grateful for the countless times you bought me lunch and dinner; as a perpetually broke UMD grad student, I always appreciated that gesture. I thank both you and Ranita ma'am for generously letting me take home leftovers from your dinner parties, sustaining me for the next couple of days without the worry of feeding myself.

I am profoundly indebted to my RA-supervisor, Prof. Jörg Drechsler. You have been more than just an academic advisor and committee member; you have imparted invaluable life lessons beyond the confines of academia. Your guidance and support have been a cornerstone of this journey.

To my parents, Maa and Baba, nothing my peanut-sized smooth brain can surmount in words will ever be adequately expressive of what I really want to convey. Thank you for bringing me into this world and for your unquestioning faith in my career choices. Your support, though more financial than emotional, has been the one singular constant throughout my entire life. I

appreciate your trust in me and for never questioning my decisions. I never required emotional inputs, and you never tried to impose them on me—perhaps that's why we get along so well!

I would also like to extend my heartfelt thanks to my friends, both within and outside of UMD and academia. You've been the glue that kept this journey together. The endless chats, banters, inside-jokes, and engagements in mutual proclivities kept my spirits high. Sincerest salutations to Aniket Da, Anurag, Babua, Charlie, Debankur, Debdipta, Kangrui the Soft, Max, Nate the Dawg, Rahul, Sungjoo, Travis, Tyler the Tiny, Yilin, Yohan, Yuting, Yufan & Zuping, and to those who are deliberately left unnamed, who have inspired me though diverse facets and phases of my life. Some of them are now estranged, by choice or by circumstances. Oh, dissonance, why must you be as relentless as time? Nonetheless, I am (hu)man enough to appreciate the impact they had on me, however fleeting or long-lasting.

My sincere appreciation goes to the staff at JPSM. Although I was an AMSC student by admission, JPSM has been my work-home for the past few years. Your support, friendliness, and assistance have made my time here incredibly enriching and enjoyable. I am grateful for the welcoming environment and the opportunities for growth that JPSM provided.

Ahh, yes, Tool. Tool! You have been my sanctuary. Your music has kept my academic life on track and helped me make sense of the drivel that is my personal life. You kept me afloat through the undertows of deadlines, crunch times, and moments of tumbling in and out of love. The relevance of your music at every phase of my journey is genuinely astounding!

Thanks to NSF for funding the research that forms the basis of the second chapter.

Lastly, I extend my love and gratitude to the city of College Park—my home away from home during these last few and significant years. You provided me with a sense of belonging and allowed me to choose the family we choose for ourselves, our friends. My only hope is that your

rapid gentrification doesn't end up confining grad students' diet to ramen and canned foods.

May serenity, tranquility, and fulfillment be unto all!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AMSC	The Applied Mathematics & Statistics, and Scientific Computation
DEFF	Design Effect
EB	Empirical Bayes
ESS	Effective Sample Size
HB	Hierarchical Bayes
JPSM	Joint Program in Survey Methodology
MCMC	Markov Chain Monte Carlo
MRP	Multilevel Regression with Post-stratification
NEF-QVF	Natural Exponential Family with Quadratic Variance Functions
NSF	National Science Foundation
SAE	Small Area Estimation
UMD	University of Maryland, College Park

Chapter 1: Introduction

1.1 Why Small Area Estimation?

Historically, sample surveys have been cost-effective tools for gathering information on a wide array of topics, ranging from socioeconomic indicators to health outcomes and market behavior. They are widely utilized to generate estimates not only for the overall population of interest but also for a diverse range of subpopulations (or domains) across different classifications. These domains may encompass geographic areas, socio-demographic groups, or other specified subsets (or subpopulations). For instance, geographic domains may include regions such as urban neighborhoods, rural townships, school catchment areas, postal code zones, census tracts, and voting precincts. Conversely, socio-demographic domains may delineate specific age-income-education brackets within larger geographic regions. Additionally, "other domains" could involve categorizations like sets of households categorized by housing tenure or employment status within a given locality.

In sample survey literature, domain estimation employs "direct" estimators, which rely solely on domain-specific sample data. These estimators can utilize auxiliary information, such as known totals of relevant variables. Typically, direct estimators are design-based, leveraging the sampling design's probability distribution while holding population values at a constant. Alternatively, they can be "model-assisted", incorporating "working" models to enhance "robustness"

against potential model misspecifications. Domains, or areas, are categorized as either large or small based on their sample sizes. Large domains possess ample samples for precise direct estimates, whereas small domains lack sufficient sample size for accurate direct estimation. In certain cases, domains of interest may have zero sample size, presenting challenges for traditional design-based estimation. We designate domains as "small areas" when reliable direct estimates cannot be produced for them. "Small Area Estimation" (hereafter SAE) is a collection of statistical methods utilized in sample surveys to provide precise estimates for "small areas" when traditional survey methods face limitations due to limited (or no) sample sizes [4].

SAE relies on two primary data sources: sample survey data and auxiliary information. Sample survey data are collected through traditional survey methods, offering insights into broader populations but often lacking sufficient sample sizes for smaller areas. Auxiliary information includes additional data sources, such as administrative records or remote sensing data, which provide insights at both small and large geographic levels.

Various statistical methods are employed in SAE, including direct estimation, indirect estimation, and model-based approaches–both from frequentist and Bayesian viewpoints. These methods enable researchers to combine information from sample surveys with auxiliary data to produce more accurate estimates for small areas. SAE finds applications in official statistics, public policy, market research, and environmental monitoring. It empowers decision-makers with localized insights, aiding in the formulation of policies and programs, allocation of government funds, and targeting of resources to areas in need. For instance, governments can use small area estimates to identify regions with higher poverty rates for targeted poverty alleviation programs or allocate healthcare resources to areas with higher disease prevalence. For further examples, readers are referred to [4], [5], [6].

1.2 Direct estimation

In sample surveys, direct estimation refers to the set of methods used to estimate the population parameter of interest (such as totals, means, etc.), θ, only from the sampled values of the variable of interest, y. There are a great many books on survey sampling that provide extensive treatment of direct estimation–like Cochran's seminal work [7], and Lohr's recent work [8]. In this review, we will mainly be talking about direct estimation for small areas.

Let $\mathcal{U}$ represent the index set associated with the (finite) population, or the universe of interest in a sample survey, comprising N distinct elements:

$$\mathcal{U} = \{j : j = 1, \cdots, N\}.$$

We denote the survey characteristic of interest as $y_1, \cdots, y_N$. This assumes that, for the variable of interest, the entire finite population is observable and measurable without any error (free of measurement error). Now, let us say that we want to estimate the following parameters of interest: the population total, $Y = \sum_{j \in \mathcal{U}} y_j$, or the population mean $\bar{Y} = Y/N$. To do this, we can employ a sampling design to select a subset s of $\mathcal{U}$, called a sample, with probability (hereafter w. p.) $p(s)$. In practice, $p(s)$ ultimately depends on the sampling scheme that was used to collect the data. Simple Random Sampling (SRS), Stratified Sampling, Cluster Sampling, Stratified Multi-Stage Sampling, etc.

With the assumption of complete response of the units $j \in s$ selected in the sample, s, an

estimator of Y, $\hat{Y}$ is said to be design-unbiased (or p-unbiased) if,

$$E_p(\hat{Y}) = \sum_{s \in \mathcal{P}(s)} p(s)\hat{Y}_s = Y, \tag{1.1}$$

where $\mathcal{P}(s)$ is the superset of all possible samples s under the specified sampling design, and $\hat{Y}_s$ is the value of $\hat{Y}$ calculated for a sample $s \in \mathcal{P}(s)$. We can follow similar notions to define a design-consistent estimator. For this, first define the design-variance of $\hat{Y}$ as:

$$V_p(\hat{Y}) = E_p[\hat{Y} - E_p(\hat{Y})]^2.$$

An estimator of this variance $V_p(\hat{Y})$ is denoted as: $v(\hat{Y}) = s^2(\hat{Y})$. We call this variance estimator p-unbiased if,

$$E_p[v(\hat{Y})] \equiv V_p(\hat{Y}).$$

The bias of the estimator $\hat{Y}$ in estimating Y is denoted as:

$$Bias(\hat{Y}) = E_p(\hat{Y}) - Y.$$

Now, we call $\hat{Y}$ a design-consistent (or p-consistent) estimator of Y, if, $\hat{Y}$ is approximately unbiased, and $V_p(\hat{Y}) \to 0$, as $N \to \infty$ [4]. Interested readers can refer to [9] for an even more extensive review of design-consistency. Even though we considered only one population, such concepts require consideration for a sequence of populations. We omit such discussions here, but [4] considers this aspect as well.

Let us now focus on how direct estimation usually works out for small areas using survey

weights, also called design weights. They play a vital role in constructing design-based estimators. Suppose there are m small areas indexed by $i = 1, \cdots, m$. Let U_i and s_i denote index sets of the finite population and the sampled units from that finite population, respectively, for each area $i = 1, \cdots, m$. We can think of these sets as the mutually exclusive (but not necessarily exhaustive) subsets of the $\mathcal{U}$ and s, respectively, using the notations used thus far. We also denote the cardinalities of these two index sets as $N_i = |U_i|$, and $n_i = |s_i|$ for the m areas.

Let y_{ik} denote the outcome of the survey variable of interest for the k^{th} unit within the i^{th} small area, conducted using the sample from the finite population using a possibly complex sampling design. Our parameters of interest are the m small area population totals:

$$Y_i = \sum_{k=1}^{N_i} Y_{ik},\ i = 1, \cdots, m. \tag{1.2}$$

Population means can also be similarly defined as: $\bar{Y}_i = Y_i/N_i$ for m small areas. Define the sampling weight of the sampled unit k within the area i, as $w_{ik},\ i = 1, \cdots, m,\ k = 1, \cdots, n_i$. Given any sampling design (simple or complex), these sampling weights (also referred to as design weights) are the inverse of the first-order inclusion probabilities π_{ik}: the probability that the k^{th} unit within the i^{th} small area is included in the sample s_i for that area i [1]. Thus,

$$w_{ik} = \frac{1}{\pi_{ik}}. \tag{1.3}$$

This sampling weight can be interpreted as the number of population units from U_i represented by this one sample from that area i. Now, depending on the sampling scheme, these first-order

[1] We emphasize the distinction between the usual notation of second-order inclusion probability π_{ij}, which is defined as the probability that both the i^{th} and the j^{th} units of a finite population are selected in the same sample, $i \neq j$ [7].

inclusion probabilities can either be equal or not. When they are equal, that sampling scheme is referred to as an 'Equal Probability of Selection Method' design, or in short, an EPSEM design. This essentially means that under an EPSEM design, every unit or element in the population has an equal probability of being selected into the sample. Simple Random Sampling (with or without replacement) is an example of an EPSEM design.

Now, for a small area i, with the corresponding sample size $n_i > 0$, and under an EPSEM design, the sample totals:

$$y_i = \sum_{i=1}^{n_i} y_{ik},\ i = 1, \cdots, m$$

will be unbiased estimators of the population totals Y_i in 1.2. Similarly, $\bar{y}_i = y_i/n_i$ will be unbiased estimators for the small area population means $\bar{Y}_i$. When the sampling scheme is not EPSEM, and the different units within an area have different selection probabilities, naturally, their corresponding weights must vary within the small areas as well. Under such situations, survey practitioners resort to using the sampling weights as defined in 1.3 to estimate the population totals (and means) for the small areas. The small area population total estimators are of the form:

$$y_i^{(w)} = \sum_{k=1}^{n_i} w_{ik} y_{ik},\ i = 1, \cdots, m, \tag{1.4}$$

and the corresponding small area population mean estimators are of the form:

$$\bar{y}_i^{(w)} = \frac{\sum_{k=1}^{n_i} w_{ik} y_{ik}}{\sum_{i=1}^{n_i} w_{ik}},\ i = 1, \cdots, m. \tag{1.5}$$

Note that these mean estimators are simply weighted means of the y_{ik}'s, with the weights be-

ing the survey (or design) weights w_{ik}. This approach considers the complex survey design and accounts for unequal probabilities of selection, nonresponse, and other survey features to produce design-based estimates for small area parameters. The domain mean estimator (1.5) is a ratio. Thus, linearization-based variance estimators may be used for estimating its variance and associated standard error [8]. On the other hand, various other variance estimators may be constructed using replicate weights-based methods. These include Balance Repeated Replication (BRR), Jackknife methods, Bootstrap methods, etc. The replicate-based methods are partly used for privacy reasons. For a comprehensive review of these methods, please see [10] and [11].

Design-based methods in survey sampling offer strengths in providing unbiased estimates that properly account for the survey design (often complex) features such as stratification, clustering, and unequal probabilities of selection. These methods allow for robust statistical inference by facilitating the calculation of standard errors, confidence intervals, and hypothesis tests without relying on specific distributional assumptions about the data. However, they are sensitive to assumptions about the underlying population distribution and the behavior of survey weights, and violations of these assumptions can lead to biased or inefficient estimates. Additionally, design-based methods may struggle with inefficiency, particularly with small sample sizes, resulting in wider confidence intervals and less precise estimates of means and standard errors. Please refer to the book by [4] for examples.

Model-based approaches can mitigate some of the weaknesses associated with design-based methods. These approaches offer greater flexibility in capturing complex relationships and incorporating auxiliary information, allowing for more accurate estimation, especially in situations where assumptions or design features may limit design-based methods. Model-based approaches can also provide more efficient estimates by leveraging information from auxiliary

variables and modeling complex relationships, particularly in small areas where sample sizes are limited. Furthermore, model-based approaches are better equipped to handle nonresponse and missing data, as they can incorporate imputation techniques or model nonresponse mechanisms to mitigate bias and improve the reliability of estimates.

1.3 Model-based approaches for small domains

Direct estimates for smaller domains may exhibit considerable variability when aggregated to a higher level, potentially resulting in significant deviations from the corresponding direct estimate for the higher level. This discrepancy is primarily due to the inherent sampling variability and limited sample sizes associated with smaller areas. When aggregated, these variations can either amplify (or attenuate), leading to estimates that may not accurately represent the true population characteristics at the higher level. This presents a critical challenge for statisticians and practitioners alike. For example, policymakers and decision-makers who rely on accurate data for resource allocation and policy formulation.

In addressing this issue, benchmarking serves as a crucial strategy to enhance the reliability and validity of small area estimates [12]. Benchmarking involves the comparison of direct estimates with external or auxiliary information, such as administrative records or census data, which are often available at higher levels or for larger geographic areas–and these are usually very reliable. By aligning direct estimates with benchmark data, one can identify and adjust for discrepancies, thereby improving the accuracy and consistency of estimates at lower levels. This approach helps to reduce the impact of variability in direct estimates and ensures that estimates align more closely with known population characteristics.

Small area model-based methods complement benchmarking by incorporating auxiliary information and sophisticated statistical models (refer to [6] for an excellent review) to enhance estimation accuracy further. For some situations, by leveraging information from neighboring areas and modeling spatial dependencies ([13]), these methods can provide more stable and "robust" [2] estimates for small domains. Moreover, some model-based approaches (like hierarchical Bayesian models with Markov Chain Monte Carlo) facilitate the integration of benchmark data into the estimation process, allowing for the refinement and validation of estimates based on external sources. We present such examples of benchmarking in chapters 2 and 3 of this thesis, albeit for two different applications.

The use of explicit models offers several advantages over direct design-based methods: (i) we can employ model diagnostics–such as residual analysis facilitating detection of model violation, choosing appropriate set of covariates deemed useful for the given data and the model assumed, detecting influential observations (small domains for the case of SAE), etc. (ii) Area or domain-specific measures of precision, or uncertainties can be associated with each small area. This results in better attribution, along with precise estimation. (iii) Generalized Linear Models can be utilized, accounting for (random) errors at different levels. Complex data structures, such as spatial correlations and time series, can also be handled. (iv) Random effects can also be incorporated into clever formation of modeling. In this section, we will examine key model-based methods frequently utilized in Small Area Estimation (SAE) literature. Our emphasis will primarily be on Bayesian estimation and inferential techniques, aligning with the hierarchical Bayesian methodology central to our thesis.

[2] robustness here is not be confused with the traditional robust statistical methods in the literature. More often than not, we will use this term in this thesis to mean robustness towards model mis-specifications, stability in terms of narrower confidence or credible intervals, etc.

In constructing a model appropriate for a given small area estimation problem, the availability of administrative data presents various scenarios to consider [14]. Administrative records may exist in different forms, including summaries for geographical areas containing one or more small areas, summaries specific to the small areas themselves, summaries for geographical areas entirely contained within the small areas of interest or unit-level records. Despite the complexity inherent in hierarchical modeling, its flexibility allows for the integration of information from diverse sources, such as surveys, administrative records, and census data available at different geographic levels. The specifics of the hierarchical model depend on the nature and accessibility of data from these sources. The models we review here can broadly be classified into two categories: (i) Models that relate the "true" (unknown and to be estimated) small area means with available auxiliary variables. These are called the "area level" or "aggregate level" models. (ii) Models that relate the unit level information of the study, or survey variable of interest, to unit-specific auxiliary variables. These are referred to as the "unit level" models. Such models often incorporate area-level auxiliary variables as well and are very flexible in terms of pooling information from different sources.

1.3.1 Area level models

Area level models present one improvement over direct survey estimators by introducing hierarchical models that assume a relationship between small area parameters of interest and auxiliary information, in terms of covariates, in the modeling step. These models offer the advantage of requiring only area-level aggregate summaries of covariates from administrative records, eliminating the need for unit-level covariates at either the sample or population level. Early ap-

plications of area-level models in small area estimation can be traced back to works by [15], [16], and [17]. While these authors implicitly utilized area-level covariates by grouping similar small areas, they did not explicitly incorporate them into the modeling process or utilize complex survey data.

Notations: Let there be $i = 1, \cdots, m$ small areas in the finite population of consideration. Let y_i be the direct survey or design-based estimate of the "true" population parameter of interest for the area i, θ_i. Assume that the sampling variances of such estimates, y_i, are known, and let us call them D_i. Since the area-level sample sizes, n_i, are typically quite small (sometimes there may even be no samples, $n_i = 0$), the direct survey-weighted estimates are subjected to non-negligible sampling variability. Model based approaches are often considered to remedy such shortcomings of the design-based approaches. We review 3 such models in the following.

Fay Herrior (FH) model: The earliest area level model was proposed by Fay and Herriot in their seminal work on small area estimation [1]. They considered a generalized version of the Efron-Morris and Carter-Rolph type Bayesian models in the context of estimating per-capita income for small domains (with some of the population sizes $< 1,000$). This model was used to create empirical Bayes estimates of poverty rates. For small areas $i = 1, \cdots, m$:

$$\begin{aligned} \text{Level 1 (sampling model):} \quad & y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i) \\ \text{Level 2 (linking model):} \quad & \theta_i|\beta, A \overset{ind}{\sim} N(x_i'\beta, A) \end{aligned} \tag{1.6}$$

where x_i is a $p \times 1$ vector of known covariates available for the m small areas. D_i's are the known sampling variances. The hyperparameters β–a $p \times 1$ vector, and A are unknown parameters to be estimated from the given model and the data. The two levels of the FH model can be combined into one linear mixed model, and thus, this model is an example of a matched model.

$$y_i = x_i'\beta + v_i + \epsilon_i \tag{1.7}$$

where v_i and ϵ_i are independent with $v_i \sim N(0, A)$ and $\epsilon_i \sim N(0, D_i)$ for $i = 1, \cdots, m$. Applying this model to a proportion estimation problem, which is of special interest to us in this thesis, can be troublesome–the assumption for normality cannot guarantee that the support stays between 0 and 1. Likewise, the normality assumption does not guarantee the support between 0 and 1 for the posterior distribution of θ_i. Different modeling options for the sampling level and the linking level for the area-level model are studied and discussed in a lot of different research works, like [2], [3].

Next, we consider a class of models called unmatched models (in contrast to the FH model, which is an example of a matched model), where other non-normal distributions with support $(0, 1)$ may be used at the linking level.

Normal-Logistic (NL) model: The Normal-Logistic (NL) model for modeling proportions or probabilities can be considered by changing the level 2 under the FH model in 1.6, using a logit transform [3]. This facilitates the modeling of the transformed proportions as a linear combination

[3]The logit transformation on $\theta \in (0, 1)$ is defined as: $\text{logit}(\theta) = \log_e(\frac{\theta}{1-\theta})$

of the auxiliary information. For small areas $i = 1, \cdots, m$:

$$\begin{aligned} \text{Level 1 (sampling model):} \quad & y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i) \\ \text{Level 2 (linking model):} \quad & \text{logit}(\theta_i)|\beta, A \overset{ind}{\sim} N(x_i'\beta, A). \end{aligned} \tag{1.8}$$

The sampling variances D_i's are not estimable from the aggregated area-level data $\{(y_i, x_i), i = 1, \cdots, m\}$. For the standard implementation of the FH and NL models, the sampling variances D_i are estimated using additional design information (within small areas), but errors due to the estimation of sampling variances are generally ignored in the subsequent inferences. This is a well-known deficiency of using an area-level model compared to that of a unit-level model. Even then, these area-level models are widely used in practice because aggregate statistics can usually be modeled using simpler distributional structures than individual observations. On top of that, the aggregate models offer a natural way to incorporate survey weights in the model hierarchy, which aids in accounting for the design-based features needed to be captured by the hierarchical Bayes estimators.

Next, we illustrate the Normal-Logistic random sampling variance model (NL_{rs}) model. This can be viewed as an extension of the model proposed by [3] and used in this present form by [2] in the context of analyzing smoking data from the 2008 NHIS (National Health Interview Survey).

Normal-Logistic random sampling variance (NL_{rs}) model: The Normal-Logistic random sampling variance (NL_{rs}) model can be obtained from the regular NL model in 1.8 by augmenting the sampling variances at Level 1 of both FH and NL models given in 1.6, and 1.8. For

small areas $i = 1, \cdots, m$, the model is given by:

$$\begin{aligned} &\text{Level 1 (sampling model):} && y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i = \frac{\theta_i(1-\theta_i)}{n_i}\text{DEFF}) \\ &\text{Level 2 (linking model):} && \text{logit}(\theta_i)|\beta, A \overset{ind}{\sim} N(x_i'\beta, A) \end{aligned} \tag{1.9}$$

where, 'DEFF' is the true design effect and is defined by the ratio of the variances under the given design, to the variance under Simple Random Sampling (SRS).

$$\text{DEFF} = \frac{\text{Var}_{design}}{\text{Var}_{SRS}},$$

where Var_{design} is the true sampling variance of the survey-weighted estimates y_i's under the given (possibly a complex) sampling design, and Var_{SRS} is the true sampling variance of an unweighted estimate under SRS of size n_i with replacement. The quantity $\frac{n_i}{\text{DEFF}}$ is often called the effective sample size (ESS) of area i, and denoted by $n_{i;eff}$. This concept is useful in assessing the amount of independent information (sample size) from a complex survey sample. There are ways of calculating n_{eff} for a given problem. The most commonly used way is to calculate the 'DEFF', the design effect first, and then use the following formula to get the effective sample size (n_{eff}):

$$n_{eff} = \frac{n}{\text{DEFF}}. \tag{1.10}$$

A sample of size n_{eff} from an SRS would be expected to give the same precision as the n observation units taken in the complex sample [8]. We present one novel way of calculating this $n_{i;eff}$, for each small area, when we apply these area level models to a specific small area estimation problem concerning a complex survey data in chapter 2. To maintain brevity in this

review, we conclude our examination of area level models. Interested readers are encouraged to explore the extensive collection of area-level models and their applications presented in this book by [4].

1.3.2 Unit level models

In an area level model, the uncertainty arising from the estimation of the sampling variances (D_i) for the direct estimates (y_i) cannot be directly addressed. To incorporate this additional uncertainty, it becomes necessary to model observations within each small area individually. Let y_{ik} be the k^{th} unit within an area i; $i = 1, \cdots, m$, and $k = 1, \cdots, n_i$. Where n_i is the sample size for the area i. Such y_{ik}'s are referred to as the 'units' in this context. The unit level models require availability of some unit level covariates, along with the usual area-level covariates. Unit level covariates are not necessary for these models, because the area specific random effects can still capture the presence of any correlation between the units within a given small area. A basic unit-level model, incorporating only area-specific covariates, can be structured hierarchically as follows:

$$\begin{aligned} \text{Level 1 (sampling model):} \quad & y_{ik}|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma_\epsilon^2) \\ \text{Level 2 (linking model):} \quad & \theta_i|\beta, A \overset{ind}{\sim} N(x_i'\beta, \sigma_v^2). \end{aligned} \tag{1.11}$$

This can also be re-expressed in the form of a linear mixed model:

$$y_{ik} = x_i'\beta + v_i + \epsilon_{ik}, \tag{1.12}$$

where the area specific random effects $\{v_i\}$ and the unit specific random errors, or sampling errors $\{\epsilon_{ik}\}$ are mutually independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, and $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. x_i is a $p \times 1$ vector of known area level covariates. σ_v^2 and σ_ϵ^2 can be interpreted as the between and within area unknown variance components, respectively.

The level 1 of the hierarchical model described in 1.11 inadvertently assumes exchangeability by treating the units within each small area as being exchangeable. Exchangeability implies that the order in which the units are observed within a small area does not affect the statistical properties of the model. In other words, the model assumes that the units within a small area are statistically identical or interchangeable with each other, allowing for the estimation of common parameters across units within the same area. This is a strong assumption in most of the practical scenarios. Such strong modeling assumption can be relaxed in presence of unit specific auxiliary information for the sampled units, and some aggregated summaries for the non-sampled units for each of the small areas. [18] considered such a model. This is an early application of a unit level model, which is also a nested error regression model. They proposed this model to predict areas under corn and soybean for 12 counties of North-Central Iowa using the 1978 June Enumerative Survey (JES) as the primary survey data and satellite (LANDSAT) data as the source of auxiliary variables. Refer to their paper [18] for further details. They considered the following nested error regression model:

$$y_{ik} = x_{ik}'\beta + v_i + \epsilon_{ik}, \tag{1.13}$$

where $i = 1, \cdots, m$, $k = 1 \cdots N_i$, and x_{ik} is $p \times 1$ vector of known covariates available for every sampled units. The area specific random effects $\{v_i\}$ and the sampling errors $\{\epsilon_{ik}\}$ are mutually independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, and $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. Such a modeling will naturally assume

that this model will hold for all the units of the finite population. This is a common practice for model-based small area estimation. In chapter 3, we propose an alternate way of modeling sampled and unsampled units of the finite population. The model can be posed in a hierarchical structure as well:

$$\begin{aligned} &\text{Level 1:} \quad y_{ik}|v_i \overset{ind}{\sim} N(x'_{ik}\beta + v_i, \sigma^2_\epsilon) \\ &\text{Level 2:} \quad v_i \overset{iid}{\sim} N(0, \sigma^2_v), \end{aligned} \tag{1.14}$$

for $i = 1, \cdots, m$, and $k = 1, \cdots, N_i$.

The unit models discussed thus far, that is in 1.12 and 1.13; and their hierarchical counterparts in 1.11 and 1.14 can produce design-consistent estimators (discussed previously in 1.2), only for a simple survey design–like SRS. This is because these models do not have the provision for incorporating features of a more complex survey design, such as stratification and clustering information. Such specific information are also not easily available for the publicly available datasets, mainly for privacy issues. But, they contain survey weights, which are usually constructed keeping all the said complex features of the survey in mind. But the aforementioned models cannot include survey weights either–because these models are typically posited on the entire finite population, and the survey weights are available only for the sampled units. Thus, more complex models are warranted for handling data from complex survey designs. [19] considered a two-fold nested error regression model for a certain survey data collected using a stratified two-stage cluster sampling, where the strata were the small areas. Their model is given as fol-

lows:

$$y_{ijk} = x'_{ijk}\beta + v_i + u_{ij} + \epsilon_{ijk};\ i = 1, \cdots m,\ j = 1 \cdots M_i,\ k = 1, \cdots N_{ij}, \tag{1.15}$$

where y_{ijk} is the response to the survey variable of interest for the k^{th} unit, within the j^{th} primary sampling unit (PSU, also the clusters in this case), within the i^{th} small area (strata). x_{ijk} is the corresponding $p \times 1$ vector of auxiliary variables. $\{v_i \overset{iid}{\sim} N(0, \sigma_v^2)\}$ are the random area specific effects, $\{u_{ij} \overset{iid}{\sim} N(0, \sigma_u^2)\}$ are the random within-area PSU effects, and $\{\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)\}$ are the random sampling error. $\{v_i, u_{ij}, \epsilon_{ijk}\}$ are also mutually independent. The model 1.15 is assumed on all the units of the finite population and is considered to hold for the sampled units as well. This is only generally true for the case of simple random sampling of clusters (PSU) and the subunits within the selected clusters (secondary selection units, or SSU). Typically, this holds for sampling designs that make use of the auxiliary covariate vectors x_{ijk} in the selection scheme. Different extensions of this model 1.15 have been considered by different authors. For example, [20] studied the case of $x'_{ijk}\beta = \beta,\ \forall\ i, j, k.$ [21] considered only cluster specific covariates $x'_{ijk}\beta = x'_{ij}\beta, \forall\ k = 1, \cdots N_{ij}$. Further extensions with useful examples are discussed in [4].

Generalized linear mixed effects unit level models: The models explored thus far in small area estimation (SAE) are primarily designed for continuous survey response variables. However, in applied survey statistics, binary and count response variables are frequently encountered. For such applications within SAE, generalized linear models (GLM) prove particularly useful. For an excellent overview of GLM, please refer to the book by [22].

In model-based inferences within SAE, GLM assumes independence of the responses.

However, this assumption often does not hold in real-life situations due to potential correlations among observations. For instance, in the two-fold nested error regression model used for two-stage cluster sampling (see Model 1.15), the responses of units within a cluster tend to be correlated due to shared geographical characteristics. This correlation is a common occurrence in multi-stage cluster sampling. In such cases, a simple GLM with fixed effects may not effectively capture this correlation. Generalized linear mixed-effects models (GLMM) provide a solution in such scenarios. Suppose that $y_{ik} = 0,$ or 1 be the binary outcome variable of interest for the k^{th} unit within the i^{th} small area, and let x_{ik} be the corresponding $p \times 1$ unit specific vector of auxiliary information; $i = 1, \cdots, m;\ k = 1, \cdots, n_i$. The usual parameters of interest are the small area proportions:

$$\bar{Y}_i = \frac{\sum_{k=1}^{N_i} Y_{ik}}{N_i} = P_i\ ,\ \text{say},\ i = 1, \cdots, m.$$

In this context, [23] assumed the $y_{ij}|\theta_{ij} \overset{iid}{\sim}$ Bernoulli(θ_{ij}). Then they used a logit link function to relate the transformed θ_{ij}'s with the known unit level covariates. The model in a hierarchical structure can be described as follows:

$$\begin{aligned}
&\text{Level 1:} \quad y_{ik}|\theta_{ik} \overset{ind}{\sim} \text{Bernoulli}(\theta_{ik}),\\
&\text{Level 2:} \quad \text{logit}(\theta_{ik}) = x'_{ik}\beta + v_i,\\
&\text{Level 3:} \quad v_i \overset{iid}{\sim} N(0, \sigma_v^2).
\end{aligned} \tag{1.16}$$

Naturally, the interpretation of the θ_{ik}'s can be carried out in the usual fashion:

$$\begin{aligned} y_{ik} &= 1, \text{ w. p. } \theta_{ik}, \\ &= 0, \text{ w. p. } (1 - \theta_{ik}). \end{aligned} \tag{1.17}$$

There have been many works done on applying variations of such unit level models to real-life problems. One notable variation can be found in the model considered by [24]. They considered a logistic regression model with random regression coefficients to estimate the true area level proportions using the National Health Interview Survey (NHIS) data. This is an example of a multistage, personal interview-based sample survey that is conducted annually by the National Center for Health Statistics (NCHS). Other relevant works include those by [25], [26], and [27], among others.

1.4 Inference using the small area models

So far, we have discussed various models used in applied small area estimation. We have gone through the modeling aspects but glossed over the estimation and inference procedures. Now, we aim to seize this opportunity to delve deeper into the inferential procedures associated with model-based small area estimation.

1.4.1 Parameters of interest

In small area estimation (SAE) model-based problems, the main parameters of interest encompass both survey-specific characteristics and model-related components. Survey parameters

of interest typically include small area parameters, such as means, proportions, or totals, which represent key attributes of the population within each small area. Oftentimes, we are also interested in assessing their estimation errors. This, in turn, provides us with the tools to carry out various hypothesis testing. These parameters are directly related to the survey data and serve as the focal points for the small area estimation. On the other hand, model parameters encompass various elements associated with the statistical models utilized in SAE, including fixed effects, random effects, and model hyperparameters. Fixed effects represent systematic relationships between the small area parameters and covariates (auxiliary data), while random effects capture unobserved variability, spatial dependencies, etc. Hyperparameters govern the distribution of other parameters in hierarchical Bayesian models, playing a crucial role in shaping the overall estimation process. Together, these parameters facilitate accurate inference and enable researchers to derive meaningful insights about small area characteristics from survey data, along with various other auxiliary data.

Notations: Let us now give the mathematical definitions of the parameters and the associated concepts. Remaining consistent with the notations used thus far, we shall refer to Y_{ik} as the value of the survey variable of interest for the k^{th} unit within the i^{th} small area; $i = 1, \cdots, m,\ k = 1, \cdots, N_i$. We define the small area population means as:

$$\bar{Y}_i = \frac{\sum_{k=1}^{N_i} Y_{ik}}{N_i} = \frac{Y_i}{N_i},\ i = 1, \cdots, m, \tag{1.18}$$

where Y_i is the corresponding small area population total for this particular characteristic.

In many practical scenarios, the Y_{ik}'s often take binary values, representing the presence or

absence of a certain variable of interest within each small area i. In such cases, the population total corresponds to the count of the observed characteristic within the small area, while the mean represents the population proportion of the observed characteristic within that area. Following a survey, we observe y_{ik} for $i = 1, \ldots, m$ and $k = 1, \ldots, n_i$, indicating that n_i units out of the total N_i units within each small area i have been surveyed. All estimations and associated inferences must be based on these observed values of the population characteristic of interest.

Various methods exist for estimating the parameters of interest within small areas. In chapter 1.2, we explored methods employed using the sampling design utilized in the sample survey, discussing concepts such as design-unbiasedness, consistency, and the potential for estimation errors. However, when it comes to model-based inference, the empirical best prediction (EBP) approach and the hierarchical Bayes (HB) approach stand out as the two most commonly utilized methods. The EBP approach is based on 'classical', or 'frequentist' methods, like maximum likelihood estimation (MLE), restricted maximum likelihood estimation (REML), and method of moments (MOM) estimation of the model parameters. Whereas the HB approach is based on the 'Bayesian' way of assuming priors on the model parameters and calculating the estimates based on the posterior distributions. [28], [6], and [4] give extensive reviews of the EBP and the HB approaches.

1.4.2 Empirical Best Prediction (EBP)

In the area level models in 1.6, 1.8, 1.9, we were modeling the direct estimates simultaneously with the true population parameter, using hierarchical models. Using the same notations, we call the population parameter of interest θ_i for small area i. y_i is the direct estimate

of θ_i. The error in estimation is usually quantified by the direct design-based estimate of the sampling variances D_i of y_i. In most applications of model-based SAE, we assume that the pair $y_i, D_i,\ i = 1, \ldots, m$ is known. Using the FH model in 1.7 as a reference, and by further assuming that the model parameters–β, A–are known, we can give the best predictor (BP) by minimizing the mean squared prediction error (MSPE). This quantity is defined as:

$$MSPE(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i)^2, \tag{1.19}$$

where the expectation is taken under the model. In our case, the FH mixed model is described in 1.7. The BP of θ_i thus obtained is given by:

$$\hat{\theta}_i^{BP}(y_i;\ \beta, A) = (1 - B_i)y_i + B_i x_i'\beta, \tag{1.20}$$

with the MSPE of $\hat{\theta}_i^{BP}$:

$$MSPE(\hat{\theta}_i^{BP}) = (1 - B_i)D_i = g_{1i}(A),\ \text{say}, \tag{1.21}$$

where $B_i = \frac{D_i}{D_i + A}$ is called the 'shrinkage factor'. Naturally, when $D_i \to 0$, meaning that we observe y_i with very low estimation error, the model based BP is dominated by y_i and relies less on the 'synthetic' part ([4]) $x_i'\beta$. Conversely, if for some area i, possibly with very low (or no) sample size, where the direct estimate y_i is observed with high a D_i, the model based BP will depend more on the synthetic part $x_i'\beta$, and give less weight to the direct estimate y_i. B_i determines this level of mixture between the synthetic and the direct estimate, and hence, it is called the shrinkage factor.

Next, we relax the assumption of the known regression coefficients β's and assume that only A is known. Then the best linear unbiased predictor (BLUP) of θ_i is obtained by minimizing the $MSPE(\hat{\theta}_i)$ among all linear unbiased predictors of θ_i, that is, the predictors of the form: $\{\hat{\theta}_i = \sum_{j=1}^{m} l_{ij} y_j : E(\hat{\theta}_i - \theta_i) = 0\}$. The BLUP is given by:

$$\hat{\theta}_i^{BLUP}(y_i;\ A) = (1 - B_i) y_i + B_i x_i' \hat{\beta}^{WLS}(A), \tag{1.22}$$

where $\hat{\beta}^{WLS}$ is the weighted least square (WLS) estimate of β, with the weights being $\{(1 - B_i),\ i = 1, \ldots, m\}$. The expectation here is also taken under the same linear mixed model 1.7. This BLUP can be obtained by using Henderson's theory [29] on estimation from linear (mixed) models. The usual WLS estimate of β is obtained by minimizing the weighted sum of squares of the residuals and is given by:

$$\hat{\beta}^{WLS} \equiv \hat{\beta}_{WLS}(A) = \left(\sum_{j=1}^{m} (1 - B_j) x_j x_j' \right)^{-1} \sum_{j=1}^{m} (1 - B_j) x_j y_j. \tag{1.23}$$

Note that, 1.20 and 1.22 are identical, except for the β part. We can obtain 1.22, by simply replacing β with its WLS estimate $\hat{\beta}^{WLS}$ in 1.20. The MSPE of this BLUP is given by:

$$MSPE(\hat{\theta}_i^{BLUP}) = (1 - B_i) D_i + B_i^2 \cdot x_i' \left(\sum_{j=1}^{m} (1 - B_j) x_j x_j' \right)^{-1} x_i, \tag{1.24}$$

$$= g_{1i}(A) + g_{2i}(A),\ \text{say}, \tag{1.25}$$

where $g_{2i}(A) = B_i^2 \cdot x_i' \left(\sum_{j=1}^{m} (1 - B_j) x_j x_j' \right)^{-1} x_i$ is the extra variability due to the WLS estimation of the regression parameters β, on top of $g_{1i}(A)$, reflected on the BLUP of θ_i. Now,

under the standard regularity conditions, $g_{1i}(A) = O(1)$, whereas, $g_{2i}(A) = O(m^{-1})$, for large enough m, the number of small areas. In the context of small area higher-order asymptotics, $g_{1i}(A)$ contributes more to the $MSPE(\hat{\theta}_i)$ than $g_{2i}(A)$ [14].

In practice, however, the variance components are due to the random effects in 1.7, namely A, is not known. The empirical best prediction approach for estimating θ_i, is based on replacing A with an estimated value $\hat{A}$ in the BLUP, obtained through the standard variance components' estimators. Such estimates can be obtained by using the method of moments ([1], [30]), maximum likelihood (ML), or the restricted maximum likelihood (REML) estimation. Note that these are all based on the frequentist idea. Furthermore, since such an estimate $\hat{A}$, works as a plug-in estimator (based on the empirical evidence) in the BLUP, such an estimate $\hat{\theta}_i$ is called an empirical best linear unbiased predictor or EBLUP. A critical drawback of $\hat{\theta}_i^{EBLUP}$ produced in such a way is that the MOM, ML, and REML estimates of A can often yield an estimate of 0. In such cases, the $\hat{\theta}_i^{EBLUP}$ reduces to $\hat{\theta}_i^{EBLUP} = x_i'\beta$, the synthetic estimate, for all the areas $i = 1, \ldots, m$. The direct estimate, y_i, gets no weight at all, however big the sample sizes may be for the individual small areas. Adjustment to density method (ADM) proposed by [31] can tackle this problem of 0 estimates with good asymptotic properties. For an extensive review of different ways of estimating the variance component, please see [6]. Comparing the estimation of variance components to assessing their uncertainties due to estimation, the former is generally considered a relatively simpler task. There have been many works done in the SAE literature to address this issue. [30] used a mean square error (MSE) criterion to measure the $MSPE(\hat{\theta}_i^{EBLUP})$ for a longitudinal mixed linear model that covers both the FH model and the nested error regression model (like the Battese-Harter-Fuller model in 1.13). Resampling-based methods, such as jackknife and parametric bootstrap have been proposed for approximating the MSE of the EBLUPs.

Please see these review papers by [6], and [32] for extensive reviews of estimation, and the MSE estimation of EBLUP.

1.4.3 Hierarchical Bayes (HB)

Hierarchical Bayesian methods represent a powerful framework for model-based small area inference, leveraging posterior distributions of model parameters given the observed data to produce reliable estimates for spatially disaggregated data. At the core of hierarchical Bayesian modeling lies the incorporation of hierarchical structures, which capture the multi-level nature of the data-generating process and allow for the estimation of parameters at different levels of aggregation.

In the context of small area inference, hierarchical Bayesian methods typically involve two levels: a population-level model and an area-level model. The population-level model describes the overall distribution of the target variable across all small areas, incorporating global parameters that characterize the population distribution. Mathematically, this can be represented as:

$$y_i|\theta_i \sim f(\theta_i),$$

$$\theta_i|\phi \sim \text{Population Model}(\phi),$$

where y_i represents the observed data in small area i, θ_i denotes the latent parameter associated with area i, and ϕ represents the hyperparameters governing the population-level distribution. The function f represents the likelihood of the data given the parameters, which can vary depending on the specific modeling assumptions and the nature of the data.

The area-level model captures the variability across small areas and accounts for local

characteristics that may influence the target variable. This model typically involves specifying a distribution for the area-specific parameters conditioned on the population-level parameters, thus enabling the borrowing of information across areas. Mathematically, the area-level model can be expressed as:

$$\theta_i|\eta_i \sim \text{Area Model}(\eta_i),$$

where η_i represents the hyperparameters governing the distribution of parameters at the area level. The choice of the area-level model depends on the characteristics of the data and the spatial structure of the small areas. Similar specifications can be done for unit-level models, as described previously in section 1.3.2.

One of the key advantages of hierarchical Bayesian methods is the ability to incorporate prior information and external sources of data to improve inference. Priors can be specified for both population-level and area-level, and/or unit-level parameters, allowing researchers to incorporate domain knowledge and existing information into the modeling process. This facilitates robust estimation, especially in cases where data are limited or noisy.

Furthermore, hierarchical Bayesian methods naturally accommodate uncertainty estimation through the posterior distribution, which provides a comprehensive representation of uncertainty in the estimated parameters. By sampling from the joint posterior distribution $p(\eta|y)$ of model parameters η given the observed data y, researchers can obtain credible intervals and posterior predictive distributions for small area estimates. This enables quantification of uncertainty and robust decision-making based on the level of confidence in the results. In practice, posterior samples are drawn using Markov Chain Monte Carlo (MCMC) methods, such as Hamiltonian

Monte Carlo (HMC), to explore the high-dimensional parameter space and obtain samples from the posterior distribution. These samples are then used to estimate the posterior mean, credible intervals, and other summary statistics of interest.

Hierarchical Bayesian methods provide an exceptionally versatile and powerful framework for model-based small area inference. By utilizing posterior distributions of model parameters, they offer a reliable means of producing estimations and quantifying uncertainty in spatially disaggregated data.

1.5 Markov Chain Monte Carlo using Stan

In this thesis, we carry out all the Bayesian computations using **Stan** ([33]). Stan is a probabilistic programming language that uses Hamiltonian Monte Carlo (hereinafter HMC)—a Markov Chain Monte Carlo method that utilizes the derivative of the posterior density function (being sampled) to generate efficient transitions spanning the posterior distribution. It uses an approximate Hamiltonian dynamics simulation based on numerical integration, which is then corrected by performing a Metropolis-Accept step. The goal of the sampling is to draw from the posterior density, say $p(\tau|y)$, of the model parameters τ given the data y.

For the sake of illustration, let us use one of the models we use in this thesis. Using this model, let us now briefly explain the key steps of HMC for estimating model parameters of the proposed model in 3.2. Stan can work with an unnormalized posterior density as long as it is a proper posterior density. There are ways around drawing from improper posteriors in Stan, but we refrain from furthering that discussion here. For the model 3.2, this unnormalized posterior

density of interest is given by:

$$p(\tau|y) \propto N(\alpha|0,5^2) \cdot \prod_{p=1}^{P} N(\beta_p|0,5^2) \cdot \prod_{c=1}^{C} N(\xi_c|0,5^2) \cdot N(\lambda|0,5^2) \cdot \prod_{i=1}^{m} N(v_i|0,\sigma_v^2) \cdot$$
$$\prod_{i=1}^{m}\prod_{g=1}^{G} N(u_{ig}|0,\sigma_u^2) \cdot \text{Cauchy}^{+}(\sigma_v|0,5) \cdot \text{Cauchy}^{+}(\sigma_u|0,5) \cdot$$
$$\prod_{i=1}^{m}\prod_{g=1}^{G}\prod_{k=1}^{n_{ig}} N(\phi(\theta_{igk})|\alpha,\beta,\xi_c,\lambda,\sigma_v^2,\sigma_u^2) \cdot \prod_{i=1}^{m}\prod_{g=1}^{G}\prod_{k=1}^{n_{ig}} f_1(y_{igk}|\theta_{igk},V(\theta_{igk})) \cdot \pi(V(\theta_{igk})), \tag{1.26}$$

where $\pi(V(\theta_{igk}))$ is a proper prior distribution on the QVF $\pi(V(\theta_{igk}))$. Since we have used proper prior distributions for all model parameters, the resulting posterior density will also be proper (i.e., a finite integral of the unnormalized specification). This posterior can be coded in as the 'target posterior density' distribution using the programming language specific to Stan. As is the case with many practical applications, the exact posterior distribution of 1.26 is also intractable. Stan will use HMC to draw samples from an approximate posterior and then correct those draws to improve approximation to the target posterior density $p(\tau|y)$. HMC introduces an auxiliary 'momentum' (as is understood in classical physics) vector ρ, having the same dimension as our model parameters τ, and draws from the joint distribution of (ρ,τ) given the data—$p(\rho,\tau|y)$. In most applications of HMC, including Stan, this auxiliary density is a Multivariate Normal that does not depend on the model parameters τ; $\rho \sim MN(0,M)$, where M is called a 'mass' matrix. The mass matrix is essentially an Euclidean metric and can be seen as a transformation of the parameter space that makes the sampling task more efficient. Here is a brief outline of how Stan implements the HMC technique:

HMC proceeds in a series of iterations (see [34], [33]) with each iteration having the fol-

lowing 3 steps:

1. **Introduction of a momentum:** The iteration begins by updating ρ with a random draw from $MN(0, M)$, independently of τ, the current values of model parameters. To make the algorithm more efficient, it can be useful for M to roughly scale with the inverse of the posterior covariance matrix $(var(\tau|y))^{-1}$. This is why, by default, Stan sets M^{-1} equal to a diagonal estimate of the covariance computed during the warm-up phase.

2. **Solving the joint system of the current parameter values τ and ρ:** The main part of the HMC iteration is a simultaneous update of (τ, ρ) conducted effectively via a discrete mimicking of the Hamiltonian dynamics for a physical system of a very specific structure—and the introduction of the auxiliary variable ρ makes sure of just that. This step is all about numerically solving a system of two-state differential equations induced by the paired time derivatives of (τ, ρ). These differential equations involve the gradients of the log posterior density for each of the model parameters. Stan uses the well-known *leapfrog integrator* to obtain solutions to these paired time differential equations. This second step has 3 further sub-steps that need to be repeated L times. Details are omitted for brevity, but we refer the readers to [34] for an excellent review of this step.

3. **Metropolis-Accept step:** Label τ^{t-1} and ρ^{t-1} as the values of the parameter and the momentum vectors at the start of the leapfrog step, and τ^* and ρ^* as the values obtained after using the L steps of the previous step. In this final step, make an update according to the

following rule:

$$\tau^t = \begin{cases} \tau^* & \text{, w. p. } min\left(1,\ \dfrac{p(\tau^*|y)\cdot p(\rho^*)}{p(\tau^{t-1}|y)\cdot p(\rho^{t-1})}\right), \\ \tau^{t-1} & \text{, otherwise.} \end{cases}$$

Strictly speaking, it would be necessary to set ρ^t as well, but since we are not interested in ρ, and it gets immediately updated with a fresh draw at the very beginning of the next iteration, there is no need to keep track of it after the Metropolis accept-reject step has been performed.

From equation 3.1, our parameter of interest can be written down as: $\bar{Y}_i \approx \bar{\Theta}_i$. Thus, to estimate the population parameter of interest: $\bar{Y}_i$, we use the posterior draws $\{\bar{\theta}_i^{(r)},\ r = 1, \cdots, R\}$ of θ_i. The posterior mean (Bayes estimate) and posterior standard deviation (standard error) of $\bar{\theta}_i$ are approximated by the average and standard deviation of $\{\bar{\theta}_i^{(r)},\ r = 1, \cdots, R\}$, respectively. Different quantiles of the posterior distribution of θ_i are approximated by the empirical quantiles of $\{\bar{\theta}_i^{(r)},\ r = 1, \cdots, R\}$.

1.5.1 Model comparison in Stan

Let us now briefly discuss the model comparison (or selection) capabilities that Stan has built in, simply as a by-product of the log posterior density calculation that Stan does anyway for sampling from the posterior [35]. The **R**-package `loo` [36] provides a quick and convenient way to diagnose and compare model fits, and we have used the same in this work. Although the capabilities are built-in, implementations, however, are not automatic. This is simply because there is a need to compute separate factors: $p(y_i|\tau)$ in the likelihood. Stan works with the joint

density and in its usual computations, does not "know" how to differentiate between the prior and the likelihood factors that go into the calculation. Thus, in order to compute these measures of predictive fit in Stan, we, the users, need to explicitly code the terms of the log-likelihood in a vectorized fashion. Then we can employ the package `loo`, to pull apart the separate terms and compute cross-validation measures at the very end, after all the simulations have been completed. The software computes approximate Leave-One-Out Cross-Validation (LOO-CV) using Pareto smoothed importance sampling (PSIS), a relatively new procedure for regularizing importance weights. We refer to [35] for a detailed discussion on 'loo' with Stan. As a byproduct of the calculations, `loo` can also obtain approximate standard errors for estimated predictive errors and for comparing predictive errors between two models. We have used PSIS-LOO-CV instead of WAIC (Widely Applicable Information Criterion – an alternative approach to estimating the expected log pointwise predictive density) because PSIS provides useful diagnostics and effective sample size and Monte Carlo standard error estimates (see [35]). When comparing any two fitted models, the package can estimate the difference in their expected predictive accuracy by the difference in their $\hat{elpd}_{loo}$, the Bayes Leave-One-Out estimate of the expected log pointwise predictive density (elpd). This difference will be referred to as `elpd_diff` in the subsequent sections, whenever needed. For a large enough sample size (used to fit the data) this difference is an approximate standard normal distribution. This will facilitate all the model comparisons that we have used in this research.

1.5.2 MCMC diagnostics

In the context of Markov Chain Monte Carlo (MCMC) diagnostics, effective sample size (ESS) and the Gelman-Rubin statistic (R-hat) are crucial metrics used to assess the quality of the samples generated by the MCMC algorithm and to determine if the chains have converged [34].

1. Regular Effective Sample Size (Regular_ESS): The regular ESS, often simply referred to as ESS, is a general measure of the effective sample size that considers the entire MCMC chain, including both the initial transient phase (burn-in period) and subsequent samples. It provides an overall assessment of the efficiency of the MCMC algorithm in exploring the target distribution. Mathematically, the ESS can be calculated using the formula:

$$\text{ESS} = \frac{n}{1 + 2\sum_{t=1}^{T} \rho_t}$$

where:

- n is the total number of samples.
- T is the lag or thinning parameter.
- ρ_t is the autocorrelation at lag t.

A common estimate of ρ_t is:

$$\hat{\rho}_t = \frac{\sum_{i=1}^{n-t}(X_i - \bar{X})(X_{i+t} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where X_i represents the i-th sample, and $\bar{X}$ represents the sample mean.

2. Bulk Effective Sample Size (Bulk_ESS): The bulk ESS refers to the effective sample size

calculated from the central portion of the MCMC chain, excluding the burn-in period. It measures how efficiently the chain explores the typical set of the target distribution. Mathematically, the bulk ESS can be defined as:

$$\text{Bulk\_ESS} = \frac{n}{1 + 2\sum_{t=1}^{T} \rho_t}$$

where the parameters are the same as in the regular ESS formula. The bulk ESS focuses on the central portion of the chain and excludes the initial transient phase.

3. Tail Effective Sample Size (Tail_ESS): The tail ESS measures the effective sample size considering only the tail portion of the MCMC chain, which represents the exploration of the tails of the target distribution. It is particularly useful when the tails of the distribution are of interest, such as in extreme value analysis or rare event simulation. Mathematically, the tail ESS can be defined similarly to the regular ESS, but considering only the tail portion of the chain:

$$\text{Tail\_ESS} = \frac{n_T}{1 + 2\sum_{t=1}^{T} \rho_t}$$

where n_T is the number of samples in the tail portion of the chain.

4. Gelman-Rubin Statistic ($\hat{R}$): The Gelman-Rubin statistic, commonly denoted as $\hat{R}$, is a diagnostic tool used to assess the convergence of multiple MCMC chains. It compares the variance between chains to the variance within chains, with values close to 1 indicating convergence. A common formulation for calculating $\hat{R}$ is by comparing the variance between chains (B) to the variance within chains (W):

$$\hat{R} = \sqrt{\frac{\text{Var}(W)}{B}}$$

where:

- Var(W) is the average of the within-chain variances.

- B is a measure of the variance between chains.

Each of these measures provides valuable insights into the efficiency and convergence properties of the MCMC algorithm in different contexts. Regular ESS, Bulk ESS, and Tail ESS offer assessments of sample quality, while $\hat{R}$ helps diagnose convergence issues when multiple chains are used.

According to [34], optimal choices of n_{eff} and $\hat{R}$ values play a crucial role in diagnosing the quality of posterior samples in applied Bayesian statistical analysis. They recommend running the simulation until the effective sample size (n_{eff}) is at least $5m$, where m is twice the number of independent chains or sequences. This criterion helps ensure that a sufficient number of effectively independent samples is obtained accurately spanning the posterior distribution space. Additionally, they propose using the Gelman-Rubin statistic ($\hat{R}$) to assess convergence and chain equilibrium. An $\hat{R}$ value close to 1, but less than 1.1, is indicative of chain equilibrium, suggesting that the multiple chains have converged to the target distribution.

These guidelines provide practical heuristics for MCMC practitioners to evaluate and ensure the quality of their posterior samples, ultimately enhancing the reliability of Bayesian inference.

1.6 Multilevel Regression and Poststratification (MRP)

In Bayesian statistics, Multilevel Regression and Poststratification (MRP) ([37]) is a widely used method for adjusting survey estimates to account for nonresponse bias and data sparsity. MRP operates in two stages where it first employs a multilevel regression model to predict the

outcome measure based on a set of covariates, then post-stratifies the model predictions to obtain population-level estimates.

1. **Model Specification**: In the first stage, MRP fits a multilevel regression model to the observed data. Let $\mathbf{y}$ denote the outcome variable of interest, and $\mathbf{X}$ represent the matrix of covariates. The multilevel model specifies a linear predictor for the mean response $\boldsymbol{\theta}$ within each poststratification cell j. Mathematically, this can be expressed as:

$$\boldsymbol{\theta}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{u}_j$$

where $\mathbf{X}_j$ represents the design matrix for poststratification cell j, $\boldsymbol{\beta}_j$ is the vector of fixed effects coefficients, and $\mathbf{u}_j$ denotes the random effects capturing the variability between cells.

2. **Poststratification**: In the second stage, MRP aggregates the cell-level estimates $\boldsymbol{\theta}_j$ to obtain population-level estimates. Let N_j denote the known population size of cell j. The MRP post-stratified estimate for a small area i is given by:

$$\theta_i^{\text{MRP}} = \frac{\sum_{j \in i} N_j \theta_j}{\sum_{j \in i} N_j}$$

where θ_j represents the model-based estimate for cell j. This weighted aggregation allows MRP to produce reliable estimates even for small areas with sparse data.

The core assumptions of MRP entail uniform inclusion probabilities for individuals within cells and the similarity between those included and excluded (in the sample). Moreover, MRP heavily

depends on reliable population counts sourced from extensive surveys such as the Census and the ACS (American Community Survey). These foundational assumptions are critical for the integrity of the poststratification process and the precision of the ensuing estimates. While MRP has demonstrated efficacy across various Small Area Estimation contexts, as evidenced by studies like [38], [39], [40], and [41], among others, meticulous scrutiny of the underlying assumptions and model specifications are imperative to uphold the credibility of the outcomes.

1.7 Discussion and Organization of the Dissertation

In this final subsection, we provide a brief overview of the structure and content of the dissertation.

Chapter 2 delves into pertinent issues within the poverty mapping literature. With a focus on multidimensional poverty measures, we explore the synthesis of scores from various dimensions to gauge poverty levels comprehensively. Our investigation aims to elucidate the contributions of individual dimensions to overall poverty scenarios at the small area level. Additionally, we address the challenges encountered in estimating the proportions of impoverished individuals based on household-level survey data, proposing adaptable solutions tailored to similar contexts. This chapter endeavors to bridge research gaps in multidimensional poverty mapping.

Chapter 3 presents a unified methodology for estimating small area means. Unlike conventional approaches that rely on explicit models for the finite population, our method intricately models the sampled units. This nuanced approach offers several advantages, including extensive validation of the explicit sample model and incorporation of both sample weights and model-based estimates for enhanced robustness and reduced sensitivity to model misspecifica-

tions. Moreover, our methodology effectively handles informative sampling scenarios, ensuring reliable small area estimates by leveraging information from disparate data sources.

In Chapter 4, we specialize the model introduced in Chapter 3 to accommodate survey responses with more than two categories. Furthermore, we scrutinize the issue of biased or informative sampling, demonstrating the pitfalls of employing identical models for both population and sample in such scenarios. We compare our proposed methodology with Multilevel Regression and Post-stratification (MRP), a popular approach in Small Area Estimation literature, highlighting the limitations of MRP when faced with informative sampling. Despite MRP yielding estimates with lower standard errors, our method offers more reliable inference, especially in contexts where national survey-based estimates serve as reliable benchmarks.

Finally, Chapter 5 concludes the dissertation with discussions on the findings and potential avenues for future research.

Chapter 2: Multidimensional Poverty Mapping for Small Areas

2.1 Introduction

The United Nations (UN) established the Sustainable Development Goals (SDGs) in 2015, which comprise 17 goals and 169 targets. These goals are part of the 2030 Agenda for Sustainable Development and represent a global effort to encourage all countries and stakeholders to collaborate in order to eliminate poverty, preserve the environment, and promote peace and prosperity for all people by the year 2030. These 17 SDGs have more than 210 indicators. These indicators are used to measure progress toward achieving the goals and cover a wide range of topics, including poverty, health, education, gender equality, and sustainable development. Each goal has a specific set of indicators, and some indicators are shared across multiple goals. Reliable disaggregated data is critical for achieving Sustainable Development Goals because the goals aim to be inclusive and comprehensive. Disaggregating data by factors such as age, gender, income, and location helps policymakers identify specific needs and challenges faced by different groups and design tailored policies and programs to address them. This approach ensures that all segments of society are accounted for and no one is left behind in the pursuit of the SDGs.

The very first of these 17 goals is 'No Poverty: End poverty in all its forms everywhere.' The primary indicators for this goal include measuring the proportion of the population living below the international poverty line, the proportion of the population covered by social pro-

tection systems, and the proportion of the population living in households with access to basic services. These indicators help measure progress toward the goal and inform policy development. There are multiple approaches to conceptualizing poverty: unidimensional poverty and multidimensional poverty. Unidimensional poverty measures are based on a single dimension, usually income or consumption. Such a monetary approach is the earliest attempt at defining poverty by assessing individuals or household welfare in terms of income or consumption. The multidimensional approach considers the set of different disadvantages experienced by poor people simultaneously, such as deprivation in health, education, nutrition, and lack of clean water etc.

Sen's Capability approach ([42]) provides a broad normative framework for assessing individual well-being and the nature of poverty. This laid the philosophical foundations for the multidimensional poverty measures that followed. The *Fuzzy Set* method (see [43]) and *Counting* method (see [44]) are predominantly used to measure multidimensional poverty. Recently, [45] proposed a *Synthesis* method, which is a well-organized amalgamation of these two techniques. Multidimensional poverty measures have become feasible due to the availability of household survey data. Multidimensional measures of poverty allow for the quantification of poverty from different dimensions by incorporating multiple indicators of poverty. This enables a more comprehensive understanding of poverty, including its multiple dimensions and how they interact, which, in turn, can inform more effective poverty reduction strategies. These measures can also identify the most disfavored individuals or groups, enabling targeted interventions to address their specific needs. Such multidimensional measures of poverty offer more than just an overall measure that combine various dimensions. They can also help quantify the impact of individual dimensions, which can help identify areas of concern and allocate resources accordingly to address issues associated with each dimension.

In this research, we consider the multidimensional synthesis poverty measure introduced by [45]. They selected dimensions based on popular literature and kept Goal 1 of the SDGs into consideration. The three main dimensions that they considered are Material Deprivation (MD), Deprivation of Social Dimension (SD), and Deprivation of Human Capital (HC). Material deprivation includes housing facilities, consumer durables, and clothing. Deprivation of social dimension considers social network, dignity, and autonomy. Deprivation of human capital covers education, employment, health, and nutrition.

We develop models and methods to estimate relative contributions from different dimensions of poverty as well as a composite poverty measure incorporating all dimensions for a granular level. The use of small area modeling in the context of poverty assessment is commonly known as *Poverty Mapping*. In these smaller regions, there is usually limited or no information available from surveys, making it essentially a small area problem. Numerous papers, including those by [18], [1], [2], and [46], discussed the challenges of small area estimation and potential solutions. However, despite the existing literature on poverty mapping, there is currently no study that specifically addresses the estimation of relative contributions from different dimensions of poverty, utilizing careful small area modeling tools and techniques. This research aims to bridge that gap and contribute to the poverty mapping literature by addressing these important aspects.

In order to tackle these challenges, we adopt an approach that involves incorporating area-level auxiliary information from reliable external sources. The availability of trustworthy area-level auxiliary data is crucial for accurate estimation at granular levels (small areas). We employ multivariate joint modeling to effectively estimate the relative contributions of three dimensions at small area levels. This allows us to understand the relative importance of each dimension in measuring poverty. Moreover, we develop a model for the multidimensional composite measure,

which enables us to generate poverty estimates for small areas. By following this comprehensive approach, we aim to derive more reliable and insightful poverty assessments at granular levels.

The rest of this chapter is structured as follows: We start by briefly describing the data set that we have used in this research in section 2.2. Section 2.3 discusses the necessary small area model and method for measuring relative contributions from different dimensions of poverty. Section 2.4 explores models to estimate multidimensional composite measures and associated uncertainty measures for small areas. In section 2.5, we present results obtained from applying the proposed models and methods to the Sri Lankan data. Moreover, we present key findings from these applications. Finally, we conclude this chapter by summarizing our findings and discussing potential avenues for future research and extensions.

2.2 Data Description

The primary data for the analysis conducted in this research, aiming to identify impoverished individuals residing in the Uva province of Sri Lanka, is based on the "Synthesis method" proposed by [45]. We prepared the set of aggregated-level covariates for our modeling from the Sri Lankan 2012 census data.

2.2.1 Primary survey data

The authors [45] designed both the sample and questionnaire for a household survey conducted in the Uva province of Sri Lanka. This survey employed a well-structured, pre-tested questionnaire consisting of close-ended questions tailored to the study's objectives. It was carried out over the period of November 2016 to January 2017.

The survey utilized a stratified two-stage sample design across two districts within the Uva province: Badulla and Moneragala. These districts were further divided into five strata, representing different regions. Badulla has the urban, rural, and estate regions, and Moneragala has the rural and estate regions. Within these regions, Primary Sampling Units (PSUs) were selected based on a probability proportional to size (PPS) scheme. The total sample size of 1200 housing units was allocated across the strata, with 120 PSUs chosen to represent various socio-economic groups within the province. Of these, 78 PSUs were from the Badulla district, while 42 were from the Moneragala district.

In the second stage of sampling, Secondary Sampling Units (SSUs) were systematically selected from each PSU. Ten housing units (SSUs) were systematically selected from each selected census block (PSU) for inclusion in the survey. All households within these selected housing units were surveyed, with information collected through face-to-face interviews. The survey collected data on the general characteristics of both individuals within households. The study's reference population comprised all individuals aged eighteen and above residing in the Uva Province of Sri Lanka, with the individual who responded during the interview serving as the unit of analysis. The selection of respondents within households was not based on any specific method. Respondents were typically individuals over the age of 18 residing in the household. The survey aimed to enumerate all households within the selected housing units and face-to-face interviews were conducted to collect information on the general characteristics of individuals within households.

The survey targeted a total of 1193 respondents, out of which 730 were female and 463 were male. The sample aimed to be representative of all socio-economic groups within the Uva province. However, for the analysis conducted in this research, only records with complete information on the selected variables were considered. As a result, the analysis utilized information

from 731 out of the 1193 respondents.

For the analysis, a binary response variable indicating poverty status (1 for poor, 0 for non-poor) was derived from the composite multidimensional measure. They considered multiple indicators of poverty, combining them into three different dimensions. Based on the scores and cut-offs for these 3 mutually exclusive and exhaustive dimensions, they created their final composite measure. Using a separate cut-off again, they finally created the binary variable, accounting for the person level poverty status. This variable was employed to estimate poverty at lower administrative divisions using Small Area Estimation (SAE) modeling techniques. In the sections that follow, we will talk about certain challenges we faced, even after careful design and collection of the data by the said authors.

2.2.2 Auxiliary data

Small area literature talks about the importance of the availability of good administrative records suitable for the response variable. Having these different kinds of data, possibly from different levels, such as area level aggregates or summaries, or unit-level information, increases the flexibility of the hierarchical models [4]. The auxiliary variables (x) used here in this study were derived from the 2012 Census of Population and Housing conducted by the Department of Census and Statistics of Sri Lanka. These variables included “floor”, “ownsMobile”, “waterSafe”, “ownsTV”, “wall”, and “toilet”, representing percentages of households with specific characteristics aggregated to Divisional Secretary’s Divisions (DSDs) within the Uva province. “Floor” represented the percentage of households with permanent flooring materials, “ownsMobile” indicated the percentage with mobile phone ownership, “waterSafe” represented households with

access to safe drinking water, "ownsTV" indicated households owning a television, "wall" represented households with permanent wall materials, and "toilet" indicated households with access to improved sanitation facilities. For more details on how the different indicators were chosen, how they were combined into the different dimensions, and the details about the selection of cut-offs, readers are encouraged to read their work [45].

2.3 Relative Contributions from Different Dimensions of Poverty: Notations, Definitions, and Modeling

In this section, we delve into the intricate dimensions of poverty and explore their relative contributions to the small areas' poverty landscape. By establishing clear notations, defining key concepts, and formulating appropriate models, we aim to gain deeper insights into the multifaceted nature of poverty and its underlying determinants. We seek to unravel the complex interplay between various poverty dimensions and their impacts on individuals and communities through rigorous analysis and modeling techniques.

In our study, we adopt a multidimensional approach to measure poverty, as outlined in [45]. This approach, known as the 'synthesis method,' integrates diverse poverty indicators into distinct 'dimensions.' By amalgamating these dimensions, the method generates a comprehensive assessment of poverty, represented by a singular score. Through the application of a predefined threshold, the authors dichotomized the composite measure, assigning a value of 1 to individuals classified as poor and 0 to those deemed non-poor. The dimensions identified in their study include 'material deprivation,' 'social deprivation,' and 'human capital.' For detailed insights into these dimensions, their corresponding scores, the designated thresholds, and the methodology

employed for their determination, we direct readers to the original work by [45].

During our exploration of modeling techniques for the composite measure, aimed at making inferences for smaller regions within the Uva province, a pertinent query emerged: Is it feasible to jointly model the various dimensions represented by their assigned scores? Such an approach holds the potential to address a crucial scientific inquiry regarding the relative contributions of each dimension to the individual-level poverty status observed across the small areas within the Uva province.

2.3.1 Notations

We will use of the following notations. For $i = 1, \cdots, m;\ k = 1, \cdots, K$, define

- $\underset{\sim}{\theta}_i = (\theta_{i1}, \cdots, \theta_{iK})' =$ vector of 'true' small area scores, of different dimensions of poverty, for area i. Also, $\theta_{i1}, \ldots, \theta_{iK}$ are all greater than 0 and less than 1. This will be our small area parameter of interest, which we aim to estimate through the use of modeling.

- y_{ik}: survey-weighted estimate of θ_{ik}; denote $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iK})'$. We use the survey design to calculate these estimates from the person-level data.

- $\Sigma_i = ((\sigma_{i;kk'}))$: a known $K \times K$ sampling covariance matrix of $\underset{\sim}{y}_i$. $\sigma_{i;kk'}$ is the sampling covariance between y_{ik} and $y_{ik'}$. In practice, Σ_i is often estimated using a smoothing technique.

2.3.2 Model

With the notations set forth, we are equipped to devise a model aligned with our research objective: discerning the relative contributions of various dimensions of poverty on the mul-

tidimensional poverty for small areas. Given that the dimensional scores are uniformly 0 for non-poor individuals, our model will focus solely on the impoverished population. In what follows, we present a hierarchical model that establishes a link between the true dimensional scores of small areas and the available area-level covariates. While multivariate hierarchical models have been employed in small area research previously, as seen in works like [47] and [48], our approach differs significantly. Unlike these studies, our model incorporates unmatched level 2 linking models, and our parameter of interest varies. Moreover, to the best of our knowledge, this specific type of multivariate unmatched model has not been utilized in the context of estimating relative contributions of various dimensions in multidimensional poverty mapping. Hence, we introduce the following Multivariate Normal Hierarchical Logistic model.

For small areas $i = 1, \cdots, m$,

$$\begin{aligned} &\text{Level 1 (sampling model):} && \underset{\sim}{y}_i | \underset{\sim}{\theta}_i \overset{ind}{\sim} MN(\underset{\sim}{\theta}_i, \Sigma_i), \\ &\text{Level 2 (linking model):} && \text{logit}(\underset{\sim}{\theta}_i) | \beta, A \overset{ind}{\sim} MN(x_i'\beta, A), \end{aligned} \tag{2.1}$$

where MN denotes a Multivariate Normal distribution, β is the vector of unknown regression coefficients, and A is the unknown covariance matrix. We estimate β and A from the model, given data. Note that the linking model facilitates the choice of different sets of auxiliary variables (possibly from reliable external sources, different from the primary sample survey) for different dimensions. This essentially models the direct survey-weighted estimates of $\underset{\sim}{\theta}_i = (\theta_{i1}, \cdots, \theta_{iK})'$, $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iK})'$, with the reliable area level auxiliary data.

2.3.3 Contribution of different dimensions

For the i^{th} area, we define the contribution of the k^{th} dimension to the overall poverty as η_{ik}:

$$\eta_{ik} = \frac{\theta_{ik}}{\theta_{i1} + \cdots + \theta_{iK}}.$$

This is our parameter of interest. Let us now turn our attention to the estimation of Σ_i. Define y_{ijk}: the score for the k^{th} dimension of poverty for the j^{th} person in the i^{th} small area. We propose the following smoothed estimate of $\sigma_{i;kk}$:

$$\hat{\sigma}_{i;kk} = \widehat{var}(y_{ik}) = \frac{s_{kk}}{n_{i,k;\text{eff}}},$$

where $n_{i,k;\text{eff}} = \frac{\tilde{n}_i}{\text{DEFF}_k}$ is the effective sample size of the k^{th} dimension in the i^{th} area; s_{kk} is the pooled variance of y_{ijk} and is defined as:

$$s_{kk} = \frac{\sum_{i,j}(y_{ijk} - \bar{y}_{..k})^2}{P-1},\ \bar{y}_{..k} = \frac{\sum_{ij} y_{ijk}}{\sum_{i,j} 1},\ P = \text{\# poor individuals in the sample.}$$

Note that the scores for the non-poor individuals are 0. We will discuss general concepts of design effect (DEFF), effective sample size (ESS), and the way we compute these specifically for our work in Section 2.4. The calculation of the $\tilde{n}_i$ will follow the same method proposed in Section 2.4. DEFF_k is the design effect corresponding to the k^{th} dimension of poverty. To smooth the covariance terms, $\sigma_{i;kk'}$, consider the following identity:

$$var(y_{ik} + y_{ik'}) = var(y_{ik}) + var(y_{ik'}) + 2cov(y_{ik}, y_{ik'}).$$

Using the above, we can get smoothed estimates of the $\sigma_{i;kk'}$ as:

$$\hat{\sigma}_{i;kk'} = \widehat{cov}(y_{ik}, y_{ik'}) = \frac{\widehat{var}(y_{ik} + y_{ik'}) - \widehat{var}(y_{ik}) - \widehat{var}(y_{ik'})}{2}.$$

To smooth $var(y_{ik} + y_{ik'})$ we follow the method used to smooth $var(y_{ik})$. But, we consider the new variable: $(y_{ijk} + y_{ijk'})$ instead of just y_{ijk}.

2.4 Multidimensional Poverty Rate: Modeling, and Estimation

In the preceding section, we delved into the intricacies of modeling scores derived from various dimensions of poverty, unveiling insights into the relative contributions of these dimensions to an area's overall poverty. Building upon this foundation, we now shift our focus to the modeling of a composite score, which synthesizes the diverse dimensions of poverty into a unified measure. Such modeling endeavors enable the estimation of small area means for person-level poverty, accompanied by robust assessments of estimation accuracy through standard errors. As with our previous exploration, we commence by defining certain quantities of interest that help with the formulation of the model and conclude with a discussion on the estimation of population parameters within this framework. Through this approach, we aim to illuminate the methodology underlying the estimation of Multidimensional Poverty Rates (MPRs), fostering a deeper understanding of poverty dynamics at the micro level.

2.4.1 Model

Let s_i be the household-level sample available for the i^{th} small area. For small areas $i = 1, \ldots, m$, define: $\underset{\sim}{z} = (z_1, \ldots, z_m)'$, where z_i is the estimated proportion of impoverished people

in area i. z_i is the direct design based estimator of π_i, which is the true proportion of impoverished people in the i^{th} area.

$$z_i = \frac{\sum\limits_{h,j \in s_i} w_{ihj} z_{ihj}}{\sum\limits_{h,j \in s_i} w_{ihj}}, i = 1, \ldots, m,\ h = 1, \ldots, n_i,\ \text{and}\ j = 1, \ldots, m_h. \tag{2.2}$$

where z_{ihj} is the poverty status of the j^{th} person, within the h^{th} household (HH), within the i^{th} small area. $z_{ihj} = 1$, if the said person is poor, and 0 otherwise. w_{ihj} is the corresponding person-level survey weight. Naturally, $n_i = |s_i|$ is the number of HHs in the sample for the small area i, and m_h is the HH size of the h^{th} HH.

Now, we can formulate a model that characterizes the true means across the m small areas, linking them to area-level auxiliary information x_i, through hierarchical structures. Traditional area-level models, such as the Fay-Herriot model ([1]) and its various extensions, including unmatched level 2 variations, serve as suitable candidates for this purpose. Drawing inspiration from the methodology outlined by [2], we adopt a similar approach. Moreover, to effectively model the sampling errors associated with the direct estimates in terms of design effect, we propose the Normal-Logistic Random Sampling Variance (NL$_{rs}$) model. Specifically, for small areas indexed by $i = 1, \ldots, m$, the model takes the following form:

$$\begin{aligned} &\text{Level 1 (sampling model):} && z_i|\pi_i \overset{ind}{\sim} N(\pi_i, D_i = \frac{\pi_i(1-\pi_i)}{\tilde{n}_i}\text{DEFF}), \\ &\text{Level 2 (linking model):} && \text{logit}(\pi_i)|\gamma, \sigma_v^2 \overset{ind}{\sim} N(x_i'\gamma, \sigma_v^2), \end{aligned} \tag{2.3}$$

where γ and σ_v^2 are unknown model parameters to be estimated from the model, given data. In

model 2.3, DEFF represents the true design effect, given by:

$$\text{DEFF} = \frac{\text{Var}_{\text{design}}}{\text{Var}_{\text{SRS}}},$$

where $\text{Var}_{\text{design}}$ is the true sampling variance of the survey-weighted estimate z_i, of ϕ_i, under the given complex sampling design. While Var_{SRS} is the true sampling variance of an unweighted estimate under Simple Random Sampling (SRS) of size n_i with replacement. $\tilde{n}_i$ is the 'adjusted' sample size of individuals for the area i. We will shortly explain this in detail. This adjustment was done on top of calculating the DEFF. The other models, for comparison purposes, that we have considered here are the well-celebrated Fay-Herriot model, the logit extension of the Fay-Herriot model–known as the Normal-Logistic model, and finally, the Normal-Logistic plugin model. The details of these well-known models are deferred to external references for brevity, such as the paper by [2] and this book by [4]. For a brief review, however, please refer to subsection 1.3.1 of Chapter 1.

Let us now explore the quantity used in the first level of model 2.3: $\tilde{n}_i$. Traditionally, sampling variances of the direct estimators, D_i, if modeled, are often expressed as:

$$D_i = \frac{\pi_i(1-\pi_i)}{n_{i;\text{eff}}} = \frac{\pi_i(1-\pi_i)}{n_i}\text{DEFF} \tag{2.4}$$

where DEFF represents the design effect and n_i is the sample size for the small area i under simple random sampling. This form assumes simple random sampling of unit-level samples and then corrects it for the complex sampling design by multiplying the variance under SRS by the DEFF. Alternatively, we can define DEFF as:

$$\text{DEFF} = \frac{n}{n_{\text{eff}}} \tag{2.5}$$

In the context of a complex sampling design with n samples, $n_{\text{eff}} = \frac{n}{\text{DEFF}}$ is often referred to as the effective sample size (ESS). This concept proves useful in evaluating the amount of independent information obtained from a complex survey sample. We have used a different form of sampling variance:

$$D_i = \frac{\pi_i(1-\pi_i)}{\tilde{n}_i}\text{DEFF} \tag{2.6}$$

as opposed to the usual form given in equation 2.4. This approach was inspired by the data collection methodology employed by [45] in their study. In their survey, household poverty status aligns with the person-level poverty status for all individuals within the same household. Hence, the sample size for an area i essentially represents the number of households in the survey data available for that area i. Since our parameter of interest, π_i, denotes the proportion of impoverished people for the i^{th} small area, we are specifically interested in person-level poverty. Therefore, in level 1 of equation 2.3, we adjusted this household-level n_i, for the person-level sample size, denoted as $\tilde{n}_i$.

It is worth noting that despite these considerations, the survey weights associated with individuals remain valid due to the complete enumeration conducted within the SSUs. Therefore, the inclusion probabilities, calculated up to the SSUs, ensure these individuals retain consistent weights. Additionally, the final weights were calibrated to the small area population sizes obtained from the census data, as determined by the authors.

2.4.2 Calculation of $\tilde{n}$

We now discuss the calculation of the adjusted sample size, $\tilde{n}$, for any small area. Recall that, z_i given in equation 2.2, is the direct survey-weighted estimate based on s_i, of π_i. We want to calculate the variance of this estimator z_i under the complex sampling. We drop the small area index i in the following steps for ease of explanation.

Initially, we compute the variance under the simple random sampling (SRS) framework. Subsequently, we adjust this variance by multiplying it with the design effect (DEFF) to account for the complexity of the sampling design. For instance, suppose n households were sampled using the SRS scheme. Dropping the small area index i, let z_h represent the poverty status at the household level, where $z_h = 1$ indicates that the HH is poor and 0 otherwise. This definition aligns with that of z_{ihj} in equation 2.2. Recall that all individuals residing within a HH h have the same poverty status, which in turn is labeled as the HH-level poverty status. Thus, we have that $z_{hj} = z_h, \ \forall j$ within that household h (dropping the small area index i uniformly). Now, under SRS, with equal weights, we can express equation 2.2 (again dropping the small area index i uniformly) as:

$$z = \frac{\sum\limits_{h,j \in s} z_{hj}}{\sum\limits_{h,j} 1} = \frac{\sum\limits_{h=1}^{m_h} m_h z_h}{\sum\limits_{h=1}^{m_h} m_h}, \text{ since } z_{hj} = z_h, \ \forall j. \tag{2.7}$$

Under SRS, we can assume that HHs were sampled from a superpopulation of households with $z_h \overset{iid}{\sim} B(\pi)$, where B is the Bernoulli distribution. Then, under SRS, we can obtain the variance of z as:

$$\text{Var}(z) = v(z) \overset{iid}{=} \frac{\sum\limits_{h=1}^{m_h} m_h^2 v(z_h)}{(\sum\limits_{h=1}^{m_h} m_h)^2} = \frac{\sum\limits_{h=1}^{m_h} m_h^2 \pi(1-\pi)}{(\sum\limits_{h=1}^{m_h} m_h)^2} = \frac{\pi(1-\pi)}{\tilde{n}}, \text{ say.} \tag{2.8}$$

From the above equation 2.8, we can obtain the adjusted $\tilde{n}$ as:

$$\tilde{n} = \frac{(\sum_{h=1}^{n} m_h)^2}{\sum_{h=1}^{n} m_h^2}.$$

If $m_h = m, \ \forall \ h$, then $v(z) = \frac{\pi(1-\pi)}{n}$. We will have $\frac{\tilde{n}}{n} \approx 1$, if the HH sizes do not vary much. Using Cauchy-Schwarz inequality, we can show that $\tilde{n} \leq n$. We can interpret $\tilde{n}$ as:

$$\tilde{n} = \frac{n}{\Delta},$$

where Δ is the adjustment factor for estimating the $v(z)$ under person-level poverty, instead of HH-level poverty. Δ can be interpreted as the ratio of the variance of the estimator under HH-level poverty to the variance of the estimator under person-level poverty. The similarity in defining n_{eff} in equation 2.5, with this adjustment factor is noteworthy. We can follow the same steps for calculating $\tilde{n}_i$ for a particular small area i.

2.4.3 Hierarchical Bayesian Estimation

In line with typical Hierarchical Bayesian (HB) methodology, we employ Markov Chain Monte Carlo (MCMC) methods to estimate the model parameters. Let θ_i represent the population parameter of interest for area i. Subsequently, we utilize the set of R MCMC replicates of θ_i, denoted as $\{\theta_i^{(r)}, r = 1, \ldots, R\}$, for making inferences regarding θ_i, the true population parameter. For the Multivariate Normal Hierarchical Logistic model (see Section 2.3), we parameterize

matrix A as follows:

$$A = \sigma^2 \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ & 1 & \dots & \rho_{2k} \\ & & & 1 \end{pmatrix}$$

Here, σ^2 denotes the common variance component, and $\rho_{i;kk'} = \text{corr}(\text{logit}(\theta_{ik}), \text{logit}(\theta_{ik'}))$ for all small areas i and dimensions (k, k'). We adopt weakly informative proper priors for the components of matrix A, specifically: $\sigma \sim \text{Cauchy}^{+}(0, 5)$, a Half-Cauchy prior [49], and the correlation coefficients $\rho_{ik} \overset{iid}{\sim} \text{Uniform}(0, 1)$. Although correlation coefficients could theoretically be negative, [45] suggest they are all positive. Thus, we use a positive uniform prior ranging from 0 to 1 for these correlation entries in matrix A. Finally, for each MCMC replicate $r = 1, \cdots, R$, we compute:

$$\eta_{ik}^{(r)} = \frac{\theta_{ik}^{(r)}}{\theta_{i1}^{(r)} + \cdots + \theta_{iK}^{(r)}}$$

This set of R replicates facilitates posterior inference about the contributions.

2.5 Data Analysis

We carry out all the analyses using the probabilistic programming language **Stan**, along with one of Stan's helper packages for **R** (`loo`).

2.5.1 Joint modeling of the dimensional scores for estimating relative contributions of three dimensions towards the poverty status

In this subsection, we report our findings from joint modeling of the three dimensions that were used to create the final composite measure of poverty status. Our goal here is to attribute

different dimensions towards the final poverty status of the small areas.

As already discussed in Section 2.3, the joint modeling of the 3 dimensions requires data only for impoverished people as determined by the composite score and the 'Synthesis Method'. We note that for the small area 'Meegahakivula', there is only one sample with a 'not poor' poverty status. For this particular area, we use our model and the posterior estimates of the model parameters to create the contribution scores for this particular small area. This, again, demonstrates one merit of the small area modeling techniques. With the help of auxiliary variables from the Census, and along with the MCMC posterior samples of the model parameters, we could provide the estimate (with the desired uncertainty statistics) even for this small area with no sample from the primary survey data. We restricted all prior choices to the proper Weakly Informative Priors (WIP) [49]. The hyperparameters at Level 2 of the Multivariate Normal Hierarchical Logistic model 2.1 are given the following priors:

$$\beta_p \overset{ind}{\sim} N(0, 5^2),\ p = 1, \cdots 7,\ \sigma \sim \text{Cauchy}^{+}(0, 5), \text{and } \rho_{ik} \overset{iid}{\sim} \text{Uniform}(0, 1).$$

We run each model in 4 parallel chains, with $10,000$ iterations in each chain. We discard the first $5,000$ as burn-in samples from each chain. Thus, all the posterior analyses are based on $20,000$ post-warm-up MCMC draws. Table 2.1 displays estimates of the contributions of all three 3 dimensions for all 26 small areas. From this table, we observe that for all the 26 areas, 'Social Deprivation' (sd) dimension contributes the most to an area's person-level poverty status while each of the other two dimensions—'Material Deprivation' (md), and 'Human Capital' (hc)—has about equal contribution to the same.

Table 2.1: Bayes estimates of relative contributions of three dimensions—'Material Deprivation' (md), 'Social Deprivation' (sd) and 'Human Capital' (hc)—for all the 26 small areas from the Uva province of Sri Lanka. The table also gives the standard errors for the estimates (denoted by md_se, sd_se and hc_se) along with their corresponding 95% Bayesian Credible Intervals (_2.5%, _97.5%) for each dimension for 26 small areas.

	small areas	md	md_se	md_2.5%	md_97.5%	sd	sd_se	sd_2.5%	sd_97.5%	hc	hc_se	hc_2.5%	hc_97.5%
1	Badalkumbura	0.242	0.010	0.224	0.263	0.470	0.011	0.447	0.491	0.288	0.010	0.269	0.308
2	Badulla	0.234	0.010	0.214	0.255	0.476	0.018	0.445	0.516	0.290	0.018	0.245	0.320
3	Bandarawela	0.234	0.010	0.217	0.256	0.486	0.012	0.462	0.509	0.280	0.011	0.258	0.300
4	Bibile	0.240	0.010	0.220	0.262	0.457	0.013	0.429	0.480	0.303	0.012	0.282	0.326
5	Buttala	0.230	0.010	0.210	0.248	0.492	0.012	0.471	0.517	0.278	0.011	0.256	0.297
6	Ella	0.250	0.011	0.230	0.272	0.487	0.013	0.463	0.512	0.263	0.011	0.241	0.284
7	Haldummulla	0.248	0.010	0.228	0.270	0.490	0.014	0.463	0.517	0.262	0.015	0.234	0.291
8	Hali-Ela	0.253	0.011	0.234	0.276	0.482	0.011	0.459	0.505	0.265	0.011	0.242	0.286
9	Haputale	0.233	0.009	0.213	0.250	0.499	0.012	0.477	0.525	0.268	0.010	0.247	0.288
10	Kandaketiya	0.249	0.010	0.230	0.270	0.449	0.014	0.421	0.475	0.302	0.012	0.280	0.326
11	Katharagama	0.227	0.011	0.205	0.247	0.477	0.012	0.454	0.503	0.296	0.011	0.274	0.319
12	Lunugala	0.255	0.011	0.232	0.276	0.482	0.017	0.447	0.515	0.263	0.017	0.228	0.293
13	Madulla	0.253	0.010	0.235	0.275	0.452	0.014	0.423	0.478	0.295	0.011	0.273	0.319
14	Mahiyanganaya	0.222	0.010	0.201	0.241	0.491	0.011	0.468	0.511	0.288	0.011	0.268	0.308
15	Medagama	0.242	0.010	0.222	0.261	0.469	0.013	0.444	0.495	0.289	0.012	0.266	0.313
16	Meegahakivula	0.262	0.009	0.244	0.281	0.455	0.014	0.427	0.481	0.283	0.012	0.259	0.307
17	Moneragala	0.246	0.009	0.230	0.266	0.462	0.011	0.440	0.483	0.292	0.009	0.273	0.310
18	Passara	0.250	0.010	0.230	0.272	0.480	0.015	0.452	0.507	0.270	0.016	0.241	0.300
19	Rideemaliyadda	0.233	0.011	0.210	0.254	0.500	0.016	0.473	0.532	0.267	0.014	0.239	0.294
20	Sevanagala	0.223	0.011	0.201	0.242	0.485	0.012	0.463	0.509	0.293	0.010	0.273	0.314
21	Siyambalanduwa	0.262	0.013	0.239	0.288	0.456	0.016	0.423	0.489	0.282	0.015	0.252	0.310
22	Soranathota	0.260	0.013	0.238	0.286	0.459	0.014	0.432	0.485	0.281	0.011	0.259	0.304
23	Thanamalvila	0.239	0.010	0.218	0.260	0.497	0.016	0.467	0.529	0.264	0.015	0.235	0.291
24	Uva Paranagama	0.268	0.014	0.242	0.296	0.443	0.013	0.418	0.468	0.289	0.012	0.267	0.311
25	Welimada	0.241	0.009	0.223	0.259	0.480	0.011	0.461	0.503	0.279	0.009	0.260	0.296
26	Wellawaya	0.231	0.010	0.211	0.249	0.472	0.010	0.451	0.492	0.298	0.009	0.281	0.318

2.5.2 Modeling of the multidimensional poverty status to create the small area estimates of poverty for the Uva province

The main goal of this subsection is to provide the proportions of poor people for each of the 26 small areas. We discuss how the different models performed when they are used to create small area estimates using the real-life data collected and compiled (as discussed in the previous section 2.2) by [45]. Again, we restrict all prior choices to the proper Weakly Informative Priors (WIP). That is, hyperparameters at Level 2 of the described hierarchical model in 2.3 are given the following priors:

$$\gamma_p \overset{ind}{\sim} N(0, 5^2),\ p = 1, \cdots 7,\ \text{and}\ \sigma_v \sim \text{Cauchy}^{+}(0, 5).$$

We run each model in 4 parallel chains, with $10,000$ iterations in each chain. We discard the first $5,000$ as burn-in samples from each chain. Thus, all the posterior analyses are based on $20,000$ post-warm-up MCMC draws.

After fitting all the 3 models, we use a Model Selection criterion to select one 'best' working model. The following Table 2.2 is used to select our working model.

Refer to subsection 1.5 for the details of the measures used in Table 2.2. From this table, we see that the LOO information criterion selects NL_{rs} as the best working model, and it is better than the other 3 models as seen by the `elpd_diff`. Now, even though the NL_{rs} model is not significantly better for the given data, we still prefer this model over the other three because of the various advantages of having a random-sampling variance component in the model. We discuss such advantages in detail in the following paragraphs.

Table 2.2: LOO-CV comparison table for four competing models – NL_{rs} is the Normal-Logistic Random Sampling model given in Section 2.4; FH is the model proposed by [1] (also see [2], [3]); NL is the univariate normal-logistic model, a special case of the model proposed in Section 2.3 (also see [3] and [2]); NL_{plugin} is the random sampling variance model in Section 2.4 where π_i in D_i is replaced by the Bayes estimate of π_i. LOO-IC selects NL_{rs} as the best of the 4 models.

	elpd_diff	se_diff
NL_{rs}	0	0
FH	-0.059	0.226
NL	-0.116	0.232
NL_{plugin}	-0.835	0.284

We had 6 area level covariates from the Census data, as described in section 2.2. We start by including all of them as the covariates, but the posterior analysis suggests that not all of them are significant. In Table 2.3, we report the Analysis of Variance (ANOVA) table from the full-model fit. The 'mean' column gives the posterior means, 'se_mean' column gives the Monte-Carlo standard errors, the 'sd' column gives the posterior standard deviations (standard errors of the Bayes estimates), the next 2 columns give the 2 percentiles (15^{th} and 85^{th} percentiles respectively), the column n_eff gives the Effective Sample Sizes (ESS, n_{eff}) – the number of independent samples that is used to replace the total n dependent MCMC draws having the same estimation power as the n autocorrelated samples. Finally, the column 'Rhat' gives a measure of chain-equilibrium ($\hat{R}$) – a value near 1 (but < 1.05) indicates that the chains have mixed well, and the posterior samples can be used with confidence for the posterior analyses (see [34] for details). This ANOVA table suggests that only the 'floor' and 'wall' are significant at 70%.

Table 3.1 gives a checklist summary of the 'Full' model and the 'Subset' model. Table 2.5 provides the ANOVA table of the subset model with the 2 significant covariates from Table 2.3. So we use the only 2 covariates that were significant at 70% – 'floor' and 'wall'. Since LOO

Table 2.3: Posterior summary statistics for the Normal-Logistic Random Sampling model parameters of the full model—with all 6 area-level covariates and 1 intercept

	mean	se_mean	sd	15%	85%	n_eff	Rhat
Intercept	0.237	0.004	0.167	0.073	0.402	1,649.193	1.007
floor	-0.553	0.018	0.364	-0.919	-0.203	407.363	1.014
ownsMobile	-0.057	0.012	0.458	-0.504	0.395	1,393.808	1.006
waterSafe	0.023	0.011	0.350	-0.341	0.367	1,092.918	1.007
ownsTV	0.324	0.015	0.455	-0.127	0.774	915.239	1.006
wall	0.304	0.005	0.213	0.093	0.508	1,960.126	1.004
toilet	0.227	0.009	0.266	-0.042	0.496	969.290	1.003
σ_v^2	0.210	0.016	0.290	0.011	0.429	324.210	1.012

Table 2.4: A list of competing models

Model	Intercept	floor	ownsMobile	waterSafe	ownsTV	wall	toilet
Full Model	✓	✓	✓	✓	✓	✓	✓
Subset Model	✓	✓				✓	

Table 2.5: Posterior summary statistics for the Normal-Logistic Random Sampling model parameters of the subset model—the best Model chosen by LOO criterion with the significant (at 70%) model parameters

	mean	se_mean	sd	15%	85%	n_eff	Rhat
Intercept	0.253	0.003	0.158	0.097	0.413	2,894.011	1.001
floor	-0.178	0.003	0.164	-0.338	-0.013	3,323.882	1.001
wall	0.128	0.003	0.140	-0.012	0.266	2,222.158	1.001
σ_v^2	0.152	0.007	0.202	0.009	0.311	739.094	1.001

selects the NL_{rs} model, we report another LOO table comparing the NL_{rs} full model with all 6 covariates with the subset model with the 2 aforementioned covariates. Clearly, from table 2.6 we can see that the subset model with 2 covariates is significantly better than the full model with all 6 covariates.

Table 2.6: LOO-CV comparison table for the 'best' model – NL_{rs} subset model using only the 2 significant (at 70%) covariates—with its counterpart using all the 6 small-area level covariates. Both models include an intercept.

	elpd_diff	se_diff
NL_{rs} (subset)	0	0
NL_{rs} (full)	-2.850	1.909

Figure 2.1 presents estimates from the 3 hierarchical Bayesian (HB) models (see Table 2.2) and the direct survey-weighted estimates for the 26 small areas, organized by increasing sample sizes for each area. The figure also includes the overall estimate across all areas. It illustrates the limitations of direct survey-weighted estimates and the advantages of modeling for obtaining stable estimates. Notably, the model-based estimates alleviate issues seen in certain areas, such as 'Meegahakivula', where the direct survey-weighted estimate of 0 may seem implausible due to the limited sample size and the poverty status of the sole individual sampled. Overall, the HB model-based estimates offer reasonable and consistent results, even in areas with limited sample sizes.

Figure 2.2 demonstrates superiority of the NL_{rs} model compared to the other 2 HB models. NL_{rs} provides an idea of the variability of the sampling standard deviations when the sample survey data is expected to vary/change. We can observe less variability of the sampling standard deviations for all the small areas over the smoothed design-based standard deviations as were

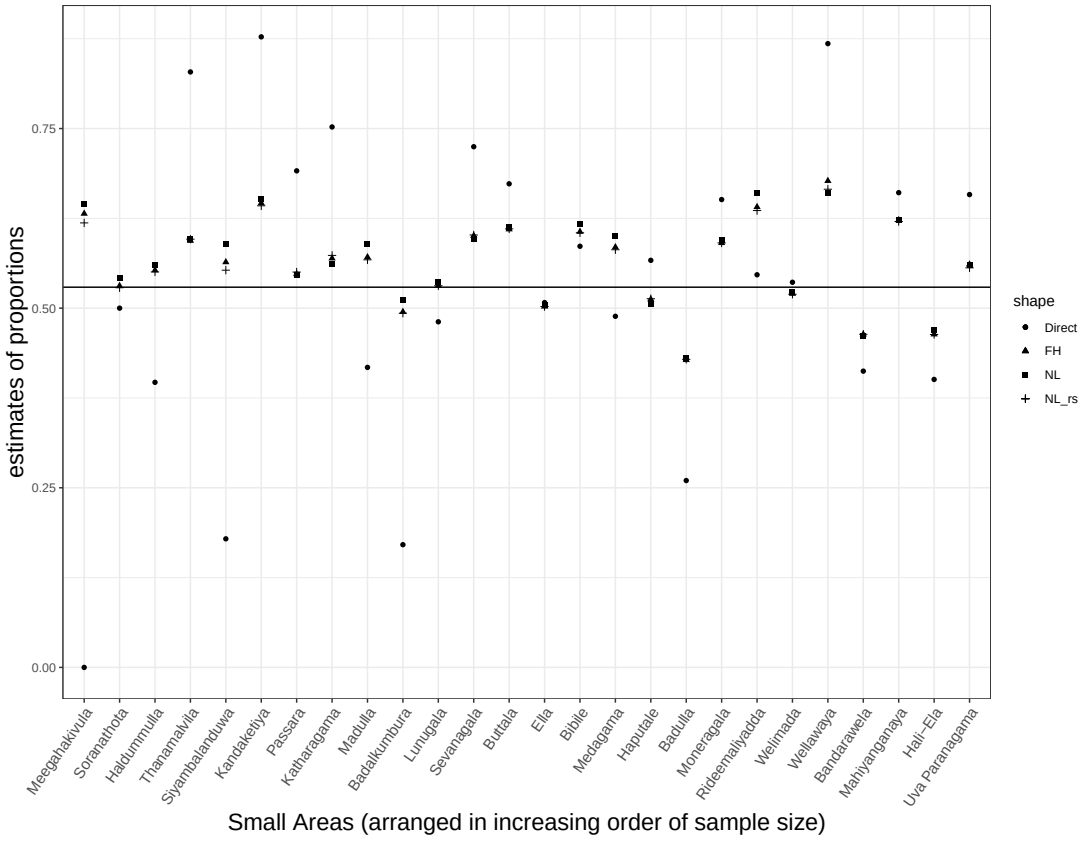


Figure 2.1: Comparison of direct survey-based estimates (Direct) and 3 Hierarchical model based estimates – Hierarchical Fay-Herriot (FH), Normal-Logistic (NL) and Normal-Logistic-Random-Sampling (NL_{rs}) models. The horizontal line refers to the overall estimate for all the areas. In the x-axis, the small areas are arranged in increasing order of sample size.

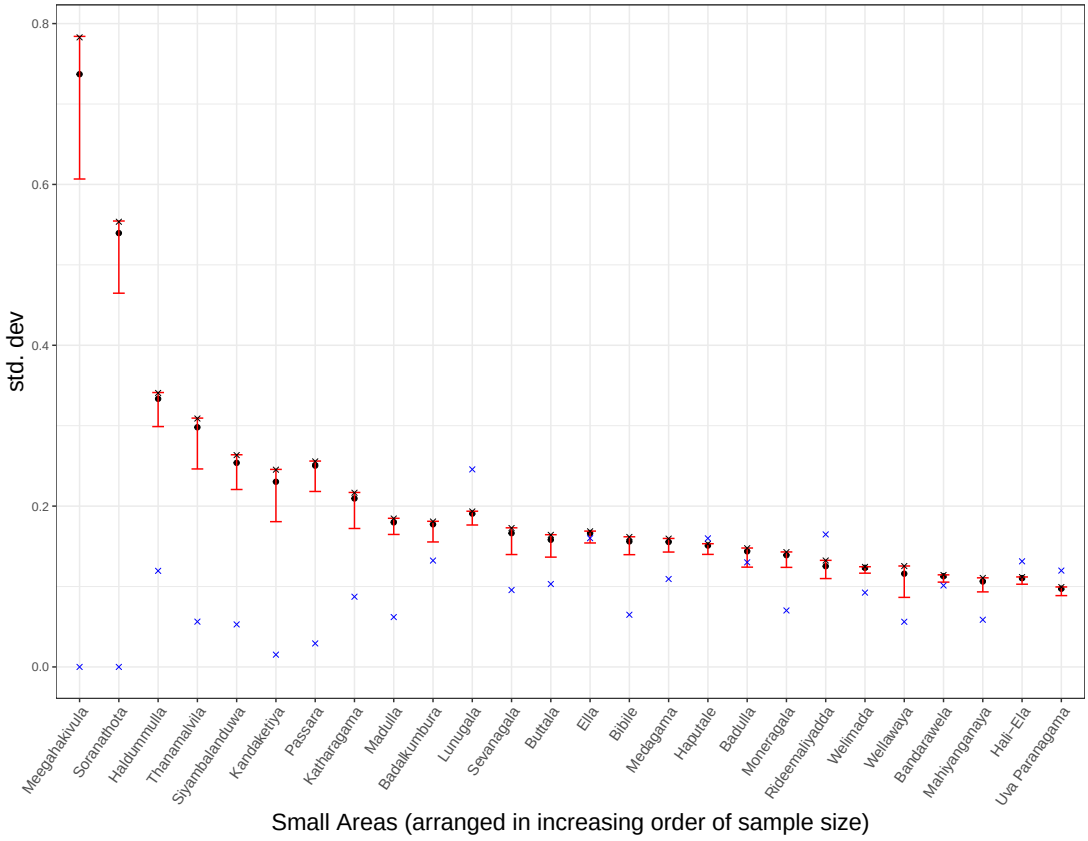


Figure 2.2: Interval plot of the sampling standard deviations—the red band gives the 95% posterior Credible Interval of the sampling standard deviations obtained from the NL_{rs} model. The black dot gives the posterior mean of the sampling standard deviation, and the black cross gives the smoothed sampling standard deviations used in the Fay-Herriot model and the Normal-Logistic model. The blue cross gives the design-based standard deviations. In the x-axis, the small areas are arranged in increasing order of sample size.

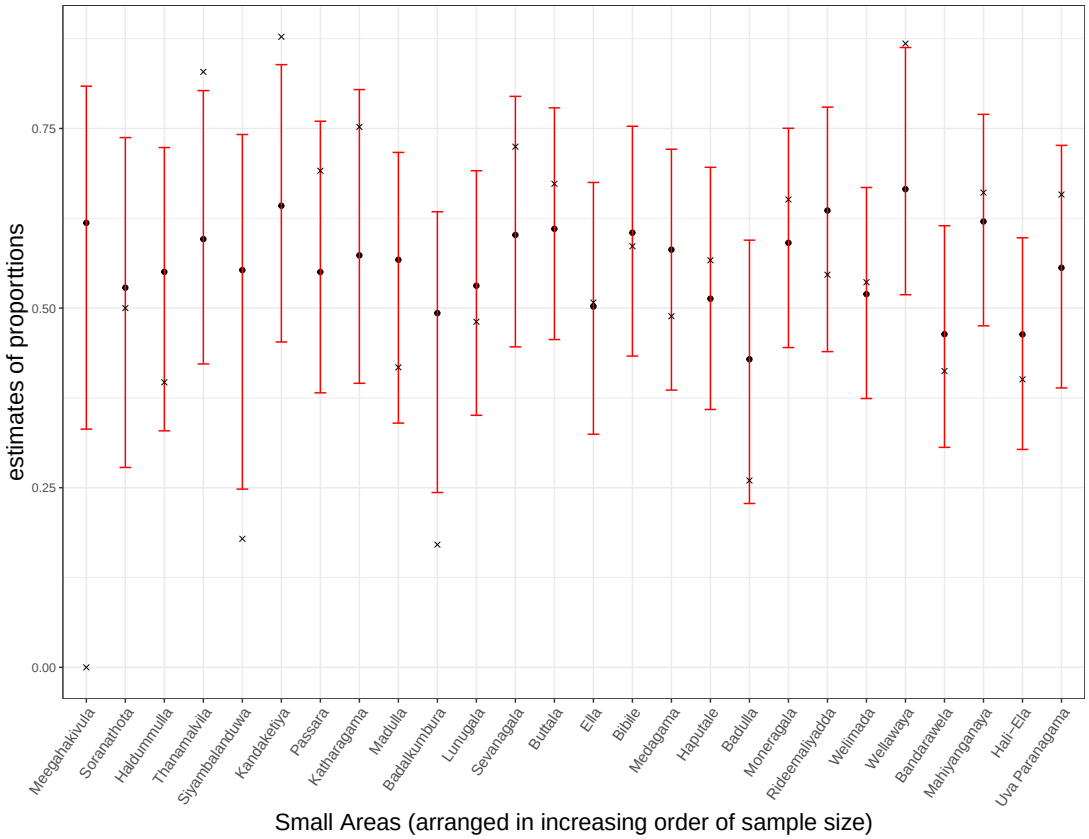


Figure 2.3: Interval plot of the estimates for 26 small areas – the red band gives the 95% posterior Credible Interval of the proportions obtained from the NL_{rs} model. The black dot gives the posterior mean of the proportions, and the black cross gives the survey-based estimates of the proportions. In the X-axis, the small areas are arranged in increasing order of sample size.

used in the other 2 HB models. Moreover, we can get an estimate of the variability and produce such plots with 95% posterior credible intervals of the standard deviations. The plot also shows the advantage of using model-based estimates instead of survey-weighted estimates for small areas. The direct estimates of the standard errors are very unstable, especially for areas with smaller sample sizes.

Figure 2.3 compares small area means, proportions in this case, from the NL_{rs} model (black dots) along with its 95% Bayesian credible interval and the estimates from the direct survey-weighted method. NL_{rs} not only gives us more believable and reliable estimates for areas with small (or no) samples, but it also gives us Bayesian credible intervals of the estimates.

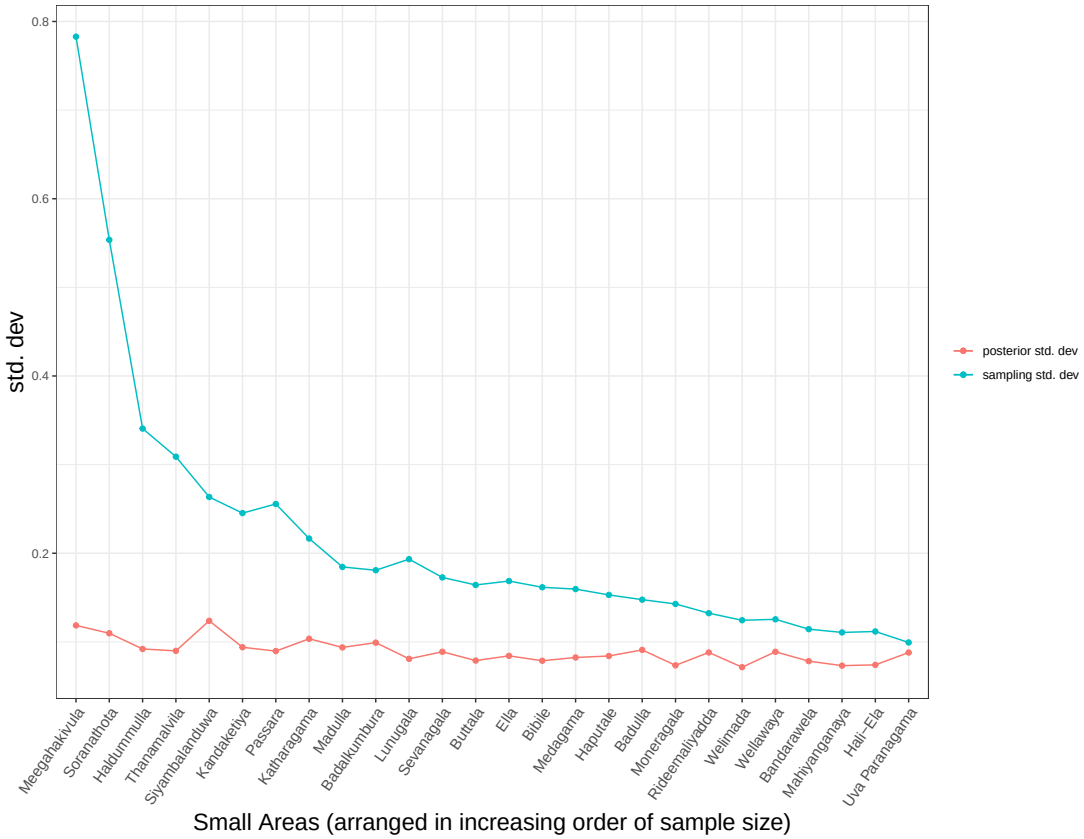


Figure 2.4: The red line gives the posterior standard deviations of the small area proportions obtained from the 'best' fitting NL_{rs} model. On the other hand, the turquoise line gives the smoothed sampling standard deviations ($\sqrt{D_i}$) that were used at level 1 (sampling level) of the FH and the NL models. In the x-axis, the small areas are arranged in increasing order of sample size.

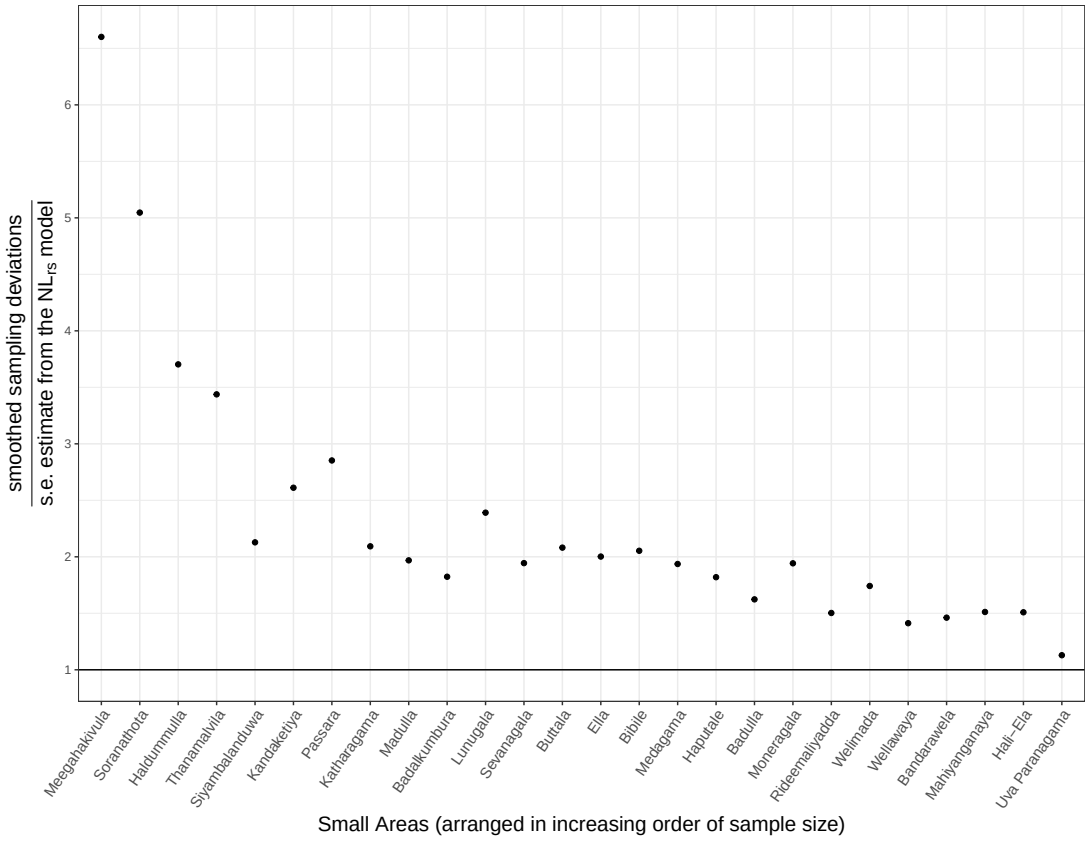


Figure 2.5: The horizontal line intersects the Y-axis at 1. In the X-axis, the small areas are arranged in increasing order of sample size. The ratio of the smoothed sampling deviations ($\sqrt{D_i}$) and the standard errors of the posterior small area means from the NL$_{rs}$ model.

As previously discussed, both the Fay-Herriot (FH) model and the Normal-Logistic (NL) model assume known sampling variances of the area-level estimates modeled at level 1. However, in practice, these variances are not known a priori and must be estimated. Figure 2.4 highlights the superiority of the NL$_{rs}$ model, which directly produces posterior standard error estimates. We compare these estimates with the smoothed sampling errors used in fitting the Hierarchical FH and NL models, noting the stability of standard error estimates, particularly in areas with smaller sample sizes. The efficiency gains are more pronounced for such areas, as evident in the ratios depicted in Figure 2.5, where all ratios exceed 1, with the highest observed for smaller sample sizes.

Table 2.7: Direct survey based estimates for the two districts with standard errors, 68% and 95% Confidence Intervals.

	district	design based estimate	se	2.5%	16%	84%	97.5%
1	Badulla	0.523	0.035	0.453	0.488	0.558	0.593
2	Moneragala	0.623	0.042	0.539	0.581	0.665	0.706

Table 2.8: Estimates for the two districts from the Normal Logistic random sampling model with standard errors, 68%, and 95% Confidence Intervals. The benchmarking ratio is displayed in the last column.

	district	NL_{rs} HB estimate	se	2.5%	16%	84%	97.5%	bm_ratio
1	Badulla	0.529	0.034	0.463	0.495	0.563	0.597	1.011
2	Moneragala	0.589	0.039	0.507	0.551	0.628	0.663	0.946

We also produce district-wise estimates for the two districts within the Uva province. Table 2.7 gives the district-wise direct survey-based estimates along with their standard errors and 68% and 95% confidence intervals. Table 2.8 gives the district-wise estimates from our Hierarchical Bayes NL_{rs} model. This table provides the same statistics as the previous one, but along with these, we also report the posterior mean of the benchmarking ratios (benchmarked to the direct district-wise estimates) for the two districts. These numbers are very close to 1. At least at the higher levels (districts are at a higher level than the DSDs–the small areas), model-based estimates are similar to the corresponding survey-weighted estimates, which are known to be reliable at higher levels with the availability of more samples [4].

Finally, in Table 2.9, we display the direct design-based estimates and our NL_{rs} model based estimates for a few interesting small areas.

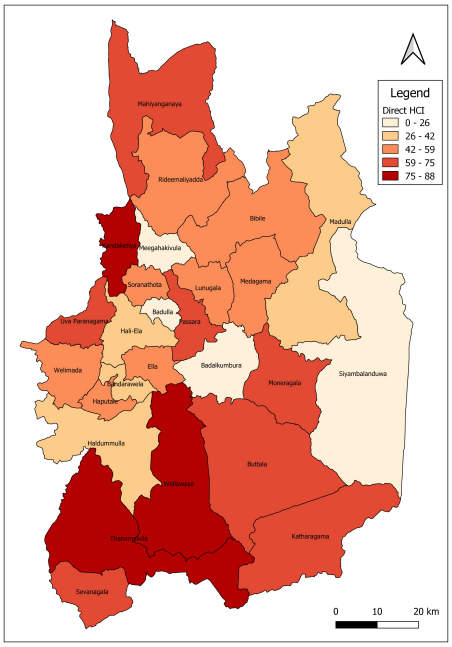
Table 2.9: Direct and HB model based estimates along with their standard errors for selected few small areas.

	small areas	n_i	$\tilde{n}_i$	Direct	Direct_se	NL_{rs}	NL_{rs}_se	$\sqrt{D_i}$
1	Meegahakivula	1	1	0	0	0.619	0.119	0.783
2	Soranathota	2	2	0.500	0	0.528	0.110	0.554
3	Kandaketiya	11	10.188	0.878	0.015	0.642	0.094	0.245
4	Lunugala	23	16.400	0.481	0.246	0.531	0.081	0.193
5	Sevanagala	23	20.544	0.725	0.096	0.602	0.089	0.173
6	Welimada	45	39.607	0.536	0.092	0.519	0.071	0.124
7	Uva Paranagama	73	62.146	0.658	0.120	0.556	0.088	0.099

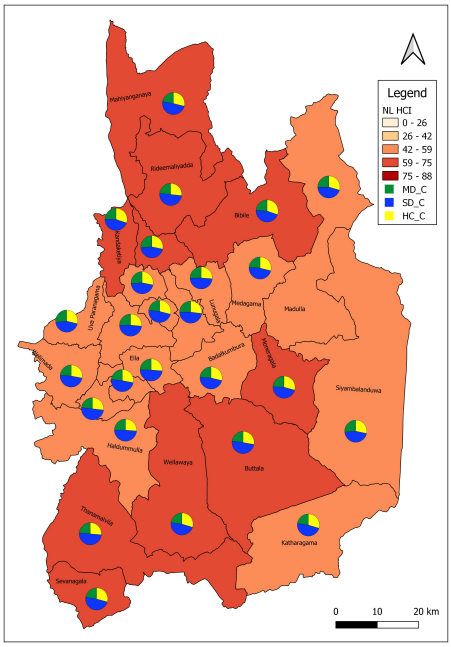
2.5.3 Maps at the lower administrative levels

In this final subsection of our analyses, we produce maps of different estimates of the proportions of poor people in the small areas of the Uva province in Sri Lanka. All of these maps were created using the free and open-source GIS software: 'QGIS version 2.18.15'.

Figure 2.6a gives the map of the direct survey-based estimates of the percentages of poor people for all the 26 small areas or the Divisional Secretary's Divisions (DSDs) in the Uva province. The problems with the direct survey-based estimates for small areas can be observed through this map—highly variable colors (estimates), anywhere from 0% to ∼88%, especially for the areas with fewer sample sizes. Figure 2.6b gives the map of the estimates of the percentages of poor people for all the 26 small areas from the 'best' chosen NL_{rs} model. In contrast to the highly variable direct estimates, we observe much more stabilized estimates, ranging from 43% to ∼67%. The pie-charts, placed inside the maps of the DSDs, give the estimated contribution (to the poverty percentages) of the 3 different dimensions that were used to create the composite measure of the multidimensional poverty [45]. 'MD_C' indicates the contribution share of the

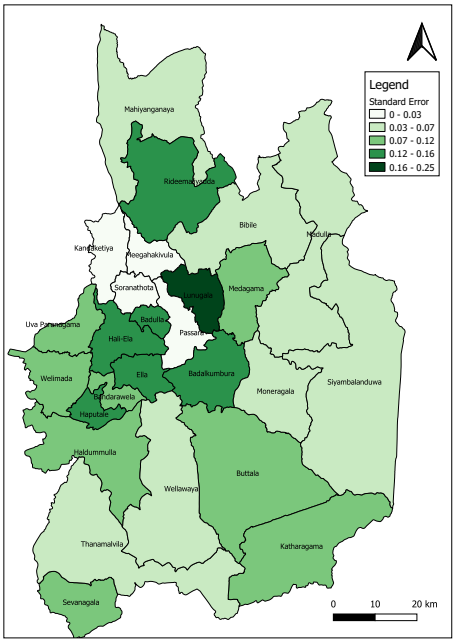


(a) The map of the direct survey based estimates of the percentage of poor people in all the 26 DSDs within the Uva province.

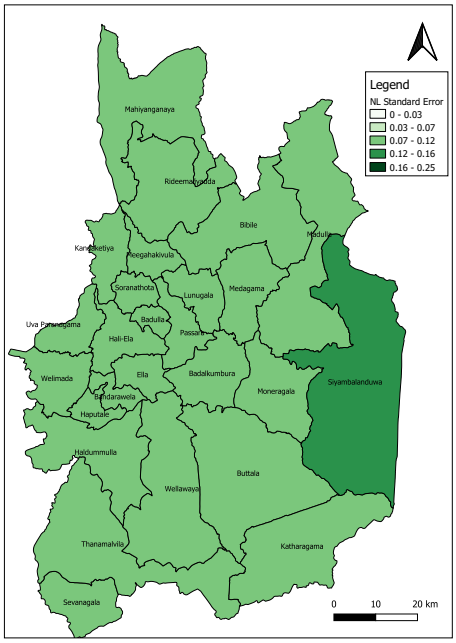


(b) The map of the Hierarchical Bayesian model (NL_{rs}) based estimates of the percentages of poor people in all the 26 DSDs within the Uva province. The pie charts within each area report the estimated contribution of the 3 different dimensions towards the composite multidimensional poverty measure.

Figure 2.6: comparison of Direct, and HB model based estimates



(a) The map of the standard errors of the direct survey-based estimates of the proportions of poor people in all the 26 DSDs within the Uva province.



(b) The map of the standard errors of the Hierarchical Bayesian model (NL_{rs}) based estimates of the proportions of poor people in all the 26 DSDs within the Uva province.

Figure 2.7: comparison of standard errors of Direct and HB model based estimates

'Material Deprivation' dimension, 'SD_C' indicates the contribution of the 'Social Deprivation' dimension, and finally, 'HC_C' indicates the contribution of the 'Human Capital' dimension. Figure 2.7a gives the map of the standard errors of the direct survey-based estimates for all the 26 small areas or the Divisional Secretary's Divisions (DSDs) of the Uva province. As with the map of the direct estimates, we can also observe highly unstable values of their corresponding standard errors, ranging from 0% to ∼25%. Figure 2.7b gives the map of the standard errors of the HB model (NL_{rs}) based on estimates of proportions of people who are poor according to the multidimensional measure of poverty for all the 26 DSDs–the small areas of the Uva province in Sri Lanka. We observe much more stabilized values of the standard errors as compared to that of the direct survey based estimates. They range from ∼7% to ∼12%.

2.6 Conclusion

There is now a sizable literature on developing multidimensional poverty measures. However, the estimation of these measures, especially for small areas, has received little importance. We need both reasonable multidimensional poverty measures and their estimation for making good and effective evidence-based public policies. In this chapter, we attempt to fill in the important research gap concerning the estimation of multidimensional poverty measures for small areas. Using a survey data specifically designed for obtaining information on multidimensional poverty, we demonstrate how a Bayesian method can be useful in estimating these poverty rates as well as the relative contributions of different dimensions towards the individual-level poverty for small areas. We considered the synthesis method for multidimensional poverty measure. However, our approach of small area estimation can be extended when counting or the fuzzy set method is used to define multidimensional poverty. In our multivariate normal hierarchical logistic model, we used a synthetic method for smoothing sampling covariance matrices. For estimating the proportion of impoverished people in small areas, using the multidimensional poverty status, we had to employ a unique way of determining effective sample sizes, motivated by the complexity of the sampling design of the survey data. In the future, we will explore extending the model/s to incorporate random covariance matrices.

Chapter 3: A Unified Approach to Small Area Estimation for Finite Population Sampling

3.1 Introduction

[50] laid a foundation for a subjective Bayesian approach to finite population sampling. In this approach, the entire matrix of values of characteristics of interest for all units of the finite population can be viewed as the finite population parameter matrix. In practice, a function (e.g., the finite population means or proportion of a characteristic of interest) or a vector of functions (e.g., finite population means or proportions of several characteristics) of this finite population parameter matrix is considered for inference. Using a subjective prior on the finite population parameter matrix, inferences on the finite population parameter(s) of interest can be drawn using the posterior predictive distribution of the unobserved units of the finite population given the observed sample.

Using this basic idea of [50], several papers were written on the estimation of finite population parameters. The methodology developed can be used to solve problems in small area estimation, repeated surveys, and other important applications. The papers can be broadly classified as Empirical Bayesian or EB (e.g., [51], [52], [53], [54], [55], [56], among others) and Hierarchical Bayesian or HB (e.g., [57], [58], [59], [60], [24], [61], [62], [63], [64], [65], [66],

and others). In an empirical Bayesian approach, hyperparameters are estimated using a classical method (e.g., maximum likelihood). In contrast, in the hierarchical Bayesian approach, priors – usually noninformative or weakly informative – are put on the hyperparameters.

The greater accessibility of administrative and Big Data and advances in technology are now providing new opportunities for researchers to solve a wide range of problems that would not be possible using a single data source. However, these databases are often unstructured and are available in disparate forms, making inferences using data linkages quite challenging. There is, therefore, a growing need to develop innovative statistical data linkage tools to link such complex multiple datasets. Using only one primary survey to answer scientific questions about the whole population or large geographical areas or subpopulations may be effective and reliable, but using it for smaller domains or small areas can often lead to unreliable estimation and unrealistic measures of uncertainty. One can often obtain reliable estimates for small areas using an appropriate statistical model to combine information from multiple data sources. A good review of different approaches to small area estimation can be found in [27], [67], [68], [4], [69], among others.

[70], [71], [72], and others discussed the concept of informative sampling under which the distribution of the sample could be markedly different from the one assumed for the finite population even after conditioning on related auxiliary variables. Most of the papers cited in the preceding paragraphs assume non-informative sampling, so the distribution of the sample is assumed to be identical to the distribution assumed on the finite population. This could be a strong assumption in many applications. The informative sampling approach suggested by [71] and [73] is one possible solution, but this approach requires additional modeling of survey weights. [74] suggested modeling the finite population given the inclusion probabilities (or, equivalently, basic

weights) for making inferences about the finite population parameters of interest. However, in practice, inclusion probabilities for all finite population units or detailed design information may not be available. Also, survey data may contain information only on the final weights for the respondents that incorporate nonresponse and calibration adjustments. [75] proposed an alternative approach in which inclusion probabilities are used to augment the sample model. Unlike [74], their approach needs weights only for the sample when estimating area means. Both [73] and [75] considered empirical best linear unbiased prediction (EBLUP) of small area means.

Our proposed methodology here differs from Multilevel Regression and Poststratification (MRP) [37] in the poststratification stage of MRP. Our way of estimating the population parameter of interest is inherently different from MRP. MRP only utilizes the known N_j's – the known population size in the j^{th} poststratified cell – whereas we utilize such N_j's along with the survey weights provided in the primary survey. Including the survey weights directly into the modeling step may reduce the extent of informativeness of the survey sample. Besides, MRP relies entirely on the model for the estimation. It uses the model to impute (predict) the values for all the poststratified cells. We do not do that. Instead, we take a different approach, as we have discussed in detail in section 3.2 and in its subsections, specifically in subsection 3.2.3. We also give a detailed explanation of how we go from the model parameters to the small-area population parameter of interest. Although the modeling of the MRP and the models we use here are similar, we emphasize how different and disparate relevant sources of information may be utilized at the modeling stage itself. Our method can be viewed as an alternative to MRP, building upon a comparable modeling technique (but different) and serving the same exact goal of reliably estimating the population parameter(s) of interest for small areas. In addition, we want to emphasize how similar models (as are used in MRP and in our methodology) can be employed to statistically

link disparate sources of data to make robust inferences for small area parameters of interest. And lastly, our way of estimating the population parameters is markedly different from that used in MRP.

In this research, we propose a hierarchical Bayesian approach to estimate finite population means for small areas. Our proposed method combines information from multiple disparate data sources such as probability surveys, non-probability surveys, administrative data, census records, social media big data, and/or any sources of available relevant information. We propose to reduce the degree of informativeness in sampling by incorporating important auxiliary variables that potentially affect the outcome variable of interest. Following [75], we also add the survey weights to our model. Our approach differs inherently from the existing small area Bayesian approach to the finite population sampling, which typically assumes a hierarchical model for all units of the finite population. We assume such an elaborate model only for the units of the finite population in which the outcome variable was observed. The reason is intuitive; for these units, the model can be checked using existing statistical tools.

To make reasonable modeling assumptions, we propose to form several cells for each of the small areas using factors that potentially influence our outcome variable of interest. The number of cells is carefully chosen so the cell population sizes can be obtained from reliable sources (e.g., population projection data). This strategy is expected to bring some degree of homogeneity within a given cell and also among cells from different small areas that are constructed with the same factor-level combination. However, unlike Multilevel Regression and Poststratification (MRP) [37], our cell construction will not achieve full homogeneity either in terms of the outcome variable of interest or survey weights. However, we stress that achieving full homogeneity in the outcome variable and weights is likely to require numerous cells for which reliable population

size data may not be available from population projection data, and one may have to rely on some unstable survey data for population counts of the small cells.

In contrast to the usual modeling approaches under similar scenarios (e.g., MRP), we do not assume an exchangeable model within cell or an elaborate model for unobserved individual units because the assumed model cannot be checked in the absence of data on the outcome variable. Instead, drawing inspiration from synthetic methods for small areas, we assume that the population means of the study variable in the cells with the same combination of factor levels are identical across small areas and the population mean for a cell is identical to the mean of true values for the observed units in that cell. Since the sample size in a given area is small, many such cells are unrepresented in the sample. If a sample for that cell can be found from other areas, the assumed model is used to produce a hierarchical Bayes synthetic estimate of the finite population mean of the study variable in that cell for the area. This assumes that the observations within a given cell and area are similar to those for the cell from other areas. If the cell is unrepresented in all areas, the cell mean can be predicted from the assumed population model. But to simplify the methodology without having the knowledge of the sampling's informativeness and to induce greater stability towards model misspecifications, we can simply ignore that cell when the contribution from that cell is negligible, as is the case in our illustrative example. The basic idea is to make the proposed methodology resistant to a possible violation of the assumed model.

As an application of the proposed methodology, we use a real-life example, using a COVID-19 probability survey representing the entire US adult population, a non-probability survey representing only active US adult Facebook users, and Census Bureau estimates of adult population counts at granular levels along with data from an independent COVID-19 data reporting website, to estimate the vaccine hesitancy rates (proportions) for the US states and the District of

Columbia (small areas). Through this example and application, we will demonstrate the problems with regular design-based estimates when used for small areas and how our methodology may be employed to get better estimates along with stable measures of uncertainty.

We now give an outline for the rest of the chapter. In section 3.2, we describe our methodology in detail. In section 3.3, we describe an application of our method to a real-life survey. In section 3.4, we show the results we obtained with this real-life application. Finally, we summarize and conclude our chapter by giving future research directions for our methodology.

3.2 Methodology

For each small area i $(i = 1, \cdots, m)$, we partition the finite population into G cells using factors available in the survey data that potentially influence the outcome variable of interest. We assume that population sizes for these cells are known from the census data. Let N_i and N_{ig} be the known population sizes of the area i, and of the g^{th} cell within the i^{th} area respectively, $i = 1, \cdots, m; g = 1, \cdots, G$. Naturally, $\sum_g N_{ig} = N_i$, and $\sum_i N_i = N$, N being the finite population size.

Let y_{igk} denote our outcome variable of interest for the k^{th} unit belonging to the g^{th} cell belonging to the i^{th} small area $(i = 1, \cdots, m;\ g = 1, \cdots, G,\ k = 1, \cdots, N_{ig})$. We assume that we have access to different types of known auxiliary variables to model our outcome variable. Let $x_{g(1)}$ be a vector of indicator variables corresponding to cell g, $x_{i(2)}$ be a vector of known auxiliary variables specific to area i and $x_{igk(3)}$ be a vector of individual level auxiliary variables $(i = 1, \cdots, m;\ g = 1, \cdots, G,\ k = 1, \cdots, N_{ig})$.

Suppose we have a sample of size n from the finite population. Let G_i be the number of

cells represented in the sample for area i ($i = 1, \cdots, m$). Let n_i and n_{ig} be the sample sizes for area i, and for cell g within area i, respectively, so that $\sum_{g=1}^{G_i} n_{ig} = n_i$ and $\sum_{i=1}^{m} n_i = n$. It is possible to have $n_i = 0$ for some area i, or $n_{ig} = 0$ for some cell g within an area i; $i = 1, \cdots, m; g = 1, \cdots, G$. Let w_{igk} denote the survey weight for k^{th} sampled respondent belonging to the g^{th} cell residing in the i^{th} small area ($i = 1, \cdots, m;\ g = 1, \cdots, G_i,\ k = 1, \cdots, n_{ig}$).

3.2.1 Parameter of interest

We are interested in estimating finite population means for all areas, i.e.,

$$\bar{Y}_i = N_i^{-1} \sum_{g=1}^{G} \sum_{k=1}^{N_{ig}} y_{igk} = \sum_{g=1}^{G} \frac{N_{ig}}{N_i} \bar{Y}_{ig}, \tag{3.1}$$

where $N_i = \sum_g N_{ig}$, population size for the i^{th} area ($i = 1, \cdots, m$).

3.2.2 Proposed hierarchical models for sampled respondents and the finite population

Our approach calls for the specification of an explicit model for the outcome variable y for all the N units of the finite population. Given θ_{igk}, we assume y_{igk} are independent with mean θ_{igk} $i = 1, \cdots, m;\ g = 1, \cdots, G,\ k = 1, \cdots, N_{ig}$.

We will assume an elaborate model for all respondents because we can evaluate such a model through careful model building steps in the actual data analysis. For $i = 1, \cdots, m;\ g =$

$1, \cdots, G_i,\ k = 1, \cdots, n_{ig}$, we propose the following model for the sampled respondents:

$$\begin{aligned}
&\text{Level 1:} \quad y_{igk}|\theta_{igk} \overset{ind}{\sim} f_1[\theta_{igk}, V(\theta_{igk})], \\
&\text{Level 2:} \quad \phi(\theta_{igk}) = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk}\xi_c + \lambda h(w_{igk}) + v_i + u_{ig}, \\
&\text{Level 3:} \quad v_i \text{ and } u_{ij} \text{ are independent with } v_i \overset{iid}{\sim} f_2[0, \sigma_v^2],\ u_{ig} \overset{iid}{\sim} f_3[0, \sigma_u^2].
\end{aligned} \tag{3.2}$$

where $f_j[a, b]$ $(j = 1, 2, 3)$ denotes a known mass or density function with mean a and variance b. For example, we can choose f_1 from NEF-QVF$[\mu, V(\mu)]$, the Natural Exponential Family with mean μ and Quadratic Variance Function $V(\mu) = v_0 + v_1\mu + v_2\mu^2$, where v_0, v_1 and v_2 are constants. The family includes the Gaussian distribution: $N(\mu, \sigma^2)$ $(v_0 = \sigma^2, v_1 = v_2 = 0)$, the Gamma distribution: $G(r, \lambda)$ $(\mu = r\lambda, v_0 = v_1 = 0, v_2 = \frac{1}{r})$, the Poisson distribution: $P(\lambda)$ $(\mu = \lambda, v_0 = v_2 = 0, v_1 = 1)$, the Binomial distribution: $B(r, \theta)$ $(\mu = r\theta, v_0 = 0, v_1 = 1, v_2 = -\frac{1}{r})$, the Negative-Binomial distribution: $NB(r, \theta)$ $(\mu = r\frac{\theta}{1-\theta},\ v_0 = 0, v_1 = 1, v_2 = \frac{1}{r})$, etc. For the theoretical properties of NEF-QVF, please refer to [76]. As for f_2 and f_3, we can choose a normal density.

As for level 2, we can choose $\phi()$ to be a suitable known link function (e.g., if f_1 is a Binomial distribution, we can choose $\phi()$ as logit link, and if f_1 is a Normal distribution, $\phi()$ then can be the identity link, etc.), and ξ_c a vector of fixed unknown coefficients (the suffix c in ξ_c allows dependence of the regression coefficients on some combination of the factors that form the demographic cells – c will be determined through the model selection step); $h(\cdot)$ is a known function of survey weights to be determined by a model selection step and λ is a fixed unknown coefficient. [75] considered multiple choices of $h(\cdot)$. If the data analysis suggests $\lambda = 0$, we can assume non-informative sampling and carry out the analysis without this additional term. In

Level 2 we added the survey weights as an additional auxiliary variable to reduce the extent of informativeness. As for level 3, we can take f_2 and f_3 as normal densities for the area specific random effects v_i, and area and groups specific random effects u_{ig} respectively. These random errors with a mean of 0 will take care of any leftover variabilities that were not captured by the fixed main effects, and their interactions that we had incorporated in the model at the level 2.

We now discuss our modeling assumptions for the remaining $N-n$ unobserved units of the finite population. To this end, for the small areas $i = 1, \cdots, m$, define:

$\mathcal{G}_{1i} = \{g \ni n_{ig} > 0,\ g = 1, \cdots, G\}$: the set of cells represented in the sample for the area i.

$\mathcal{G}_{2i} = \{g \ni n_{ig} = 0, n_{jg} > 0, \text{for some } j \neq i, g = 1, \cdots, G\}$: the set of cells not represented in the sample for the area i, but represented by one or more other areas $j \neq i$. And finally,

$\mathcal{G}_{3i} = \{g \ni n_{ig} = 0, \forall i,\ g = 1, \cdots, G\}$: the set of cells not represented in the sample for area i.

Define the population mean for the set of cells $\mathcal{G}_{1i}$ as,

$$\bar{\Theta}_{1i} = \frac{\sum_{g\in\mathcal{G}_{1i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{1i}}N_{ig}} = \sum_{g\in\mathcal{G}_{1i}} b_{ig;1}\bar{\Theta}_{ig},\ \text{where } b_{ig;1} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{1i}}N_{ig}}\ \text{ and } \bar{\Theta}_{ig} = \frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}}.$$

Similarly, we define

$$\bar{\Theta}_{2i} = \frac{\sum_{g\in\mathcal{G}_{2i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\bar{\Theta}_{ig},\ \text{where } b_{ig;2} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}} \text{ and } \bar{\Theta}_{ig} = \frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}},$$

and

$$\bar{\Theta}_{3i} = \frac{\sum_{g\in\mathcal{G}_{3i}}\sum_{k=1}^{N_{ig}}\theta_{igk}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}} = \sum_{g\in\mathcal{G}_{3i}} b_{ig;3}\bar{\Theta}_{ig},\ \text{where } b_{ig;3} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}} \text{ and } \bar{\Theta}_{ig} = \frac{\sum_{k=1}^{N_{ig}}\theta_{igk}}{N_{ig}}.$$

as the population means for the set of cells $\mathcal{G}_{2i}$ and the set of cells $\mathcal{G}_{3i}$, respectively. Here, we

make a synthetic assumption that:

$$\bar{\Theta}_{2i} = \sum_{g \in \mathcal{G}_{2i}} b_{ig;2} \bar{\Theta}_{ig;\text{syn}} \text{ where, } \bar{\Theta}_{ig;\text{syn}} = \frac{\sum_{j \neq i} \sum_k \theta_{jgk}}{\sum_{j \neq i} N_{jg}}$$

i.e, $\bar{\Theta}_{2i}$ is identical to the overall mean of the cells in $\mathcal{G}_{2i}$ obtainable from one or more areas (other than area i). We make no assumption on the rest of the finite population units. Since population sizes N_{ig} are in general large, appealing to the law of large numbers as in [28], we can approximate the finite population means $\bar{Y}_i$ in equation 3.1, with the following functions of parameters of the assumed finite population model.

$$\begin{aligned} \bar{Y}_i \approx \bar{\Theta}_i = \sum_{g=1}^{G} \frac{N_{ig}}{N_i} \bar{\Theta}_{ig} &= a_{1i}\bar{\Theta}_{1i} + a_{2i}\bar{\Theta}_{2i} + (1 - a_{1i} - a_{2i})\bar{\Theta}_{3i}, \text{(say)} \\ &\approx a_{1i}\bar{\Theta}_{1i} + a_{2i}\bar{\Theta}_{2i}, \end{aligned} \tag{3.3}$$

where $a_{1i} = N_i^{-1} \sum_{g \in \mathcal{G}_{1i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{1i}$; $a_{2i} = N_i^{-1} \sum_{g \in \mathcal{G}_{2i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{2i}$; $a_{3i} = N_i^{-1} \sum_{g \in \mathcal{G}_{3i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{3i}$. The approximation in equation 3.3 will only be reasonable when $(1 - a_{1i} - a_{2i}) \approx 0$.

3.2.3 A Bayesian implementation of the hierarchical model

As is the common practice with most implementations of HB methodology, the model parameters will be estimated using Monte Carlo Markov Chain (MCMC) method of sampling from posterior distributions. Assume weakly informative priors on the hyperparameters: β, and σ_v. At

each MCMC iteration $(r = 1, \ldots, R)$ step, the following is generated for the area i:

$$\bar{\theta}_i^{(r)} = a_{1i}\bar{\theta}_{1i}^{(r)} + a_{2i}\bar{\theta}_{2i}^{(r)} + (1 - a_{1i} - a_{2i})\bar{\theta}_{3i}^{(r)} \approx a_{1i}\bar{\theta}_{1i}^{(r)} + a_{2i}\bar{\theta}_{2i}^{(r)},\ r = 1, ..., R,$$

where

$$\begin{aligned}
\bar{\theta}_{1i}^{(r)} &= \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\theta}_{ig}^{(r)},\ \text{with } \bar{\theta}_{ig}^{(r)} = \frac{\sum_k w_{igk}\theta_{igk}^{(r)}}{\sum_k w_{igk}}, \\
\bar{\theta}_{2i}^{(r)} &= \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\theta}_{ig;\text{syn}}^{(r)},\ \text{with } \bar{\theta}_{ig;\text{syn}}^{(r)} = \frac{\sum_{j \neq i}\sum_k w_{jgk}\theta_{jgk}^{(r)}}{\sum_{j \neq i}\sum_k w_{jgk}}, \\
\theta_{igk}^{(r)} &= \phi^{-1}\left(x'_{g(1)}\alpha^{(r)} + x'_{i(2)}\beta^{(r)} + x'_{igk}\xi_c^{(r)} + \lambda^{(r)}h(w_{igk}) + v_i^{(r)} + u_{ig}^{(r)}\right).
\end{aligned}$$

To justify, note that

$$\bar{\Theta}_{1i} = \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\Theta}_{ig} \approx \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\theta}_{ig} = \bar{\theta}_{1i},\ \text{say},$$

where $\bar{\theta}_{ig} = \frac{\sum_k w_{igk}\theta_{igk}}{\sum_k w_{igk}}$; see [28] and [77] for similar justifications in empirical best prediction and hierarchical Bayes approaches, respectively. Also,

$$\bar{\Theta}_{2i} = \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\Theta}_{ig;\text{syn}} \approx \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\theta}_{ig;\text{syn}} = \bar{\theta}_{2i},\ \text{say},$$

where $\bar{\theta}_{ig;\text{syn}} = \frac{\sum_{j \neq i}\sum_k w_{jgk}\theta_{jgk}}{\sum_{j \neq i}\sum_k w_{jgk}}$.

The set of R MCMC replicates for the i^{th} area, i.e., $\{\bar{\theta}_i^{(r)}, r = 1, ..., R\}$, will be used for inferences about $\bar{\theta}_i$.

3.2.4 A special closed form of our proposed Bayes estimate

We now consider the following special case of our general small area model to understand our estimator.

$$\text{Level 1:} \quad y_{igk}|\theta_{igk} \overset{ind}{\sim} N(\theta_{igk}, \sigma_y^2),$$

$$\text{Level 2:} \quad \theta_{igk}|\Lambda_{igk}, v_i \overset{ind}{\sim} N(\Lambda_{igk} + v_i, \sigma_u^2)$$

$$\text{Level 3:} \quad v_i \text{ and } u_{ig} \text{ are independent with } v_i \overset{iid}{\sim} N(0, \sigma_v^2),\ u_{ig} \overset{iid}{\sim} N(0, \sigma_u^2),$$

where $\Lambda_{igk} = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk(3)}\xi_c + \lambda h(w_{igk})$. Level 2 of the same model can be reexpressed as: $\theta_{igk} = \Lambda_{igk} + v_i + u_{ig}$, where, $u_{ig} \sim N(0, \sigma_u^2)$. Under this consideration, and for the sake of simplicity of illustration, with an additional assumption of $\alpha, \beta, \xi_c, \lambda, \sigma_u^2$ and σ_v^2 being known; we can obtain a closed form of the proposed small area Bayes estimate given in equation 3.3. Naturally, this will be the 'Best Predictor' (BP) of the small area means ([4]). As already mentioned in the subsections 3.2.2, 3.2.3, and using the same notations and terminologies, the Bayes estimate for a given area i will be given by $E(a_{1i}\bar{\theta}_{1i} + a_{2i}\bar{\theta}_{2i}|\text{data})$. Under the Normality

of each level of the illustrative model, we obtain:

$$
\begin{aligned}
E(a_{1i}\bar{\theta}_{1i} + a_{2i}\bar{\theta}_{2i}|\text{data}) &= a_{1i}E(\bar{\theta}_{1i}|\text{data}) + a_{2i}E(\bar{\theta}_{2i}|\text{data}) \\
&= a_{1i}\sum_{g\in\mathcal{G}_{1i}} b_{ig;1} \times \frac{1}{\sum_k w_{igk}}\left[\sum_k w_{igk}\left\{\frac{\sigma_u^2 y_{igk} + \sigma_y^2\Lambda_{igk}}{\sigma_u^2+\sigma_y^2} + \right.\right. \\
&\left.\left. \frac{n_i\sigma_y^2\sigma_v^2(\bar{y}_{i..} - \bar{\Lambda}_{i..})}{(\sigma_u^2+\sigma_y^2)(n_i\sigma_v^2+\sigma_u^2+\sigma_y^2)}\right\}\right] + \\
&a_{2i}\sum_{g\in\mathcal{G}_{2i}} b_{ig;2} \times \frac{1}{\sum_{j\neq i}\sum_k w_{jgk}}\left[\sum_{j\neq i}\sum_k w_{jgk}\left\{\frac{\sigma_u^2 y_{jgk} + \sigma_y^2\Lambda_{jgk}}{\sigma_u^2+\sigma_y^2} + \right.\right. \\
&\left.\left. \frac{n_j\sigma_y^2\sigma_v^2(\bar{y}_{j..} - \bar{\Lambda}_{j..})}{(\sigma_u^2+\sigma_y^2)(n_j\sigma_v^2+\sigma_u^2+\sigma_y^2)}\right\}\right]
\end{aligned}
$$

where, $\bar{y}_{i..} = \frac{1}{n_i}\sum_{g,k} y_{igk}$. $\bar{\Lambda}_{i..}$, $\bar{y}_{j..}$, and $\bar{\Lambda}_{j..}$ are similarly defined.

From this simple illustration, under the usual assumptions made for estimating the best predictor, we can get an insight in the form of our Bayes estimate. We observe how the y_i's available from the groups in the sample for area i enter the estimator through the first summand. The groups, for which there is no representation in the sample for the same area, enter the estimator through the second summand–in terms of the samples available for those groups at one or more other areas.

3.3 A real-life application

In elucidating our proposed methodology, we examine a topic of contemporary research significance. Our focus is directed towards furnishing reasonably stable, state-wise (our small areas of consideration) estimates of the proportions of people who are "hesitant" to get vaccinated for COVID-19. In this section, we first provide an outline of disparate data sources that we have

considered in this application.

3.3.1 Primary data: Understanding America Study (UAS)

The Understanding America Study (UAS), undertaken by the University of Southern California (USC), comprises approximately 9,000 respondents, representing the entirety of the United States' adult population. Functioning as an Internet Panel, participants engage in surveys via internet-enabled devices at their convenience. Households, broadly defined, involve individuals residing with the primary participant. To extract valuable insights on attitudes, behaviors, and societal aspects, the Center for Economic and Social Research (CESR) at USC initiated the Understanding Coronavirus in America tracking survey on March 10, 2020. Utilizing the UAS panel, this ongoing survey, supported by the Bill & Melinda Gates Foundation and the National Institute on Aging, delves into various dimensions, including mental health, healthcare behavior, and the economic crisis during the COVID-19 pandemic.

The survey commenced with 8,547 eligible participants from the UAS panel, out of 9,603 respondents who expressed interest. Our study focuses on data collected during wave 16 (October 14–November 11, 2020), where 7,832 UAS participants received the survey. Among them, 6,081 completed the survey, constituting our respondents. The response rate was approximately 78%, with participants given a 14-day window for completion and compensated with $15 upon survey submission.

3.3.2 Direct UAS estimates

For the "hesitancy towards vaccination" analysis, we focused on responses to the survey question: 'How likely are you to get vaccinated for coronavirus once a vaccine is available to the public?' This question offered five response options: very unlikely, somewhat unlikely, somewhat likely, very likely, and unsure. We transformed the responses into a binary variable, assigning a value of 1 if the respondent answered anything other than 'very likely' and 0 otherwise. Utilizing this binary variable and the corresponding survey weights, we computed the direct survey-based estimates for the proportions of individuals who are *not* 'very likely' to get vaccinated across the 50 states and the District of Columbia. While the national estimate met high-quality standards, with margin of errors below 3%, estimates for smaller geographic areas (50 states and DC) were less reliable.

Table 3.7 presents direct survey estimates of vaccine hesitancy proportions for all 50 states and the District of Columbia, accompanied by their standard errors. Notably, we encountered an impractical estimate of 1 with an associated estimated standard error of 0 for Delaware. This anomaly arises because none of the respondents from Delaware chose 'very likely' in response to the survey question. However, this does not imply universal hesitancy in Delaware. The estimate for Alaska is 0.21 with a high associated standard error of 0.19, exceeding the typical range of 0.02. Similar patterns emerge for Vermont, Rhode Island, the District of Columbia, and other states.

3.3.3 Supplementary data descriptions

In the subsequent subsections, we delineate various supplementary data sources examined in the context of applying our methodology.

The COVID-19 Symptom Survey:

The COVID-19 Symptom Survey, a collaboration between public health scholars and Facebook, was hosted in the US by the Carnegie Mellon Delphi Research Center and internationally by the University of Maryland's Joint Program in Survey Methodology (JPSM). This non-probability survey, representative of adult Facebook users, gathered daily data. Seeking reliable state-level covariates, we incorporated static variables from the survey, focusing on responses to two questions pertaining to: (1) Pre-existing medical conditions, encoding 1 if respondents reported having diabetes, asthma, or chronic lung disease, and 0 otherwise; (2) Flu shots received in the last 12 months, encoding 1 for a "yes" response and 0 otherwise. Combining survey responses from August 5 to September 9, 2020, we calculated weighted state-wise estimates for individuals with specified diseases and those who received a flu shot. These estimates, derived as survey-weighted means of binary encoded variables, serve as continuous covariates in our model.

Census Bureau's Population Estimates Program (PEP):

The US Census Bureau's Population Estimates Program (PEP) provides annual population estimates for various geographic levels, including states, counties, cities, and towns, along with demographic components like births, deaths, migration, and housing units. Utilizing the rich PEP data, we generate state-wise and demographic cell-wise population counts, focusing on the `race x ethnicity x gender x age category` demographic cells. The demographics considered include four levels of `race` (White, Black, Asian, and Others), two levels

of `ethnicity` (Hispanic and Non-Hispanic), two levels of `gender` (Male and Female), and seven levels of `age category` ($18-24, 25-34, 35-44, 45-54, 55-64, 65-74, \&75+$). The data used pertains to 'Annual State Resident Population Estimates for 5 race groups (5 race alone or in combination groups) by age, sex, and Hispanic origin: April 1, 2010, to July 1, 2019'. Pre-processing involved grouping and collapsing data to obtain counts for the specified demographic cells (poststratified cells).

The COVID Tracking project: Data from `https://covidtracking.com` is utilized for our model, providing freely downloadable COVID-19 data for all 50 states and the District of Columbia. Two state-specific continuous covariates were created: 'Testing rate', calculated as the total tests with confirmed outcomes divided by the overall test counts in the state (acknowledging potential values exceeding 1 due to multiple tests for an individual), and 'Positivity rate', obtained by dividing total positive test outcomes by the total tests with confirmatory outcomes.

3.4 Data analysis and evaluation

We commence this section by noting that, given the available data, we undertook a comprehensive exploration of several variables. Through a series of trial and error steps, we arrived at a specific set of covariates deemed pertinent for our analysis. Subsequent considerations were also given to the selection of models for the model selection step, all of which were grounded in a careful assessment of the given problem and datasets. We advocate for applied scientists to exercise their judgment, drawing on considerations such as the nature of the problem and subject matter expertise, when conducting similar steps in their analyses. As previously discussed in subsection 3.3, we refer to 'RACE' by a having 4 levels – White, Black, Asian and Others,

'ETHNICITY' by b having 2 levels – Hispanic or Non-Hispanic and 'GENDER' by c having 2 levels – Male and Female. For the variable, 'AGE' we have divided the whole adult population into 7 cells, e.g., $18-24,\ 25-34,\ \ldots,\ 65-74$ & $75+$; denoted by d.

We consider fixed effect intercepts by 'RACE' x 'ETHNICITY' – so, in total, 8 fixed effect intercepts for all the models we considered. A list of competing models is described in Table 3.1.

As for the state level covariates $x_{i(2)}$ (refer to section 3.2 for reference to this notation), we use Facebook survey data, and data from the COVID tracking project. Facebook COVID symptom survey data is used to calculate two state-level summary variables – survey-weighted proportions of people having one of three pre-existing conditions or co-morbidity rates and survey-weighted proportions of people who took a flu shot in the last 12 months or Flu-Shot rates (as discussed in subsection 3.3). We use logit-transforms of these two rates as the covariates in our models described in Table 3.1. Other state level covariates we have used came from the COVID Tracking project, e.g., testing rates and positivity rates as defined previously. Finally, we have also used the state level percentages of republican votes that were cast in the 2020 United States presidential election as a state level covariate – 'Percent Republican' – in our model (also referred to as 'rep %' in table 3.6).

For the first respondent level covariate $x_{igk(3)}$, we used the respondents' 7 age-levels and converted them into numbers $1, \ldots, 7$ and used them as a continuous-scale covariate in all our models. We use gender-specific slopes – one for males and the other for females – for the age covariate in our model. Such a choice was motivated based on empirical studies that we had carried out. For the second respondent level covariate, we use the survey weights associated with every respondent from the UAS-survey data.

Table 3.1: A list of competing models

Model	Race x Ethnicity specific intercept	Gender specific slope of Age	Comorbidity Rate (Facebook)	Flu Shot Rate (Facebook)	Test Rate (CovidTrack)	Positivity Rate (CovidTrack)	Percent Republican	Survey Weight
M1	✓	✓	✓	✓	✓	✓	✓	✓
M2	✓	✓	✓	×	✓	×	✓	✓
M3	✓	✓	✓	×	×	×	✓	✓
M4	✓	✓	✓	×	×	×	✓	×

3.4.1 Model fit

Recall, the approximation we did to the population (here, 50 states and DC) parameter of interest $\bar{Y}_i$ in equation 3.3 would only hold when the quantity $(1 - a_{1i} - a_{2i}) \approx 0$. Table 3.2 summarizes the quantity $(1 - a_{1i} - a_{2i})$ for 50 states and Washington, DC. We can reasonably justify our approximation, since both the maximum value (0.0120) and the 3^{rd} quartile (0.0031) are negligible.

Table 3.2: Summary statistics of $(1 - a_{1i} - a_{2i})$ for 50 states and DC.

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
$1 - a_{i1} - a_{i2}$	0.0004	0.0010	0.0016	0.0026	0.0031	0.0120

Now we discuss the model fits. We consider the following special case of the general model proposed in equation 3.2:

$$
\begin{aligned}
&\text{Level 1:} \quad y_{igk}|\theta_{igk} \overset{ind}{\sim} \text{Bernoulli}(\theta_{igk}),\\
&\text{Level 2:} \quad \text{logit}(\theta_{igk}) = x'_{g(1)}\alpha + x'_{i(2)}\beta + x'_{igk(3)}\xi_c + \lambda h(w_{igk}) + v_i + u_{ig},\\
&\text{Level 3:} \quad v_i \overset{iid}{\sim} N(0, \sigma_v^2),\ u_{ig} \overset{iid}{\sim} N(0, \sigma_u^2).
\end{aligned} \tag{3.4}
$$

As already discussed in section 1.5, to fit the models in Table 3.1, we used the probabilistic programming language **Stan** [78] with the Statistical software **R** [79]. The **R**-package `rstan` [80] provides the **R** interface to **Stan**. For the intercept and the slopes, we have used $N(0, 5^2)$ prior. For σ_v and σ_u, we have used the Cauchy$^+(0, 5)$ prior. Note that, such priors are often called 'weakly informative priors', or WIP in literature [34]. These prior choices made sure that the posterior density would be proper. We ran each model for $4,000$ iterations, and the first $2,000$ were discarded as burn-in samples. Thus, all the posterior analyses were based on $2,000$ post-warm-up MCMC draws.

3.4.2 Model selection

After fitting all four models, we have used a Model Selection criterion to select one 'best' working model. As already discussed in 1.5.1, the **R**-package `loo` was used for this purpose; see [36]. The following Table 3.3 provides the LOO-CV comparison results for the 4 models.

Table 3.3: loo-cv comparison for the 4 models. Model 3 is selected as the best model.

	elpd_diff	se_diff
M3	0	0
M2	-0.958	0.804
M1	-1.672	0.836
M4	-64.216	18.477

From Table 3.3, we see that the LOO criterion selects Model 3 as the best model, and it is better than Model 2 as seen by the `elpd_diff`, which is significant at $\sim$70%. Note that the survey weights are entered linearly in model M3. Following [75], we wanted to see how M3 would perform when the survey weights are entered in the model non-linearly, keeping everything

else unchanged, viz., $\lambda \ln(w_{igk})$ and λw_{igk}^{-1} instead of λw_{igk}. Naturally, to check the performance of these two new competing models, we again employed the LOO-criterion. From Table 3.4, we observe that neither log-transformed survey weights nor the inverse of the survey weights brings about any improvement to the already existing M3. So, we keep M3 as our best working model and follow up with subsequent analyses with Model 3 (M3).

Table 3.4: loo-cv comparison of Model 3 varying weight functions

	elpd_diff	se_diff
M3	0	0
M3(ln(w_{igk}))	-0.329	1.874
M3(w_{igk}^{-1})	-3.644	2.576

Looking at the posterior analyses from Model 3, outlined in Table 3.6, we see that the variable co-morbidity from the Facebook survey seems to be insignificant at 70%. That is why we decided to run another model comparison between our chosen Model 3 and Model 3 without that particular variable. LOO analysis described in Table 3.5 suggests that removing the co-morbidity from the model does not bring about any significant difference. So we decided to leave that variable in our model.

Table 3.5: loo-cv comparison of Model 3 and Model 3 without the co-morbidity rate from Facebook-CMU-Delphi COVID-19 Trends and Impact Survey

	elpd_diff	se_diff
M3	0	0
M3 (without co-morb from FB)	-0.740	0.747

3.4.3 Findings from the selected model

In the last subsection, we decided on one model that best fits our data with the help of LOO criterion using **Stan**, `rstan` and `loo`. Now we present important findings from the selected 'best' working model, Model 3.

Table 3.6: Posterior summary statistics for the model parameters of Model 3–the best Model chosen by LOO criterion

	mean	se_mean	sd	10%	15%	85%	90%	n_{eff}	$\hat{R}$
α_1	-2.457	0.063	0.888	-3.608	-3.361	-1.566	-1.349	199.651	1.002
α_2	-1.232	0.061	0.870	-2.345	-2.132	-0.357	-0.141	201.311	1.003
α_3	-0.916	0.064	0.878	-2.078	-1.843	-0.034	0.159	187.865	1.001
α_4	-1.422	0.060	0.873	-2.540	-2.347	-0.545	-0.316	210.060	1.003
α_5	-1.391	0.063	0.885	-2.522	-2.326	-0.500	-0.294	198.465	1.002
α_6	-1.997	0.060	0.896	-3.157	-2.932	-1.096	-0.905	221.261	1.002
α_7	-1.121	0.062	0.979	-2.385	-2.140	-0.113	0.116	245.541	1.002
α_8	-1.574	0.066	1.153	-3.044	-2.762	-0.368	-0.154	302.163	1.002
β_1(co-morb)	-0.573	0.048	0.682	-1.439	-1.273	0.109	0.280	202.383	1.002
β_2(survey wts)	-0.112	0.001	0.040	-0.161	-0.153	-0.070	-0.060	$1,040.340$	1.000
ξ_{male}	0.209	0.001	0.019	0.184	0.189	0.229	0.233	$1,122.518$	1.001
ξ_{female}	0.118	0.001	0.020	0.092	0.096	0.139	0.144	952.271	1.001
λ(rep %)	-0.833	0.029	0.575	-1.571	-1.427	-0.248	-0.103	385.690	1.001
σ_v	0.121	0.013	0.072	0.026	0.035	0.196	0.215	31.222	1.039
σ_u	0.162	0.009	0.048	0.108	0.116	0.212	0.228	29.895	1.059

Table 3.6 gives the posterior summary statistics for the selected best model, Model 3. The 'mean' column gives the posterior means, 'se_mean' column gives the Monte-Carlo standard errors, the 'sd' column gives the posterior standard deviations (standard errors), the next four columns give the four percentiles, the column n_{eff} gives the Effective Sample Sizes (ESS, n_{eff}) – the number of independent samples that was used to replace the total n dependent MCMC draws having the same estimation power as the n autocorrelated samples, and finally the column $\hat{R}$ gives a measure of chain-equilibrium – a value near 1 (but < 1.1) means that the chains have

mixed well, and the posterior samples can be used with confidence for the posterior analyses ([33], [34]). As a default rule, [34] recommend running the simulation until n_{eff} is at least $5m$, where m is twice the number of independent chains/sequences. For our calculations, we ran 2 independent chains, each for $2,000$ iterations. This makes $5m = 5 \times 2 \times 2 = 20$ for our case. As we can see from table 3.6, all the $n_{\text{eff}} > 20$, and $\hat{R} < 1.1$. Thus, we can use these posterior analyses with fair confidence that the chains attained stationarity and as well as achieved proper mixing. We use the posterior estimates to calculate our state-wise parameter of interest $\bar{Y}_i$ for all the 50 states and DC – this essentially gives us the state-wise vaccine hesitancy estimates. These numbers, along with their corresponding standard errors, are displayed in Table 3.7. For easy comparison, the direct survey based estimates are also given in the same table.

Next, we compare our HB model-based state-wise estimates with the UAS-survey-weighted state-wise estimates. We plot the point estimates in the same graph. From the Figure 3.1 we observe that even though state-level UAS estimates for which enough sample are available are reliable, some of them, for which little to no samples were available in the UAS data, are unreasonably low/high due to the low sample sizes. Compared to that, our HB method produced stable estimates even for states with small or no samples. The corresponding 95% credible intervals of the HB estimates are also given in the same plot.

From figure 3.2, we can see how much improvement our HB estimates bring in over the direct UAS-survey-based estimates in terms of standard errors of the hesitancy estimates. The largest improvements came for the states that had fewer samples in the UAS survey data. We can still observe smaller standard error estimates even for the states with larger sample sizes.

Clearly, in Table 3.7, we can see improvements over the regular survey-weighted estimates. Delaware does not have an unrealistic estimate of 1 (with an absurd s.e. of 0) anymore and

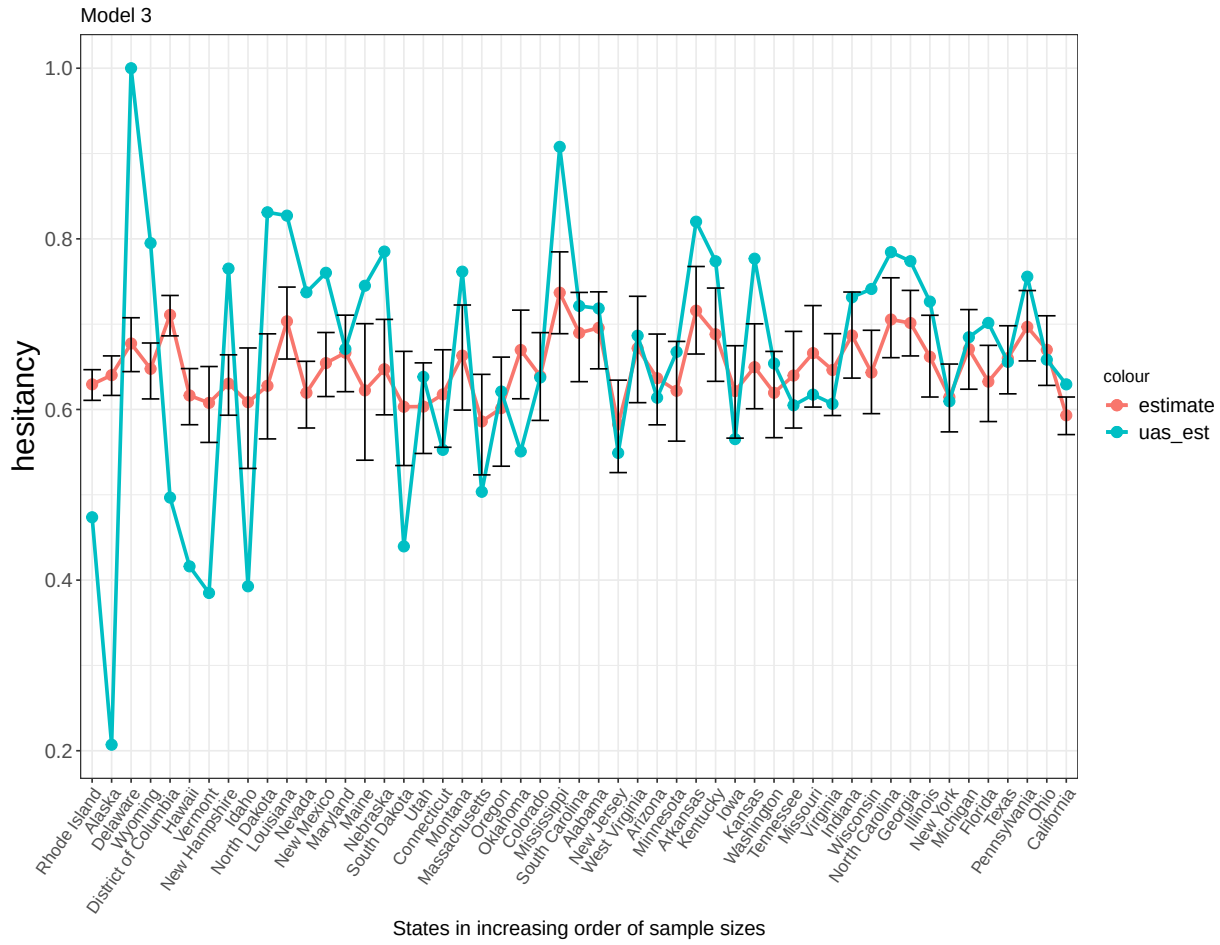


Figure 3.1: In x-axis, we have states in increasing order of UAS sample sizes. In y-axis, we have the point estimates of proportion of people who are "hesitant" to get vaccinated from UAS survey (in blue) and estimates from our HB Model – Model 3 (in red). The vertical lines around every HB estimates are their corresponding 95% Credible Intervals.

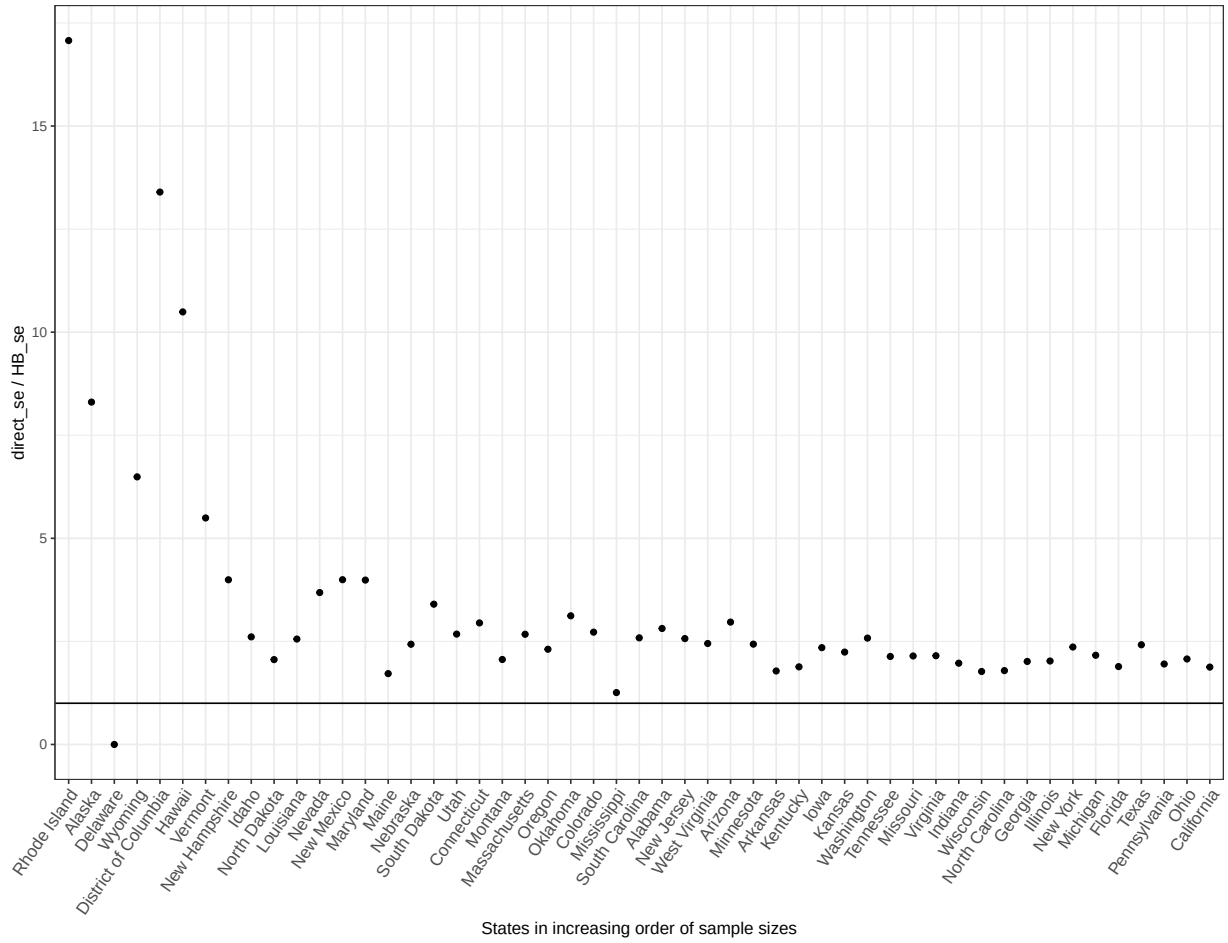


Figure 3.2: In x-axis, we have states in increasing order of UAS sample sizes. In y-axis, we have ratios of standard errors of direct UAS-survey based estimates to our HB model estimates. The horizontal dark line intersects the y-axis at 1.

Table 3.7: HB hesitancy estimates and direct UAS design-based estimates along with their standard errors for all the 50 states and Washington, D.C.

	State	HB_hesitancy	HB_se	direct	dir_se		State	HB_hesitancy	HB_se	direct	dir_se
1	Alabama	0.721	0.021	0.718	0.060	26	Missouri	0.686	0.027	0.617	0.057
2	Alaska	0.611	0.022	0.207	0.187	27	Montana	0.689	0.031	0.762	0.063
3	Arizona	0.668	0.024	0.614	0.072	28	Nebraska	0.643	0.032	0.785	0.078
4	Arkansas	0.752	0.026	0.820	0.046	29	Nevada	0.615	0.023	0.737	0.085
5	California	0.601	0.009	0.629	0.016	30	New Hampshire	0.576	0.033	0.765	0.131
6	Colorado	0.631	0.027	0.638	0.074	31	New Jersey	0.583	0.026	0.549	0.067
7	Connecticut	0.566	0.030	0.553	0.088	32	New Mexico	0.669	0.024	0.760	0.096
8	Delaware	0.751	0.026	1	0	33	New York	0.628	0.018	0.610	0.042
9	District of Columbia	0.701	0.016	0.497	0.219	34	North Carolina	0.719	0.021	0.784	0.037
10	Florida	0.662	0.019	0.701	0.035	35	North Dakota	0.669	0.042	0.831	0.086
11	Georgia	0.735	0.019	0.774	0.038	36	Ohio	0.689	0.017	0.658	0.036
12	Hawaii	0.573	0.019	0.416	0.203	37	Oklahoma	0.626	0.028	0.551	0.089
13	Idaho	0.506	0.045	0.393	0.118	38	Oregon	0.551	0.034	0.621	0.079
14	Illinois	0.674	0.022	0.726	0.045	39	Pennsylvania	0.733	0.015	0.756	0.030
15	Indiana	0.687	0.024	0.731	0.048	40	Rhode Island	0.594	0.019	0.474	0.321
16	Iowa	0.602	0.030	0.565	0.071	41	South Carolina	0.701	0.023	0.721	0.058
17	Kansas	0.676	0.022	0.777	0.050	42	South Dakota	0.543	0.036	0.440	0.122
18	Kentucky	0.680	0.026	0.774	0.049	43	Tennessee	0.622	0.026	0.605	0.055
19	Louisiana	0.721	0.028	0.827	0.071	44	Texas	0.661	0.015	0.656	0.037
20	Maine	0.591	0.044	0.745	0.076	45	Utah	0.630	0.034	0.638	0.090
21	Maryland	0.670	0.024	0.671	0.096	46	Vermont	0.558	0.037	0.385	0.204
22	Massachusetts	0.531	0.029	0.504	0.077	47	Virginia	0.602	0.022	0.607	0.048
23	Michigan	0.701	0.020	0.685	0.043	48	Washington	0.613	0.021	0.654	0.053
24	Minnesota	0.674	0.027	0.667	0.065	49	West Virginia	0.671	0.026	0.686	0.064
25	Mississippi	0.769	0.024	0.908	0.031	50	Wisconsin	0.644	0.023	0.741	0.041
						51	Wyoming	0.659	0.028	0.795	0.184

Alaska, DC, etc. – they all have much more believable estimates and smaller values of standard errors.

Table 3.8 gives a subset of table 3.7, detailing a few important states for which the differences between our HB and direct UAS-survey-based estimates are most prominent. We can see how having moderate to large sample sizes makes the direct estimates and our HB model-based estimates close to one another. As opposed to the states having smaller sample sizes, where the use of such modeling really makes a lot of sense in terms of having stable estimates and increased accuracy in their estimation. We also provide summary statistics of Direct UAS estimates and our HB estimates in Table 3.9. The summary statistics for standard errors of the Direct UAS estimates and our HB estimates are given in Table 3.10. From these latter two tables, we observe that our HB method improves over the direct method both in terms of point estimates and associated

Table 3.8: HB hesitancy estimates and direct UAS design-based estimates along with their standard errors for a subset of the states (small areas).

	State	HB_hes	HB_se	direct	direct_se	sample size
1	Alabama	0.721	0.021	0.718	0.060	73
5	California	0.601	0.009	0.629	0.016	1,879
8	Delaware	0.751	0.026	1	0	5
9	District of Columbia	0.701	0.016	0.497	0.219	7
10	Florida	0.662	0.019	0.701	0.035	241
11	Georgia	0.735	0.019	0.774	0.038	164
14	Illinois	0.674	0.022	0.726	0.045	176

Table 3.9: Summary statistics of direct UAS-survey based estimates of hesitancy and our HB based estimates of hesitancy.

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
direct	0.207	0.606	0.671	0.661	0.761	1
HB	0.506	0.602	0.661	0.648	0.688	0.769

standard errors.

3.5 Conclusion

In conclusion, this research introduces a novel hierarchical Bayes estimation procedure for finite population means in small areas, employing a sophisticated data linkage technique that integrates a primary (probability) survey with diverse data sources, including non-probability survey

Table 3.10: Summary statistics of *standard errors* of direct UAS-survey based estimates of hesitancy and standard errors of our HB based estimates of hesitancy.

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
direct	0	0.046	0.065	0.081	0.089	0.321
HB	0.009	0.021	0.024	0.026	0.029	0.045

data from social media. In our elaborate model, assumed only for the sample respondents, we include weights from the primary survey and factors potentially influencing the outcome variable of interest in order to reduce the informativeness of the sample. If the applied data analysis suggests that the weights are significant in the model, then this will imply that the sampling is informative. One should no longer assume the validity of the same model for the respondents and the population. Unlike many existing works, we acknowledge and address this critical step, assuming a simple model for the finite population regardless of the informativeness of sampling. The resulting hierarchical synthetic estimate demonstrates reduced sensitivity to model misspecifications, diminishing reliance on assumed models. Drawing inspiration from seminal works, our methodology is versatile and applicable to various finite population mean estimation challenges. However, it should be noted that our approach may fail to produce small area estimates in specific scenarios, requiring the validation of a crucial condition before proceeding with the data analysis. Using the same notations as we did in section 3.2, for a probability survey data, it is essential that the condition $(1 - a_{1i} - a_{2i}) \approx 0$ is met. This condition needs to be checked before proceeding with the estimation procedure. Finally, we illustrate the application of our methodology in generating proportion estimates (small area means of a binary outcome variable) from a COVID-19 survey, statistically linking information from diverse data sources. Future extensions of this work may explore similar estimation procedures for jointly modeling categorical and continuous response variables.

Chapter 4: Hierarchical Bayes Estimation of Small Area Proportions for Multinomial Response Variables

4.1 Introduction

Analyzing survey responses with more than two categories holds significant practical relevance in various scenarios. In many survey sampling contexts, especially those concerning socio-economic or health-related inquiries, respondents often provide nuanced responses that binary choices cannot adequately capture. For instance, when assessing public opinion on government policies, respondents may express varying degrees of agreement or disagreement rather than simply indicating a "yes" or "no" stance. Similarly, in health surveys, individuals may report their overall well-being on a scale ranging from "poor" to "excellent," encompassing multiple categories.

In the small area estimation (SAE) literature, which aims to provide reliable estimates for characteristics of small geographic areas, responses with more than two categories are common. However, due to methodological constraints or simplicity considerations, researchers typically dichotomize these responses into binary outcomes. While this approach may seem pragmatic, it overlooks the nuanced information embedded within the original multi-category responses, leading to the loss of valuable insight.

A method capable of handling multinomial responses in survey sampling is therefore of paramount interest to statisticians. Particularly in the context of small area estimation, where the aim is to produce accurate estimates despite limited sample sizes and complex survey designs, the ability to preserve the richness of multi-category responses can yield more precise and informative results. By embracing the complexity of multinomial responses, statisticians can better capture the true nature of respondents' perceptions, opinions, or behaviors, enhancing the quality and depth of survey-based analyses. Moreover, such an approach enables researchers to explore subtle variations in responses across different geographic areas, thereby facilitating more nuanced policymaking and targeted interventions. Overall, incorporating multinomial response analysis into survey sampling methodologies represents a promising avenue for advancing the field of small area estimation and enriching the understanding of complex social phenomena.

In this section, we aim to build upon the groundwork laid in the previous chapter 3 by extending our methodology to accommodate survey response variables with more than two categories, eliminating the need for dichotomization. The general model, defined for a broad family of NEF-QVF as delineated in equation 3.2, remains applicable in this context. Please refer to section 3.2 for a refresher on the topic. Previously, we showcased the practical utility of our model by addressing a binary outcome variable in the context of small area mean (proportion) estimation, as illustrated in equation 3.4. Now, we delve into the intricacies of adapting this model to effectively handle response variables with multiple categories, thereby preserving the richness of the original data without resorting to simplifying dichotomization methods.

The remainder of the chapter is structured as follows: In Section 4.2, we address the challenges arising from assuming a single model for both the population and the sample in the context of biased or informative sampling. Section 4.3 introduces the methodology tailored for extend-

ing the approach developed in Chapter 3 to handle multinomial response variables. Subsequently, Section 4.4 presents empirical findings obtained from applying our methodology to the datasets under consideration. Furthermore, we conduct a comparative analysis with Multilevel Regression with Poststratification (MRP) and highlight the biases introduced by MRP under informative sampling conditions. Finally, Section 4.5 provides concluding remarks for the chapter.

4.2 Unit level models and informative sampling

Assume unit-specific auxiliary data: $\underset{\sim}{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$, is available for the entire population, for element j within the i^{th} small area, $i = 1, \ldots, m,\ j = 1, \ldots, N_i$. Oftentimes, we have the area level summaries of these p covariates, $\bar{X}_i$. In most scenarios, this may be enough to formulate a working model. Assume the following model, which relates the variable of interest y_{ij} with the $\underset{\sim}{x}_{ij}$ through a nested error linear regression structure:

$$y_{ij} = \underset{\sim}{x}'_{ij}\underset{\sim}{\beta} + v_i + e_{ij},\ j = 1, \ldots, N_i,\ i = 1, \ldots, m. \tag{4.1}$$

Usual parameters of interests are: small-area population means $\bar{Y}_i$, or the small-area population total Y_i. Now assume that a sample s_i, of size n_i,, was drawn from the small-area population of the size of N_i units. In model-based small area estimation, it is often assumed that the population model given in equation 4.1, also holds for the sampled units. This is, of course, easily satisfied by the simple random sampling (SRS) within each small area [4]. For this particular small area i, we can rewrite equation 4.1 as:

$$\underset{\sim}{y}_i^P = \mathbf{X}_i^P\underset{\sim}{\beta} + v_i\mathbf{1}_i^P + \underset{\sim}{e}_i^P,\ i = 1, \ldots, m. \tag{4.2}$$

where the superscript P denotes all the units of the population. $\mathbf{1}_i^P$ is the indicator variable taking the value of 1, if the j^{th} unit is included in the sample, and 0 otherwise. We can partition equation 4.2 into sampled population units who responded, sampled population units who did not respond, and finally, the non-sampled population units as follows:

$$\underset{\sim}{y}_i^P = \begin{bmatrix} \underset{\sim}{y}_i^r \\ \underset{\sim}{y}_i^{nr} \\ \underset{\sim}{y}_i^{ns} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i^r \\ \mathbf{X}_i^{nr} \\ \mathbf{X}_i^{ns} \end{bmatrix} \underset{\sim}{\beta} + v_i \begin{bmatrix} \mathbf{1}_i^r \\ \mathbf{1}_i^{nr} \\ \mathbf{1}_i^{ns} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i^r \\ \mathbf{e}_i^{nr} \\ \mathbf{e}_i^{ns} \end{bmatrix}, \tag{4.3}$$

where the superscript r stands for sampled respondents, nr stands for sampled nonrespondents, and ns stands for non-sampled population units. Now, under the absence of selection bias, or under non-informative sampling, we can say that the population model also holds for the sample respondents. Subsequently, the inference on model parameters $\underset{\sim}{\lambda} = (\underset{\sim}{\beta}', \sigma_v^2, \sigma_e^2)$ can be based on:

$$f(\boldsymbol{y}_i^r | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) = \int \int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns}. \tag{4.4}$$

Now, let us introduce the following two indicator variables:

1. $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iN_i})'$ is the vector of indicator variables, defined for all the population units, where $a_{ij} = 1$, if the j^{th} population unit of the i^{th} area was included in the sample, and 0 otherwise.

2. In the same spirit, define another vector of indicator variables $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{in_i})'$, defined only for the sampled units, where $u_{ij} = 1$, if the j^{th} sampled unit responded to the survey questionnaire, and 0 otherwise.

Now, the joint distribution of the sample data $(\boldsymbol{y}_i^r, \boldsymbol{a}_i, \boldsymbol{u}_i)$ can be written down as:

$$
\begin{aligned}
& f(\boldsymbol{y}_i^r, \boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) \\
&= \int\int f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns}, \boldsymbol{X}_i^P) f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \\
&= \int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns}, \boldsymbol{X}_i^P) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \qquad (4.5) \\
&= \left[\int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \right] f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{X}_i^P) \qquad (4.6)
\end{aligned}
$$

assuming:

$$
f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns}, \boldsymbol{X}_i^P) = f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{X}_i^P). \qquad (4.7)
$$

That is, the joint distribution of $(\boldsymbol{a_i}, \boldsymbol{u_i}) | \boldsymbol{X}_i^P$ is independent of $\boldsymbol{y}_i^P$, but may depend on $\boldsymbol{X}_i^P$. If this is the case, then we can base our inferences for $\boldsymbol{\lambda}$ on $f(\boldsymbol{y}_i^r | \boldsymbol{X}_i^P, \boldsymbol{\lambda})$ given in equation 4.4. Alternatively, equation 4.5 can be further expanded into the following:

$$
\begin{aligned}
& f(\boldsymbol{y}_i^r, \boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) \\
&= \int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) f(\boldsymbol{a}_i, \boldsymbol{u}_i | \boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns}, \boldsymbol{X}_i^P) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \\
&= \int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) f(\boldsymbol{u}_i | \boldsymbol{a}_i, \boldsymbol{y}_i^P, \boldsymbol{X}_i^P) f(\boldsymbol{a}_i | \boldsymbol{y}_i^P, \boldsymbol{X}_i^P) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \qquad (4.8) \\
&= \left[\int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns} | \boldsymbol{X}_i^P, \boldsymbol{\lambda}) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns} \right] f(\boldsymbol{u}_i | \boldsymbol{a}_i, \boldsymbol{X}_i^P) f(\boldsymbol{a}_i | \boldsymbol{X}_i^P) \qquad (4.9)
\end{aligned}
$$

assuming:

$$f(\boldsymbol{u}_i|\boldsymbol{a}_i, \boldsymbol{y}_i^P, \boldsymbol{X}_i^P) = f(\boldsymbol{u}_i|\boldsymbol{a}_i, \boldsymbol{X}_i^P) \tag{4.10}$$

$$f(\boldsymbol{a}_i|\boldsymbol{y}_i^P, \boldsymbol{X}_i^P) = f(\boldsymbol{a}_i|\boldsymbol{X}_i^P) \tag{4.11}$$

That is, assuming given $\boldsymbol{X}_i^P$, the conditional distribution of the response indicator vector given the sample selection vector, is independent of $\boldsymbol{y}_i^P$, but may depend on $\boldsymbol{X}_i^P$. Furthermore, given $\boldsymbol{X}_i^P$, the sample selection vector is independent of $\boldsymbol{y}_i^P$, but may depend on $\boldsymbol{X}_i^P$. Then, we do not have a case of 'informative sampling', and we may proceed with the inferences for $\boldsymbol{\lambda}$ based on $f(\boldsymbol{y}_i^r|\boldsymbol{X}_i^P, \boldsymbol{\lambda})$ given in equation 4.4.

Now, if $(\boldsymbol{u}_i|\boldsymbol{a}_i)$, and $\boldsymbol{a}_i$ depend on auxiliary variable/s $\boldsymbol{Z}_i^P$, which is not included in $\boldsymbol{X}_i^P$, then the distribution of the sample data: $(\boldsymbol{y}_i^r, \boldsymbol{a}_i, \boldsymbol{u}_i)$ is given by:

$$\begin{aligned} &f(\boldsymbol{y}_i^r, \boldsymbol{a}_i, \boldsymbol{u}_i|\boldsymbol{X}_i^P, \boldsymbol{Z}_i^P, \boldsymbol{\lambda}) \\ &= \left[\int\int f(\boldsymbol{y}_i^r, \boldsymbol{y}_i^{nr}, \boldsymbol{y}_i^{ns}|\boldsymbol{X}_i^P, \boldsymbol{Z}_i^P, \boldsymbol{\lambda}) d\boldsymbol{y}_i^{nr} d\boldsymbol{y}_i^{ns}\right] f(\boldsymbol{u}_i|\boldsymbol{a}_i, \boldsymbol{X}_i^P, \boldsymbol{Z}_i^P) f(\boldsymbol{a}_i|\boldsymbol{X}_i^P, \boldsymbol{Z}_i^P) \end{aligned} \tag{4.12}$$

In such a case, inferences on $\boldsymbol{\lambda}$ must be based on $f(\boldsymbol{y}_i^r|\boldsymbol{X}_i^P, \boldsymbol{Z}_i^P, \boldsymbol{\lambda}) \neq f(\boldsymbol{y}_i^r|\boldsymbol{X}_i^P, \boldsymbol{\lambda})$, as was the case in equation 4.4.

4.3 Methodology

Similar to the notations outlined in Section 3.2, we begin by partitioning the finite population into G cells for each small area i $(i = 1, \ldots, m)$, leveraging factors available in the survey

data that potentially influence the outcome variable of interest. We assume access to known population sizes for these cells derived from census data. Let N_i denote the known population size of area i, and N_{ig} represent the population size of the g^{th} cell within the i^{th} area, where $i = 1, \ldots, m$ and $g = 1, \ldots, G$. Naturally, it follows that $\sum_g N_{ig} = N_i$ and $\sum_i N_i = N$, with N representing the total finite population size.

Consider y_{igk} as the response variable representing the k^{th} individual's response to a survey question with multiple categories within the g^{th} cell of the i^{th} small area ($i = 1, \ldots, m;\ g = 1, \ldots, G;\ k = 1, \ldots, N_{ig}$). We assume the availability of various types of known auxiliary variables to model our outcome variable. Let $x_{g(1)}$ denote a vector of indicator variables corresponding to cell g, $x_{i(2)}$ represent a vector of known auxiliary variables specific to area i, and $x_{igk(3)}$ signify a vector of individual-level auxiliary variables.

Let us further consider a sample of size n drawn from the finite population of size N. Define G_i as the number of cells represented in the sample for area i ($i = 1, \ldots, m$). Let n_i and n_{ig} represent the sample sizes for area i and cell g within area i, respectively. This means $\sum_{g=1}^{G_i} n_{ig} = n_i$ and $\sum_{i=1}^{m} n_i = n$. It is possible to have $n_i = 0$ for some area i, or $n_{ig} = 0$ for some cell g within an area i; $i = 1, \ldots, m$; $g = 1, \ldots, G$. Denote w_{igk} as the survey weight for the k^{th} sampled respondent belonging to the g^{th} cell residing in the i^{th} small area ($i = 1, \ldots, m;\ g = 1, \ldots, G_i;\ k = 1, \ldots, n_{ig}$).

In subsection 4.2, we explored the implications of informative or biased sampling on the assumed population model for the sampled respondents. To illustrate this, we introduced two indicator variables: $\{\boldsymbol{a}_i,\ i = 1, \ldots, m\}$ at the population-level, indicating whether a population unit was included in the sample for the i^{th} area, and $\{\boldsymbol{u}_i,\ i = 1, \ldots, m\}$, at the sample-level indicating whether a sampled unit responded to the survey. We demonstrated that in the case of

informative sampling, these variables are not independent of $\boldsymbol{y}_i^P$, the survey response variable for the i^{th} small area. To address this informativeness of the sampling design, we can introduce a set of auxiliary variable/s $\boldsymbol{Z}_i^P$ that were not initially included in the auxiliary variables $\boldsymbol{X}_i^P$, while devising a model for the population. Such $\boldsymbol{Z}_i^p$ is expected to contain information regarding $(\boldsymbol{a}_i, \boldsymbol{u}_i)$. However, obtaining such variables can be challenging due to the complexity of sample selection processes and respondent behavior. In such scenarios, survey weights become valuable as they adjust for complex sampling schemes and non-response, among other factors. By incorporating these weights into the model and testing the significance of associated regression coefficients with the sampled data, we can assess the informativeness of the sampling design. If these coefficients are significant, it indicates that the population model cannot be assumed to hold for the sampled respondents. However, it is important to note that survey weights are only available for the sampled data, requiring careful modifications to the model and corresponding inference for population parameters.

4.3.1 Parameter of interest

We are interested in estimating finite population proportions of different categories of responses for all areas. If we consider that the survey question had C categories, then the population parameters of interest are $\{\bar{Y}_i^{(c)},\ c = 1, \ldots, C\}$, where

$$\bar{Y}_i^{(c)} = N_i^{-1} \sum_{g=1}^{G} \sum_{k=1}^{N_{ig}} y_{igk} 1\{y_{igk} = c\} = \sum_{g=1}^{G} \frac{N_{ig}}{N_i} \bar{Y}_{ig}^{(c)}, \tag{4.13}$$

where $N_i = \sum_g N_{ig}$, population size for the i^{th} area ($i = 1, \cdots, m$), and $1\{y_{igk} = c\} = 1$, when k^{th} individual of the g^{th} cell of the i^{th} small area responds to the survey question with category

c, $c = 1, \ldots, C$, and 0 otherwise.

4.3.2 Model

Up to this point, we have introduced the notation for the different components of the model and have defined our parameters of interest within the finite population framework. With this foundation, we can now construct a detailed model specifically tailored to the sampled respondents. Focusing on the sample data alone allows us to conduct thorough model validation and selection procedures. Define: $\boldsymbol{\pi}_{igk} = (\pi_{igk}^{(1)}, \ldots, \pi_{igk}^{(C)})$ as the vector of probabilities, with $\{\pi_{igk}^{(c)} = P(y_{igk} = c) : \sum_{c=1}^{C} \pi_{igk}^{(c)} = 1\}$. Also, we designate category C to serve as the baseline or the pivotal category. For $i = 1, \cdots, m;\ g = 1, \cdots, G_i,\ k = 1, \cdots, n_{ig}$,

$$
\begin{aligned}
&\text{Level 1:} \quad y_{igk} | \boldsymbol{\pi}_{igk} \overset{ind}{\sim} \text{Multinomial}(1;\ \pi_{igk}^{(1)}, \ldots, \pi_{igk}^{(C)}), \\
&\text{Level 2:} \quad \log_e \left(\frac{\pi_{igk}^{(c)}}{\pi_{igk}^{(C)}} \right) = x'_{g(1)} \alpha^{(c)} + x'_{i(2)} \beta^{(c)} + x'_{igk(3)} \xi^{(c)} + \lambda^{(c)} h(w_{igk}) + v_i^{(c)} + u_{ig}^{(c)}, \\
&\text{Level 3:} \quad v_i^{(c)} \overset{iid}{\sim} N(0, \sigma_v^2), \text{ independently of, } u_{ig}^{(c)} \overset{iid}{\sim} N(0, \sigma_u^2),\ \forall\, c = 1, \ldots, C-1. \qquad (4.14)
\end{aligned}
$$

This approach for specifying level 2, along with setting the model parameters for the baseline category to all 0, help circumvent the challenge of model identifiability. It is worth noting that the distribution specified at level 1 is also known as the categorical distribution, given that the multinomial distribution has only one trial. The probabilities of group membership, denoted as $\pi_{igk}^{(c)}$, can be explicitly solved for as follows:

$$\pi_{igk}^{(c)} = \frac{\exp\left(x'_{g(1)}\alpha^{(c)} + x'_{i(2)}\beta^{(c)} + x'_{igk(3)}\xi^{(c)} + \lambda^{(c)}h(w_{igk}) + v_i^{(c)} + u_{ig}^{(c)}\right)}{1 + \sum_{c=1}^{C-1} \exp\left(x'_{g(1)}\alpha^{(c)} + x'_{i(2)}\beta^{(c)} + x'_{igk(3)}\xi^{(c)} + \lambda^{(c)}h(w_{igk}) + v_i^{(c)} + u_{ig}^{(c)}\right)},$$

for each category $c = 1, \ldots, C$. Typically, model parameters are estimated using maximum likelihood methods. However, in our approach, we will assign prior distributions to these parameters and estimate them from the posterior distribution, in line with hierarchical Bayesian methodology.

4.3.3 Inference

We proceed by defining the following three partitions of the group index set $\{1, \ldots, G\}$, for small area i in the following manner. Note that these are defined exactly as in the previous chapter, in section 3.2.

$\mathcal{G}_{1i} = \{g \ni n_{ig} > 0,\ g = 1, \cdots, G\}$: the set of cells represented in the sample for the area i.

$\mathcal{G}_{2i} = \{g \ni n_{ig} = 0, n_{jg} > 0, \text{for some } j \neq i, g = 1, \cdots, G\}$: the set of cells not represented in the sample for the area i, but represented by one or more other areas $j \neq i$. And finally, $\mathcal{G}_{3i} = \{g \ni n_{ig} = 0, \forall i,\ g = 1, \cdots, G\}$: the set of cells not represented in the sample for area i. The population proportion (or probability) for the response category c for the set of cells $\mathcal{G}_{1i}$ is defined as:

$$\bar{\Pi}_{1i}^{(c)} = \frac{\sum_{g\in\mathcal{G}_{1i}} \sum_{k=1}^{N_{ig}} \pi_{igk}^{(c)}}{\sum_{g\in\mathcal{G}_{1i}} N_{ig}} = \sum_{g\in\mathcal{G}_{1i}} b_{ig;1}\bar{\Pi}_{ig}^{(c)}, \text{ where } b_{ig;1} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{1i}} N_{ig}} \text{ and } \bar{\Pi}_{ig}^{(c)} = \frac{\sum_{k=1}^{N_{ig}} \pi_{igk}^{(c)}}{N_{ig}}.$$

Similarly, we define:

$$\bar{\Pi}_{2i}^{(c)} = \frac{\sum_{g\in\mathcal{G}_{2i}}\sum_{k=1}^{N_{ig}}\pi_{igk}^{(c)}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\bar{\Pi}_{ig}^{(c)}, \text{ where } b_{ig;2} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{2i}}N_{ig}} \text{ and } \bar{\Pi}_{ig}^{(c)} = \frac{\sum_{k=1}^{N_{ig}}\pi_{igk}^{(c)}}{N_{ig}},$$

and finally,

$$\bar{\Pi}_{3i}^{(c)} = \frac{\sum_{g\in\mathcal{G}_{3i}}\sum_{k=1}^{N_{ig}}\pi_{igk}^{(c)}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}} = \sum_{g\in\mathcal{G}_{3i}} b_{ig;3}\bar{\Pi}_{ig}^{(c)}, \text{ where } b_{ig;3} = \frac{N_{ig}}{\sum_{g\in\mathcal{G}_{3i}}N_{ig}} \text{ and } \bar{\Pi}_{ig}^{(c)} = \frac{\sum_{k=1}^{N_{ig}}\pi_{igk}^{(c)}}{N_{ig}}.$$

as the population means for the set of cells $\mathcal{G}_{2i}$ and the set of cells $\mathcal{G}_{3i}$, respectively. Here, we make a synthetic assumption that:

$$\bar{\Pi}_{2i}^{(c)} = \sum_{g\in\mathcal{G}_{2i}} b_{ig;2}\bar{\Pi}_{ig;\text{syn}}^{(c)} \text{ where, } \bar{\Pi}_{ig;\text{syn}}^{(c)} = \frac{\sum_{j\neq i}\sum_{k}\pi_{jgk}^{(c)}}{\sum_{j\neq i}N_{jg}}$$

i.e, $\bar{\Pi}_{2i}^{(c)}$ is identical to the overall mean of the cells in $\mathcal{G}_{2i}$ obtainable from one or more areas (other than area i). We make no assumption on the rest of the finite population units. Since population sizes N_{ig} are in general large, appealing to the law of large numbers as in [28], we can approximate the finite population means $\bar{Y}_i^{(c)}$ in equation 4.13, with the following functions of parameters of the assumed finite population model.

$$\begin{aligned}\bar{Y}_i^{(c)} \approx \bar{\Pi}_i^{(c)} = \sum_{g=1}^{G}\frac{N_{ig}}{N_i}\bar{\Pi}_{ig}^{(c)} &= a_{1i}\bar{\Pi}_{1i}^{(c)} + a_{2i}\bar{\Pi}_{2i}^{(c)} + (1-a_{1i}-a_{2i})\bar{\Pi}_{3i}^{(c)}, \text{(say)} \\ &\approx a_{1i}\bar{\Pi}_{1i}^{(c)} + a_{2i}\bar{\Pi}_{2i}^{(c)}, \end{aligned} \tag{4.15}$$

where $a_{1i} = N_i^{-1} \sum_{g \in \mathcal{G}_{1i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{1i}$; $a_{2i} = N_i^{-1} \sum_{g \in \mathcal{G}_{2i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{2i}$; $a_{3i} = N_i^{-1} \sum_{g \in \mathcal{G}_{3i}} N_{ig}$, the population proportion of units belonging to the set of cells $\mathcal{G}_{3i}$. The approximation in equation 4.15 will only be valid when $(1 - a_{1i} - a_{2i}) \approx 0$.

Most implementations of HB methodology estimate model parameters by sampling from posterior distributions using the Monte Carlo Markov Chain (MCMC) method. We will assume the same weakly informative priors on the model-hyperparameters as we did in the last chapter 3. At each MCMC iteration step r $(r = 1, \ldots, R)$, and for each category of responses c $(c = 1, \ldots, C)$, the following quantity is generated for the area i:

$$\bar{\pi}_{i(r)}^{(c)} = a_{1i}\bar{\pi}_{1i(r)}^{(c)} + a_{2i}\bar{\pi}_{2i(r)}^{(c)} + (1 - a_{1i} - a_{2i})\bar{\pi}_{3i(r)}^{(c)} \approx a_{1i}\bar{\pi}_{1i(r)}^{(c)} + a_{2i}\bar{\pi}_{2i(r)}^{(c)}, r = 1, \ldots, R,\ c = 1, \ldots, C,$$

where

$$\begin{aligned}
\bar{\pi}_{1i(r)}^{(c)} &= \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\pi}_{ig(r)}^{(c)}, \text{ with } \bar{\pi}_{ig(r)}^{(c)} = \frac{\sum_k w_{igk}\pi_{igk}^{(r)}}{\sum_k w_{igk}}, \\
\bar{\pi}_{2i(r)}^{(c)} &= \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\pi}_{ig;\text{syn}(r)}^{(c)}, \text{ with } \bar{\pi}_{ig;\text{syn}(r)}^{(c)} = \frac{\sum_{j \neq i} \sum_k w_{jgk}\pi_{jgk(r)}^{(c)}}{\sum_{j \neq i} \sum_k w_{jgk}}, \\
\pi_{igk(r)}^{(c)} &= \frac{\exp\left(x'_{g(1)}\alpha_{(r)}^{(c)} + x'_{i(2)}\beta_{(r)}^{(c)} + x'_{igk(3)}\xi_{(r)}^{(c)} + \lambda_{(r)}^{(c)}h(w_{igk}) + v_{i(r)}^{(c)} + u_{ig(r)}^{(c)}\right)}{\sum_{c=1}^{C} \exp\left(x'_{g(1)}\alpha_{(r)}^{(c)} + x'_{i(2)}\beta_{(r)}^{(c)} + x'_{igk(3)}\xi_{(r)}^{(c)} + \lambda_{(r)}^{(c)}h(w_{igk}) + v_{i(r)}^{(c)} + u_{ig(r)}^{(c)}\right)}.
\end{aligned}$$

To justify, note that

$$\bar{\Pi}_{1i}^{(c)} = \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\Pi}_{ig}^{(c)} \approx \sum_{g \in \mathcal{G}_{1i}} b_{ig;1}\bar{\pi}_{ig}^{(c)} = \bar{\pi}_{1i}^{(c)}, \text{ say},$$

where $\bar{\pi}_{ig}^{(c)} = \frac{\sum_k w_{igk}\pi_{igk}^{(c)}}{\sum_k w_{igk}}$; see [28] and [77] for similar justifications in empirical best prediction and hierarchical Bayes approaches, respectively. Also,

$$\bar{\Pi}_{2i}^{(c)} = \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\Pi}_{ig;\text{syn}}^{(c)} \approx \sum_{g \in \mathcal{G}_{2i}} b_{ig;2}\bar{\pi}_{ig;\text{syn}}^{(c)} = \bar{\pi}_{2i}^{(c)}, \text{ say,}$$

where $\bar{\pi}_{ig;\text{syn}}^{(c)} = \frac{\sum_{j \neq i}\sum_k w_{jgk}\pi_{jgk}^{(c)}}{\sum_{j \neq i}\sum_k w_{jgk}}$. The set of R MCMC replicates, drawn from the posterior distribution given the sample data for the i^{th} area, i.e., $\{\bar{\pi}_{i(r)}^{(c)}, r = 1, ..., R\}$, will be used for inferences about $\{\bar{\pi}_i^{(c)}, c = 1, \ldots, C\}$.

4.4 Data Analysis

In the preceding subsection, we detailed the essential components for model fitting and conducting posterior inferences. Here, we present the results derived from applying the chosen model to the data. Notably, as we have previously described the datasets in the preceding chapter, we will refrain from duplicating that information. Interested readers can refer to subsection 3.3 for a comprehensive overview of the primary survey data, including various auxiliary data sources utilized.

Moreover, in the preceding chapter 3, we conducted a thorough comparison of numerous models during the model selection phase before finalizing the model for generating small area estimates. The model employed in this subsection 3.4 closely resembles the final selected model from that process. Given the extensive model building and selection conducted with the same dataset in the previous chapter, we have omitted those steps here. Instead, we focus on demonstrating the production of small area estimates for response variables with multiple categories in

the presence of informative sampling using a single model.

Table 4.1 presents the posterior estimates of various fixed effects terms utilized in the model. Notably, all $\hat{R}$ values are below 1.05, indicating confidence in utilizing the posterior samples [34]. Moreover, it is evident that the survey weights are significant at the 95% confidence level for multiple categories. This underscores the informative nature of the sampling process. Consequently, it is imperative not to disregard this aspect and proceed with a population model lacking survey weights. This validation reinforces our approach to estimating small area means, as outlined in subsection 4.3.3. Of course, this is given that the survey data must obey $(1 - a_{1i} - a_{2i}) \approx 0$, as is the case for the dataset of our consideration (see table 3.2).

In Table 4.2, we present the posterior estimates of the random effects terms in the model. Similar to the fixed effects table, all $\hat{R}$ values remain comfortably below 1.05. Moreover, both "Bulk_ESS" and "Tail_ESS" exhibit satisfactory levels in both tables, indicating high effective sample sizes. For further insights into these measures and their significance in assessing MCMC sample quality, please refer to subsection 1.5. Figures 4.1, 4.2, 4.3, 4.4, and 4.5 display the comparison of our model based method with the direct survey estimates. The 95% credible intervals of our HB model-based estimates are also given in the same plot. The small areas are arranged in increasing order of sample sizes in the X-axis. Also, the national estimate is displayed as a solid horizontal line in each of the plots for each category. We observe pretty consistent scenarios throughout all the different categories of response. Our model produces estimates with considerable precision, even for the small areas with 0 samples for some categories.

As discussed in subsection 1.6, Multilevel Regression and Post-stratification (MRP) is commonly employed for small area estimation. However, MRP assumes that the population and sample models are identical, which is problematic in cases of informative or biased sampling, as

Table 4.1: Summary of the fixed effects of the model parameters. 1-VUL: category 1–Very Un-Likely; 2-SUL: category 2–Somewhat Un-Likely; 3-SL: category 3–Somewhat Likely; 4-VL: category 4–Very Likely; and the baseline category 5–Unsure.

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
1-VUL_Intercept	-0.249	0.822	-1.848	1.312	1.002	1656.533	2443.167
2-SUL_Intercept	1.226	0.902	-0.561	2.954	1.002	1990.340	2529.666
3-SL_Intercept	0.832	0.790	-0.767	2.385	1.004	1850.438	2615.362
4-VL_Intercept	0.475	0.785	-1.052	2.022	1.006	1466.184	2515.302
1-VUL_raceBlack	0.409	0.310	-0.199	1.016	1.002	1474.273	1998.169
1-VUL_raceOthers	0.263	0.325	-0.380	0.901	1.002	1412.228	2017.522
1-VUL_raceWhite	0.412	0.290	-0.163	0.984	1.003	1419.950	1996.929
1-VUL_hisplatinoYes	-0.288	0.159	-0.600	0.029	1.001	2472.112	2637.633
1-VUL_genderMale	0.167	0.099	-0.021	0.363	1.004	2340.005	2563.420
1-VUL_age_cat2	0.634	0.264	0.119	1.145	1.011	704.351	1538.344
1-VUL_age_cat3	0.555	0.258	0.053	1.051	1.013	552.264	1448.026
1-VUL_age_cat4	0.388	0.264	-0.112	0.916	1.013	586.898	1556.836
1-VUL_age_cat5	0.287	0.262	-0.225	0.803	1.013	601.866	1323.219
1-VUL_age_cat6	0.406	0.279	-0.143	0.961	1.009	717.049	1827.403
1-VUL_age_cat7	0.379	0.326	-0.242	1.020	1.009	719.401	1550.590
1-VUL_C1	0.165	0.716	-1.268	1.521	1.001	2501.864	3101.276
1-VUL_weight	-0.048	0.055	-0.154	0.057	1.002	1835.605	2463.328
2-SUL_raceBlack	-0.545	0.308	-1.146	0.058	1.002	1620.510	2575.726
2-SUL_raceOthers	-0.400	0.318	-1.049	0.179	1.002	1552.967	2249.710
2-SUL_raceWhite	-0.424	0.273	-0.944	0.100	1.002	1450.599	2452.281
2-SUL_hisplatinoYes	-0.150	0.172	-0.487	0.194	1.001	2768.375	3175.447
2-SUL_genderMale	0.402	0.114	0.174	0.623	1.003	2522.710	2889.931
2-SUL_age_cat2	-0.073	0.266	-0.599	0.458	1.009	713.194	1475.523
2-SUL_age_cat3	-0.287	0.259	-0.797	0.227	1.009	712.064	1421.467
2-SUL_age_cat4	-0.543	0.265	-1.071	-0.033	1.010	725.163	1534.734
2-SUL_age_cat5	-0.797	0.266	-1.313	-0.281	1.008	729.491	1327.101
2-SUL_age_cat6	-0.589	0.284	-1.155	-0.036	1.008	779.805	1723.153
2-SUL_age_cat7	-0.702	0.355	-1.396	-0.009	1.009	1000.825	2097.173
2-SUL_C1	0.440	0.809	-1.132	2.004	1.000	2737.211	2995.057
2-SUL_weight	-0.310	0.067	-0.439	-0.180	1.000	2425.197	2711.690
3-SL_raceBlack	-0.808	0.258	-1.311	-0.315	1.001	1416.546	1843.100
3-SL_raceOthers	-0.775	0.276	-1.328	-0.238	1.001	1407.175	1789.440
3-SL_raceWhite	-0.460	0.231	-0.913	-0.009	1.004	1268.959	1951.808
3-SL_hisplatinoYes	-0.236	0.148	-0.525	0.062	1.001	2472.712	2904.142
3-SL_genderMale	0.406	0.095	0.221	0.589	1.004	2133.667	2417.740
3-SL_age_cat2	0.096	0.241	-0.396	0.562	1.016	665.715	1294.748
3-SL_age_cat3	0.217	0.237	-0.253	0.668	1.017	636.064	1375.058
3-SL_age_cat4	0.002	0.239	-0.477	0.459	1.016	623.238	1520.384
3-SL_age_cat5	-0.017	0.238	-0.496	0.448	1.018	614.908	1396.900
3-SL_age_cat6	0.403	0.250	-0.080	0.889	1.017	612.977	1347.733
3-SL_age_cat7	0.361	0.293	-0.220	0.935	1.015	690.493	1672.807
3-SL_C1	-0.170	0.709	-1.573	1.252	1.001	2565.630	2567.431
3-SL_weight	-0.247	0.055	-0.354	-0.138	1.001	2028.856	2559.987
4-VL_raceBlack	-1.724	0.278	-2.255	-1.189	1.002	1457.953	2378.965
4-VL_raceOthers	-0.938	0.280	-1.485	-0.415	1.003	1210.833	2339.810
4-VL_raceWhite	-0.494	0.242	-0.965	-0.032	1.003	1164.846	2296.319
4-VL_hisplatinoYes	-0.340	0.159	-0.645	-0.015	1.002	1818.375	2584.888
4-VL_genderMale	0.691	0.088	0.518	0.864	1.002	2134.258	3089.841
4-VL_age_cat2	0.179	0.233	-0.270	0.627	1.013	697.354	1646.674
4-VL_age_cat3	0.156	0.228	-0.301	0.590	1.016	663.182	1318.004
4-VL_age_cat4	0.037	0.229	-0.410	0.482	1.015	654.054	1367.113
4-VL_age_cat5	0.188	0.228	-0.266	0.635	1.017	630.713	1329.275
4-VL_age_cat6	0.730	0.237	0.259	1.187	1.014	651.867	1328.418
4-VL_age_cat7	0.875	0.277	0.344	1.426	1.014	767.008	1679.868
4-VL_C1	-0.884	0.692	-2.261	0.443	1.002	2101.345	2867.892
4-VL_weight	-0.280	0.053	-0.385	-0.178	1.003	1746.503	2676.564

Table 4.2: Summary of the random effects of the model parameters.

State	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
$\sigma_v(1-VUL)$	0.089	0.058	0.005	0.214	1.003	1361.305	1865.940
$\sigma_v(2-SUL)$	0.130	0.078	0.007	0.297	1.001	1071.347	1400.320
$\sigma_v(3-SL)$	0.104	0.063	0.006	0.235	1.006	1018.157	1121.844
$\sigma_v(4-VL)$	0.121	0.071	0.007	0.264	1.006	689.759	1351.671
State:Race:Hisp							
$\sigma_u(1-VUL)$	0.121	0.076	0.007	0.285	1.001	866.657	1540.878
$\sigma_u(2-SUL)$	0.072	0.054	0.003	0.201	1.000	1959.455	2021.904
$\sigma_u(3-SL)$	0.069	0.051	0.002	0.189	1.002	1318.036	2235.666
$\sigma_u(4-VL)$	0.142	0.074	0.011	0.286	1.011	491.892	1360.607

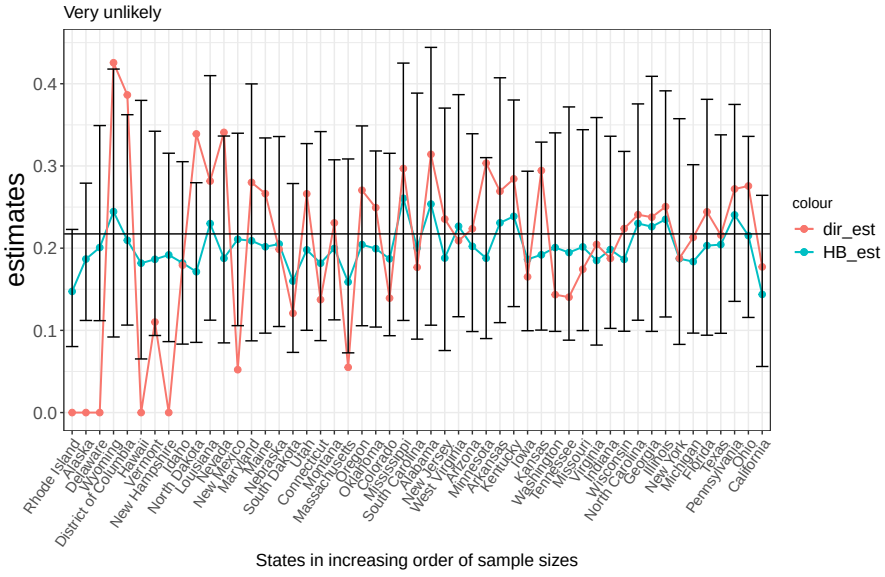


Figure 4.1: Comparison of direct estimates with our proposed HB model-based estimates for the response category 1: 'Very unlikely'. The national estimate is represented by the solid horizontal line.

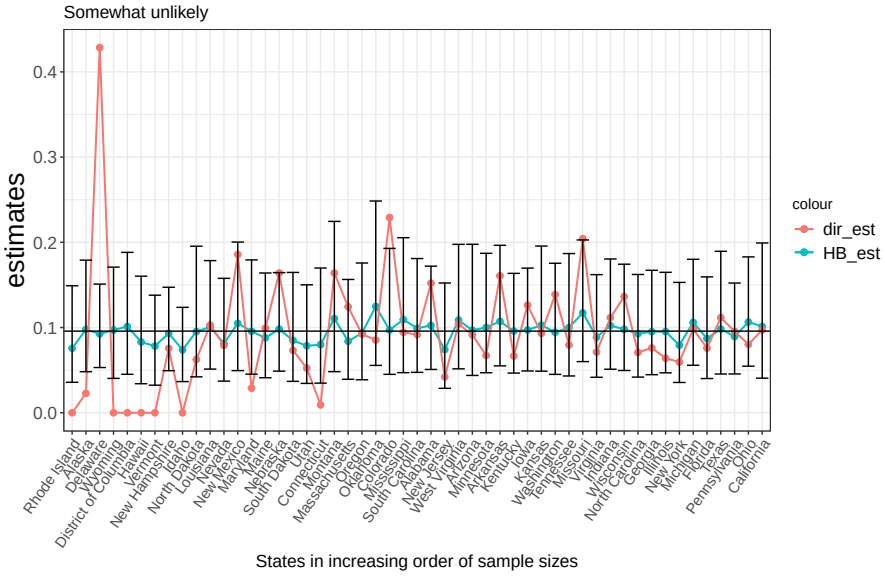


Figure 4.2: Comparison of direct estimates with our proposed HB model-based estimates for the response category 2: 'Somewhat unlikely'. The national estimate is represented by the solid horizontal line.

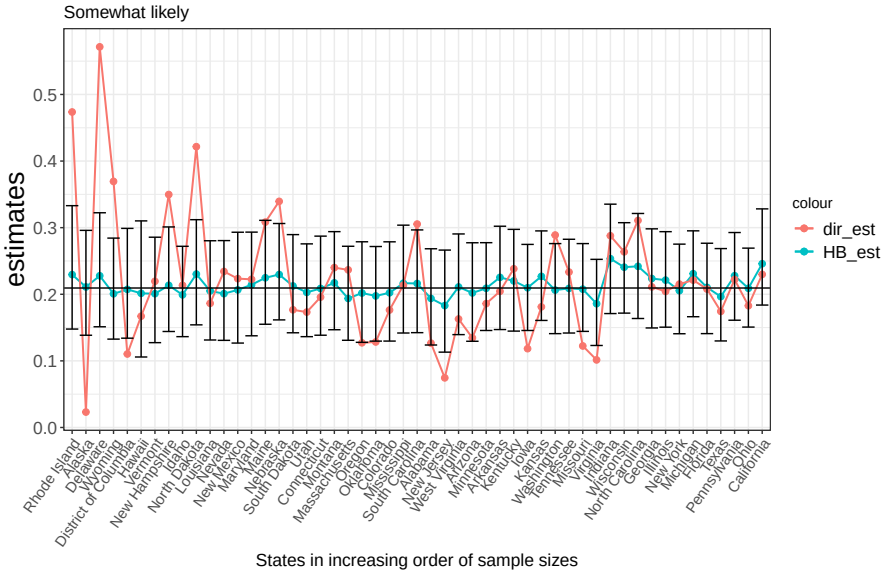


Figure 4.3: Comparison of direct estimates with our proposed HB model-based estimates for the response category 3: 'Somewhat likely'. The national estimate is represented by the solid horizontal line.

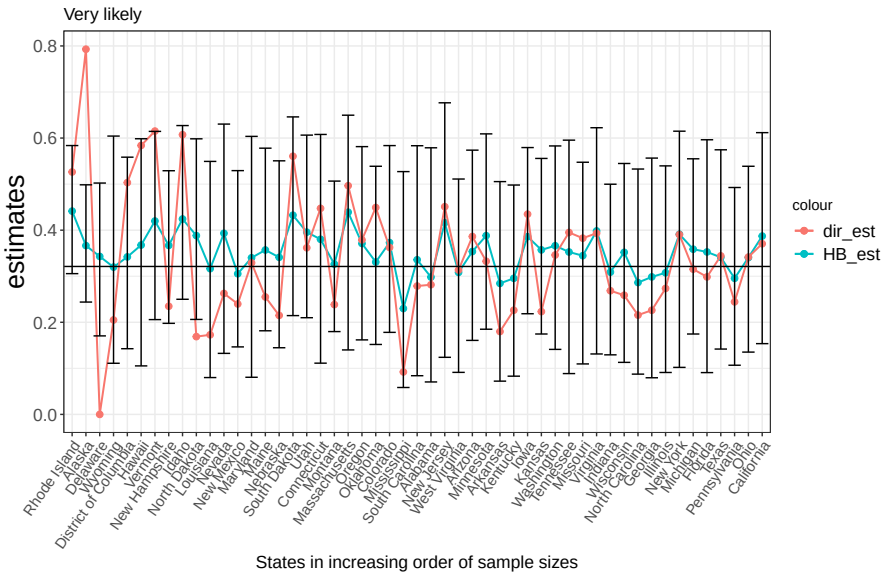


Figure 4.4: Comparison of direct estimates with our proposed HB model-based estimates for the response category 4: 'Very likely'. The national estimate is represented by the solid horizontal line.

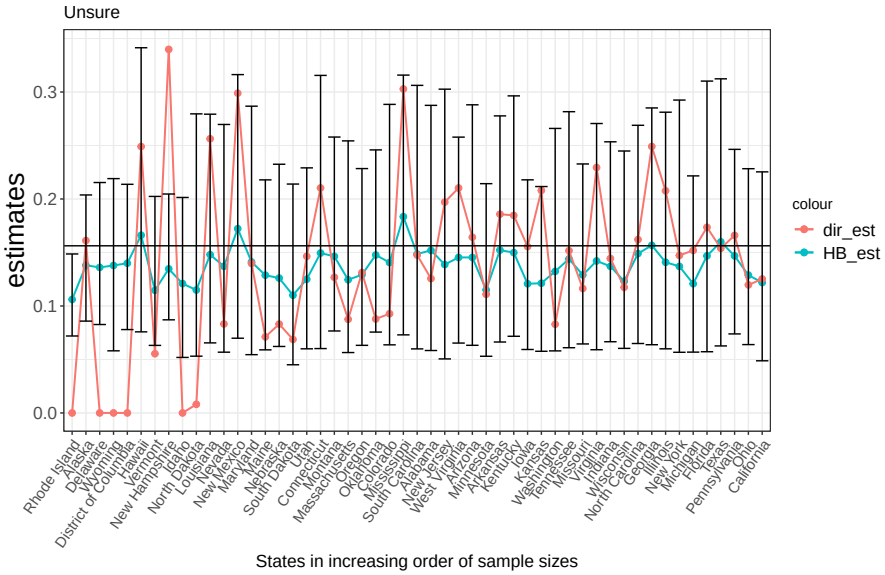


Figure 4.5: Comparison of direct estimates with our proposed HB model-based estimates for the response category 5: 'Unsure'. The national estimate is represented by the solid horizontal line.

demonstrated earlier. Despite incorporating all relevant factors influencing the sampling mechanism from the available data, significant sampling weights were observed across various response variable categories. In such situations, it is crucial to avoid using a single model for the population and the sample in estimation methods.

For illustrative purposes, we generated small area estimates using MRP for our dataset. Below, we present comparison plots of MRP with our proposed methodology. While MRP yields estimates with lower standard errors compared to our model-based approach, it is not suitable for our scenario, as discussed previously. Any method assuming a single model for both the population and the sample is inadequate in the presence of informative sampling.

Figures 4.6, 4.7, 4.8, 4.9, and 4.10 depict the comparison between MRP-based estimates and our proposed HB-model based estimates. Generally, both methods yield similar estimates, with MRP exhibiting lower standard errors. However, in figure 4.7, we observe consistently higher estimates from MRP compared to national survey-based estimates for response category 2: 'Somewhat unlikely'. It is important to note that national survey-based estimates are highly reliable due to the large sample sizes. Thus, for this category, our method outperforms MRP, despite MRP's lower standard errors. We think this is most likely because, for this category, the survey weights were highly significant.

4.5 Conclusion

In this chapter, we explore a specialized application of the general methodology outlined in Chapter 3, specifically tailored to address survey response variables with multiple categories. We investigate the potential impact of biased or informative sampling on the assumed model.

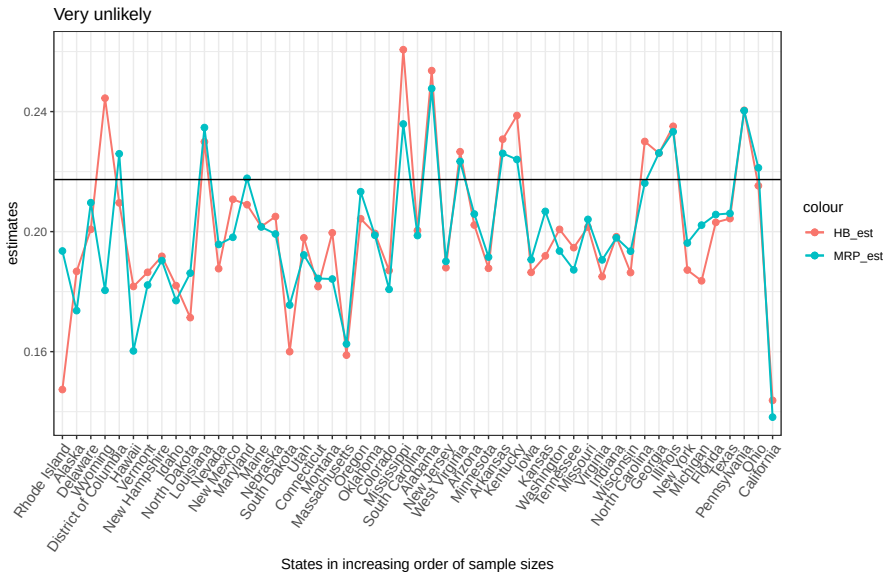


Figure 4.6: Comparison of MRP-based estimates with our proposed HB model-based estimates for the response category 1: 'Very unlikely'. The national estimate is represented by the solid horizontal line.

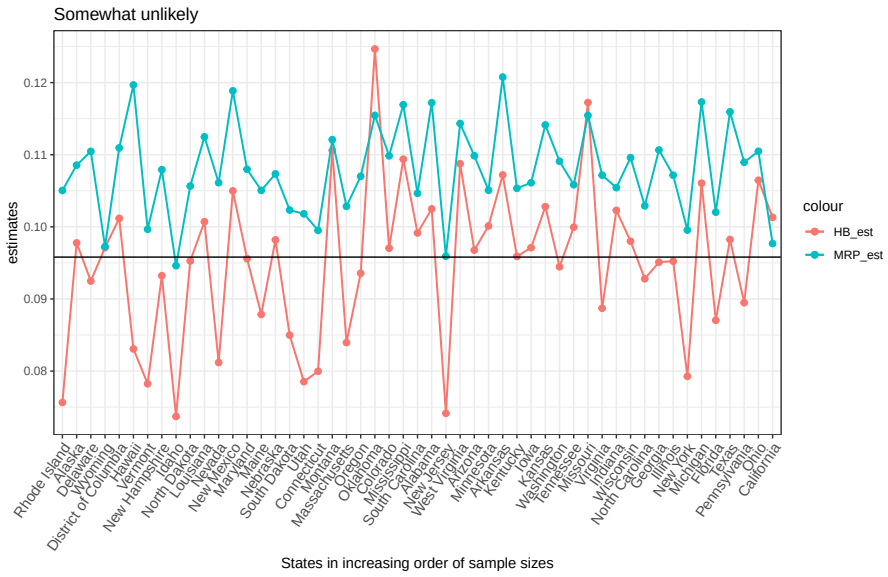


Figure 4.7: Comparison of MRP-based estimates with our proposed HB model-based estimates for the response category 2: 'Somewhat unlikely'. The national estimate is represented by the solid horizontal line.

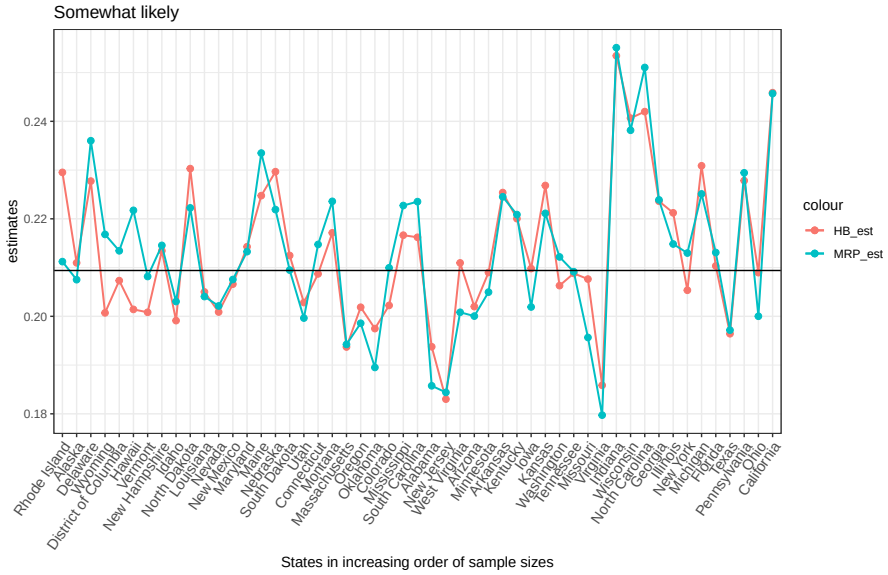


Figure 4.8: Comparison of MRP-based estimates with our proposed HB model-based estimates for the response category 3: 'Somewhat likely'. The national estimate is represented by the solid horizontal line.

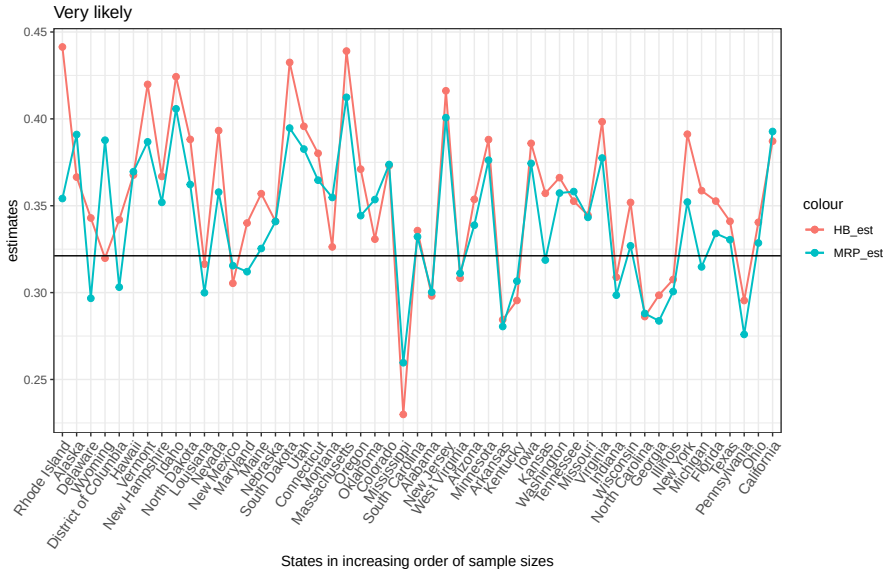


Figure 4.9: Comparison of MRP-based estimates with our proposed HB model-based estimates for the response category 4: 'Very likely'. The national estimate is represented by the solid horizontal line.

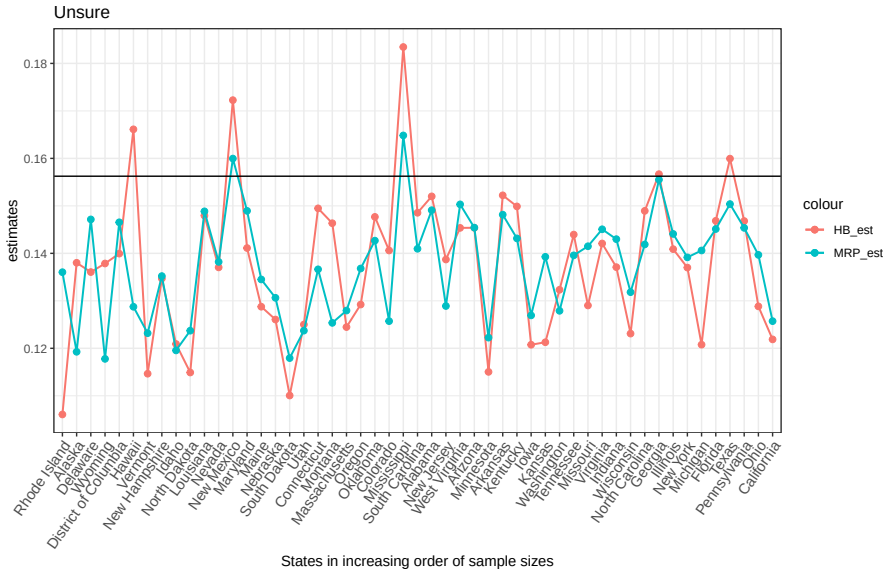


Figure 4.10: Comparison of MRP-based estimates with our proposed HB model-based estimates for the response category 5: 'Unsure'. The national estimate is represented by the solid horizontal line.

We show that when informative sampling is present, employing the assumed population model on the sampled units can yield erroneous/biased inferences. Given that many factors influencing informative sampling are unobserved or inaccessible to survey statisticians, we propose utilizing survey weights and conducting tests to evaluate the statistical significance of the associated coefficient. Our proposed model and estimation method accommodate survey weights seamlessly within the modeling and estimation processes. Furthermore, we conduct a comparative analysis between our approach and Multilevel Regression with Poststratification (MRP), a widely used technique for generating reliable small area estimates. Despite MRP's reputation for precision, we highlight its unsuitability under conditions of informative sampling, which can lead to inflated estimates for certain response categories. However, it is worth noting that MRP generally yields estimates with higher precision compared to our method. Moreover, had benchmarking been utilized, MRP estimates would eventually stabilize, akin to ours. Despite the inadvisability

of employing MRP or similar methods assuming identical models for population and observed sample data in cases of informative sampling, benchmarking can still mitigate over or underestimations within the MRP framework. Future extensions of our work may involve joint modeling of multiple survey response variables, including binary, categorical, and continuous variables, incorporating appropriate correlation structures where necessary to enhance the robustness and accuracy of the modeling framework.

Chapter 5: Discussion and concluding remarks

The dissertation aims to address significant gaps in the estimation of multidimensional poverty measures for small areas, propose novel hierarchical Bayesian estimation procedures for finite population means in small areas, and explore specialized applications of these methods for survey response variables with multiple categories. Central to the dissertation is the utilization of hierarchical Bayesian methods in the realm of small-area estimation. In this final chapter, we provide an overview of the key findings and contributions from each chapter, along with implications and future research directions.

5.1 Summary of Findings

In Chapter 2, we identified a research gap concerning the estimation of multidimensional poverty measures for small areas. We proposed a Bayesian method for estimating these poverty rates and quantifying the relative contributions of different dimensions towards individual-level poverty for small areas. Our approach is versatile and can be easily adapted to accommodate various methods for defining multidimensional poverty, including popular measures of multidimensional poverty methods like counting or fuzzy set methods. Additionally, we introduced a unique approach for determining effective sample sizes, considering the complexity of the survey data and the way it was collected.

Chapter 3 introduced a unified hierarchical Bayesian estimation procedure for finite population means in small areas, leveraging data linkage techniques to integrate diverse data sources, including social media data. We developed an elaborate model for sampled respondents, incorporating survey weights and factors influencing the outcome variable to reduce the informativeness of the sample. Our methodology demonstrated reduced sensitivity to model misspecifications and diminished reliance on assumed models, offering a robust approach for small area estimation challenges.

In Chapter 4, we explored specialized applications of our methodology for survey response variables with multiple categories. We addressed the potential impact of biased or informative sampling on assumed models and proposed strategies to evaluate and mitigate these effects. Comparative analysis with Multilevel Regression with Poststratification (MRP) highlighted the limitations of MRP under conditions of informative sampling, emphasizing the need for alternative approaches to small area estimation.

5.2 Implications and Future Directions

The findings from this dissertation have several implications for research and practice in small area estimation and multidimensional poverty measurement. Our Bayesian estimation procedures offer robust and flexible methods for addressing challenges such as informative sampling, model misspecification, and data integration. These methods have the potential to enhance the accuracy and reliability of small area estimates, providing valuable insights for evidence-based policy-making and resource allocation.

Future research directions may involve extending our methodologies to incorporate addi-

tional data sources, such as sensor data, to further improve the precision and accuracy of small area estimates. Incorporating sensor data into small area estimation (SAE) can enrich analyses by providing real-time, granular insights into environmental conditions, infrastructure usage, and human behavior. By augmenting traditional data sources, sensor data can enhance the accuracy and depth of SAE models, such as the ones considered in this thesis. This will facilitate more precise estimations of socioeconomic indicators and enable targeted interventions to address local disparities and improve community well-being. Random covariance matrices can also be considered for the joint modeling of multidimensional poverty scores. Joint modeling of multiple survey response variables, including binary, categorical, and continuous variables, could enhance the versatility and applicability of our approach across diverse domains. Additionally, exploring alternative strategies for handling informative sampling and evaluating the robustness of estimation procedures under different sampling scenarios would contribute to advancing the field of small area estimation.

5.3 Conclusion

In conclusion, this dissertation contributes to the advancement of small area estimation and multidimensional poverty measurement by introducing novel Bayesian estimation procedures, addressing key methodological challenges, and offering insights into the implications of informative sampling. By developing robust and flexible methods for small area estimation, we aim to provide policymakers and researchers with reliable tools for understanding and addressing socioeconomic disparities at local levels.

Bibliography

[1] Robert E Fay III and Roger A Herriot. Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

[2] Neung Soo Ha, P. Lahiri, and V. Parsons. Methods and results for small area estimation using smoking data from the 2008 national health interview survey. *Statistics in Medicine*, 33(22):3932–3945, 2014.

[3] B. Liu, P. Lahiri, and G. Kalton. Hierarchical bayes modeling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association, Survey Research Section*, pages 3181–3186, 2007.

[4] John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.

[5] Malay Ghosh et al. Small area estimation: Its evolution in five decades. *Statistics in Transition. New Series*, 21(4):1–22, 2020.

[6] Jiming Jiang and Partha Lahiri. Mixed model prediction and small area estimation. *Test*, 15:1–96, 2006.

[7] William Gemmell Cochran. *Sampling techniques*. John Wiley & Sons, 1977.

[8] Sharon L Lohr. *Sampling: design and analysis*. CRC press, 2021.

[9] Wayne A Fuller. *Sampling statistics*. John Wiley & Sons, 2011.

[10] Keith F Rust and JNK Rao. Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research*, 5(3):283–310, 1996.

[11] Douglas Yeo. Bootstrap variance estimation for the national population health survey. In *Proceedings of the Survey Research Methods Section Conference, August 1999*. American Statistical Association, 1999.

[12] Andreea L Erciulescu, Nathan B Cruze, and Balgobin Nandram. Benchmarking a triplet of official estimates. *Environmental and Ecological Statistics*, 25(4):523–547, 2018.

[13] Xin Wang, Emily Berg, Zhengyuan Zhu, Dongchu Sun, and Gabriel Demuth. Small area estimation of proportions with constraint for national resources inventory survey. *Journal of Agricultural, Biological and Environmental Statistics*, 23(4):509–528, 2018.

[14] Andreea L Erciulescu, Carolina Franco, and Partha Lahiri. Chapter: Use of administrative records in small area estimation. 2018.

[15] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[16] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[17] Grace M Carter and John E Rolph. Empirical bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, 69(348):880–885, 1974.

[18] George E Battese, Rachel M Harter, and Wayne A Fuller. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36, 1988.

[19] Diana Maria Stukel and JNK Rao. On small-area estimation under two-fold nested error regression models. *Journal of statistical planning and inference*, 78(1-2):131–147, 1999.

[20] M Ghosh and P Lahiri. Bayes and empirical bayes analysis in multistage sampling. *Statistical Decision Theory and Related Topics IV*, 1:195–212, 1988.

[21] Gauri Sankar Datta and Malay Ghosh. Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, pages 1748–1770, 1991.

[22] Peter McCullagh. *Generalized linear models*. Routledge, 2019.

[23] Brenda MacGibbon and Thomas J Tomberlin. *Small area estimates of proportions via empirical Bayes techniques*. Faculty of Commerce and Administration, Concordia University, 1987.

[24] Donald Malec, J Sedransk, Christopher L Moriarity, and Felicia B LeClere. Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, 92(439):815–826, 1997.

[25] TWF Stroud. Hierarchical baxes predictive means and variances with application to sample survey inference. *Communications in Statistics-Theory and Methods*, 20(1):13–36, 1991.

[26] Jiming Jiang and P Lahiri. Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53:217–243, 2001.

[27] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, 2007.

[28] J. Jiang and P. Lahiri. Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101:301–311, 2006.

[29] Charles R Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2):226–252, 1953.

[30] NG Narasimha Prasad and Jon NK Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

[31] Carl N Morris. Approximating posterior distributions and posterior moments. *Bayesian statistics*, 3:327–344, 1988.

[32] Julie Gershunskaya, Jiming Jiang, and P Lahiri. Resampling methods in surveys. In *Handbook of Statistics*, volume 29, pages 121–151. Elsevier, 2009.

[33] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. 2023.

[34] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[35] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

[36] Aki Vehtari, Jonah Gabry, Mans Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2020. R package version 2.4.1.

[37] Andrew Gelman and Thomas Little. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127, 1997.

[38] Xingyou Zhang, James B Holt, Hua Lu, Anne G Wheaton, Earl S Ford, Kurt J Greenlund, and Janet B Croft. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American journal of epidemiology*, 179(8):1025–1033, 2014.

[39] PI Eke, X Zhang, H Lu, L Wei, G Thornton-Evans, KJ Greenlund, JB Holt, and JB Croft. Predicting periodontitis at state and local levels in the united states. *Journal of dental research*, 95(5):515–522, 2016.

[40] Marnie Downes, Lyle C Gurrin, Dallas R English, Jane Pirkis, Dianne Currier, Matthew J Spittal, and John B Carlin. Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *American journal of epidemiology*, 187(8):1780–1790, 2018.

[41] Yajuan Si, Rob Trangucci, Jonah Sol Gabry, and Andrew Gelman. Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology*, page 181, 2020.

[42] Amartya Sen. Equality of what? *The Tanner Lecture on Human Values*, 1980.

[43] Andrea Cerioli and Sergio Zani. A fuzzy approach to the measurement of poverty. In *Income and wealth distribution, inequality and poverty*, pages 272–284. Springer, 1990.

[44] Sabina Alkire, José Manuel Roche, Paola Ballon, James Foster, Maria Emma Santos, and Suman Seth. *Multidimensional poverty measurement and analysis*. Oxford University Press, USA, 2015.

[45] Dilshanie Deepawansa, Priyanga Dunusinghe, Partha Lahiri, and Ramani Gunatilaka. An innovative approach to measure multidimensional poverty: A synthesis method. *Statistical Journal of the IAOS*, 38(4):1303–1323, 2022.

[46] Soumojit Das and Partha Lahiri. Hierarchical bayes estimation of small area means using statistical linkage of disparate data sources, 2023.

[47] GS Datta, RE Fay, and M Ghosh. Hierarchical and empirical bayes multivariate analysis in small area estimation. In *Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, Washington, DC*, pages 63–79, 1991.

[48] Roberto Benavent and Domingo Morales. Multivariate fay–herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94:372–390, 2016.

[49] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. 2008.

[50] William A Ericson. Subjective bayesian models in sampling finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):195–224, 1969.

[51] Malay Ghosh and Glen Meeden. Empirical bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81(396):1058–1062, 1986.

[52] Malay Ghosh and Parthasarathi Lahiri. Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82(400):1153–1162, 1987.

[53] Malay Ghosh, Parthasarthi Lahiri, and Ram C Tiwari. Nonparametric empirical bayes estimation of the distribution function and the mean. *Communications in Statistics-Theory and Methods*, 18(1):121–146, 1989.

[54] B Nandram and J Sedransk. Empirical bayes estimation for the finite population mean on the current occasion. *Journal of the American Statistical Association*, 88(423):994–1000, 1993.

[55] Vipin Arora, Partha Lahiri, and Kanchan Mukherjee. Empirical bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association*, 92(440):1555–1562, 1997.

[56] Jiming Jiang and Thuan Nguyen. Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, 40:588–663, 2012.

[57] Gauri Sankar Datta and Malay Ghosh. Bayesian Prediction in Linear Models: Applications to Small Area Estimation. *The Annals of Statistics*, 19(4):1748 – 1770, 1991.

[58] Malay Ghosh and Parthasarathi Lahiri. A hierarchical bayes approach to small area estimation with auxiliary information. In *Bayesian Analysis in Statistics and Econometrics*, pages 107–125. Springer, 1992.

[59] Gauri Sankar Datta and Malay Ghosh. Hierarchical bayes estimators of the error variance in one-way anova models. *Journal of statistical planning and inference*, 45(3):399–411, 1995.

[60] M. Ghosh and G. Meeden. *Bayesian Methods for Finite Population Sampling*. Chapman Hall, 1997.

[61] Roderick J Little. To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556, 2004.

[62] Qixuan Chen, Michael R Elliott, and Roderick JA Little. Bayesian inference for finite population quantiles from unequal probability samples. *Survey methodology*, 38(2):203, 2012.

[63] Malay Ghosh. Bayesian developments in survey sampling. In *Handbook of Statistics*, volume 29, pages 153–187. Elsevier, 2009.

[64] Benmei Liu and Partha Lahiri. Adaptive hierarchical bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, 69(2):150–164, 2017.

[65] Balgobin Nandram, Lu Chen, Shuting Fu, and Binod Manandhar. Bayesian logistic regression for small areas with numerous households. *arXiv preprint arXiv:1806.00446*, 2018.

[66] Neung Soo Ha and Joseph Sedransk. Assessing health insurance coverage in florida using the behavioral risk factor surveillance system. *Statistics in medicine*, 38(13):2332–2352, 2019.

[67] Gauri S Datta. Model-based approach to small area estimation. *Handbook of statistics*, 29:251–288, 2009.

[68] D. Pfeffermann. New important developments in small area estimation. *Statistical Science*, 28:40–68, 2013.

[69] M. Ghosh. Small area estimation: Its evolution in five decades. *Statistics in Transition New Series, Special Issue on Statistical Data Integration*, pages 1–67, 2020.

[70] Alastair J Scott. Some comments on the problem of randomisation in surveys. *Sankhya C*, 39:1–9, 1977.

[71] Danny Pfeffermann and Michail Sverchkov. Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 166–186, 1999.

[72] Daniel Bonnéry, F Jay Breidt, and François Coquet. Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli*, 18(4):1361–1385, 2012.

[73] D. Pfeffermann and M. Sverchkov. Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102 (480):1427–1439, 2007.

[74] Donald B Rubin. An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. *Journal of the American Statistical Association*, 78(384):803–805, 1983.

[75] Francois Verret, Jon NK Rao, and Michael A Hidiroglou. Model-based small area estimation under informative sampling. *Survey Methodology*, 41(2):333–348, 2015.

[76] Carl N Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

[77] P. Lahiri and J. Suntornchost. A general bayesian approach to meet different inferential goals in poverty research for small areas. *STATISTICS IN TRANSITION, Special Issue*, pages 245–261, 2020.

[78] Bob Carpenter, Daniel Lee, Marcus A. Brubaker, Allen Riddell, Andrew Gelman, Ben Goodrich, Jiqiang Guo, Matt Hoffman, Michael Betancourt, and Peter Li. Stan: A probabilistic programming language.

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[80] Stan Development Team. RStan: the R interface to Stan, 2020. R package version 2.21.2.