Estimating subnational health and demographic indicators using complex survey data

Peter A. Gao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Jonathan Wakefield, Chair

Elena Erosheva

Thomas Lumley

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Estimating subnational health and demographic indicators using complex survey data

Peter A. Gao

Chair of the Supervisory Committee:
Professor of Statistics and Biostatistics Jonathan Wakefield
Department of Statistics

Subnational estimates of health and demographic indicators such as immunization coverage rates and child mortality rates are critical for identifying regional health disparities and guiding policy design. When population data on an outcome of interest are unavailable or incomplete, many countries gather information from a sample of the population using household surveys. These surveys are typically designed for producing national estimates of key indicators, but generally do not collect sufficient data to produce reliable subnational estimates using traditional direct estimation methods, especially when estimating the prevalence of rare events. In this setting, indirect methods that use statistical models to incorporate covariate information or smooth estimates across areas using random effects can be effective for generating more precise estimates. However, national statistical offices and policymakers commonly desire estimators that are robust to model misspecification, making careful selection of methods crucial for producing estimates that are acceptable for dissemination and decision making. In recent years, geostatistical models which treat quantities of interest as continuous spatial surfaces have become popular among global health researchers for mapping key health indicators, especially for low- and middle- income countries. These approaches often compensate for limited data avail-

ability by leveraging advances in spatial modeling and incorporating newly available covariate information derived from satellite imaging, but may fail to account for features of the complex surveys used to collect data, such as informative sampling or cluster effects, potentially leading to biased estimates. On the other hand, traditional small area estimation approaches common in the survey statistics literature are typically specified with careful consideration for survey design, but have historically been adopted in countries where high-quality census data on auxiliary covariates are available and may perform suboptimally in low data settings.

In this thesis, I propose a suite of methods for estimating subnational health and demographic indicators using complex survey data. First, I propose an area level model for demographic rates that jointly models the direct estimators and associated variance estimators and induces spatial smoothing of both means and variances. This method can be viewed as an extension of the Fay-Herriot model popular for small area estimation that is adapted for estimation of small area proportions. Second, I outline a smoothed model-assisted estimator for small area means that incorporates unit level covariate information and smoothing via random effects while accounting for the survey design via the use of survey weights. Finally, I describe a method for incorporating sampling weights when estimating unit level models in order to address the effects of clustering and informative sampling.

As a whole, these methods bridge traditional small area estimation approaches and geostatistical models commonly used in global health research. These methods leverage recent advances in spatial modeling and adopt a fully Bayesian approach to estimation and inference, showing how modern Bayesian methods and software popular in global health research can be adopted for small area estimation.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables vi

Chapter 1: Introduction 1

1.1 Small area estimation 1

1.2 Motivating example 3

Chapter 2: Literature review 8

2.1 Notation and definitions 8

2.2 Sampling design 9

2.3 Inferential frameworks 11

2.4 Existing approaches for small area estimation 17

Chapter 3: A variance smoothing areal model for estimating proportions 30

3.1 Introduction 30

3.2 Existing approaches 34

3.3 A joint mean- and variance-smoothing model 40

3.4 Simulation study 45

3.5 Applications 48

3.6 Discussion 51

Chapter 4: Combining area level and unit level modeling for small area estimation of proportions . . . 55
4.1 Introduction . . . 55
4.2 Simulation study . . . 61
4.3 Application: Vaccination coverage in Nigeria . . . 68
4.4 Discussion . . . 72

Chapter 5: Unit level modeling for small area estimation under informative sampling . . . 75
5.1 Introduction . . . 75
5.2 Background and inferential framework . . . 78
5.3 Standard estimation approaches . . . 83
5.4 Proposed approach . . . 87
5.5 Simulations . . . 91
5.6 Application: Vaccination coverage in Guinea . . . 102
5.7 Discussion . . . 107

Chapter 6: Conclusion . . . 110

Appendix A: A variance smoothing areal model for estimating proportions . . . 113
A.1 Parameter estimation . . . 113
A.2 Additional results . . . 114

Appendix B: Combining area level and unit level modeling for small area estimation for proportions . . . 120
B.1 Design consistency of survey regression LGREG estimator . . . 120
B.2 Parameter estimation . . . 122
B.3 Additional results . . . 124

Appendix C: Unit level modeling for small area estimation under informative sampling . . . 126
C.1 Misspecification of the superpopulation model . . . 126
C.2 Pseudo-posterior convergence . . . 127
C.3 Estimating the multivariate design effect . . . 129
C.4 Estimation procedure . . . 130

References . . . 132

LIST OF FIGURES

Figure Number Page

1.1 Map of Nigeria with Admin-1 level boundaries. Points indicate enumeration area locations (randomly displaced for privacy) for which data on measles vaccination is available. 5

3.1 Direct weighted estimates of vaccination coverage rate for first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria, 2018 (left) and HIV prevalence rate for women aged 15–49 in Malawi, 2015–2016 (right). 33

3.2 Small area boundaries and sampled enumeration area locations for 2015–16 Malawi DHS. 35

3.3 Direct and model-based point estimates (top) and length of corresponding 90% interval estimates (bottom) of vaccination coverage rate for first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria, 2018. 49

3.4 Direct and model-based point estimates (top) and length of corresponding 90% interval estimates (bottom) of HIV prevalence rate for women aged 15–49 in Malawi, 2015–2016. 50

3.5 Comparison of model-based 90% credible interval lengths with Hájek 90% confidence interval lengths for Malawi HIV example (left) and Nigeria MCV example (right). 51

4.1 Estimated measles vaccination rates (left) and 90% prediction interval lengths for estimated measles vaccination rates (right) among children aged 12-23 months for Admin-1 areas in Nigeria in 2018. 70

4.2 Comparison of reference Hájek design-based 90% confidence interval length (x-axis) with model-assisted and model-based interval estimate lengths (y-axis) for four methods for measles vaccination rates. 71

5.1 Map of Guinea with Admin-1 level boundaries (thick borders) and Admin-2 level boundaries (thin borders). Points indicate enumeration area locations for which data on measles vaccination is available. 104

5.2 Estimated measles vaccination rates among children aged 12–23 months for Admin-2 areas in Guinea in 2018. 105

5.3 Prediction interval lengths for estimated measles vaccination rates for Admin-2 areas in Guinea in 2018. 106

A.1 Simulated cluster locations and covariate values used to generate population data. 115

LIST OF TABLES

Table Number Page

3.1 RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and mean interval length ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling. The reduced model omits one of the spatial covariates in the full model. 47

4.1 Averaged RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and MIL ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling for methods using no covariates or only the reduced set of covariates (omitting one of the spatial covariates used in population generation). The lowest RMSE and MAE are in ***bold italics***. 66

4.2 Averaged RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and MIL ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling for methods using the full set of covariates. The lowest RMSE and MAE are in ***bold italics***. 67

5.1 Averaged evaluation metrics of estimators of area level means across 1,000 continuous response simulations for SRS, PPS1, and PPS2 designs for a sample size of 30 units per area. 98

5.2 Averaged evaluation metrics of estimators of area level means across 1,000 continuous response simulations for SRS, PPS1, and PPS2 designs for a sample size of 100 units per area. 99

5.3 Averaged evaluation metrics of estimators of area level means across 1,000 binary response simulations for SRS, PPS1, and PPS2 designs for a sample size of 30 units per area. 100

5.4 Averaged evaluation metrics of estimators of area level means across 1,000 binary response simulations for SRS, PPS1, and PPS2 designs for a sample size of 100 units per area. 101

A.1 RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and mean interval length ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on large sample (25 clusters) obtained via informative sampling. The reduced model omits one of the spatial covariates in the full model. 115

A.2 Point estimates and 90% interval estimates for model parameters for Nigeria measles vaccination example. 116

A.3 Point estimates and 90% interval estimates for model parameters for Malawi HIV prevalence example. 117

A.4 Point estimates of measles vaccination rates and 90% interval estimates for Admin-1 areas among children aged 12–23 months in Nigeria in 2018. . . . 118

A.5 Point estimates of HIV prevalence and 90% interval estimates for Admin-1 areas among women aged 15–49 in Nigeria in 2018. 119

B.1 Estimated measles vaccination rates (left) with 90% prediction intervals for Admin-1 areas in Nigeria in 2018. 125

ACKNOWLEDGMENTS

First, I owe my thanks to the individuals who agreed to be interviewed and shared their family histories with the Demographic and Health Surveys for the questionnaires used throughout this dissertation. Their contributions, along with those of the interviewers and survey specialists working with the Demographic and Health Surveys have made this work possible.

I thank my advisor, Jon Wakefield, for his support and guidance. There were many times when I felt lost in the weeds throughout the research process, but thanks to Jon's patience and encouragement, it was usually the good kind of being lost.

I also want to thank my other members of my committee, Elena Erosheva, Thomas Lumley, and Zack Almquist, who have supported me through the dissertation process. I have also benefited greatly from having wonderful mentors at the University of Washington and the University of Chicago, especially Abel Rodriguez, June Morita, Adrian Raftery, and Dan Nicolae. The Statistics Department staff, particularly Kristine Chan, Tracy Pham, Ellen Reynolds, and Asa Sourdiffe, have helped and supported me over the past six years.

In graduate school, my classmates have been an immense source of support and advice throughout the years, including Wenyu Chen, Hannah Director, Kristof Glauninger, Richard Li, Daphne Liu, Bryan Martin, Alan Min, Anna Neufeld, Michael Pearce, Anupreet Porwal, Max Schneider, Sarah Teichman, Steven Wilkins-Reeves, and Nathan Welch. I am especially thankful for the members of the STAB lab, including Serge Aleshin-Guendel, Jessica Godwin, Taylor Okonek, John Paige, and Austin Schumacher, for being exceptional role models and teachers. I also want to thank my friends from

other departments including Nicasia Beebe-Wang, Mark Boyer, Derrick Ho, Tia Nguyen, Hannah Lee, Selen Güler, Lukas Hager, and Nick Hadjimichael.

I thank my friends from outside of graduate school who supported me enthusiastically in Seattle and elsewhere, especially Matt Basile, Sam Baugh, Christian Belanger, Katrina Deloso, Amelia Dmowska, Kenzo Esquivel, Julia Guo, Mike Hua, Aaron Jacobs, Courtney Kan, Clara Kao, Jack Liang, Maggy Liu, Henry Lewis, Kristin Lin, Ian Morse, Eunice Park, Joon Pyun, Chris Williams, Andrew Yang, and Ron Yehoshua. Finally, I thank Crystal Liu, whose perspective, care, and encouragement enabled me to reach the end of this journey.

My parents granted me the freedom to pursue my interests and I will always be grateful for their love and support. I was always chasing after my brother David growing up and I still learn from him every day. This dissertation is dedicated to my mother, who returned to school in pursuit of her own doctorate when I was in high school.

DEDICATION

For my mother

Chapter 1

INTRODUCTION

1.1 Small area estimation

Reliable estimates of subnational health-related indicators such as child mortality rates and vaccination coverage rates are crucial for guiding health policy and decisions related to resource allocation. In 2015, the United Nations General Assembly prioritized the elimination of preventable deaths under five years of age under its Sustainable Development Goals agenda (United Nations 2015). As part of tracking progress towards these goals, subnational estimates can shed light on regional health disparities in achieving reductions in mortality and identify key targets for health interventions (United Nations Inter-agency Group for Child and Mortality Estimation 2021).

Policymakers and researchers often desire subnational estimates of outcomes of interest for which complete population data are unavailable and many low- and middle- income countries (LMIC) lack vital registration systems with full coverage, so household surveys are often used to collect data from a sample of the population. These surveys, often carried out by multinational programs such as the Multiple Indicator Cluster Surveys (MICS) or the Demographic and Health Surveys (DHS), are typically designed to produce high-quality national level estimates, but generally do not collect sufficient data to produce reliable estimates at the resolution desired for making decisions, especially for outcomes that are rare.

This problem, of estimating quantities for domains such as subnational areas using limited data, is well-known in the survey statistics literature, where it is called small area

estimation (SAE). Pfeffermann (2013), Rao and Molina (2015) and Ghosh (2020) review common approaches and recent developments in SAE. SAE have been adopted for subnational mapping of a wide variety of outcomes, including poverty indicators (Bell et al. 2016; Corral Rodas et al. 2021; Marhuenda et al. 2017), health outcomes (Congdon and Lloyd 2010) and crop production estimates (Erciulescu et al. 2019). Wakefield et al. (2020) discuss the use of SAE methods for mapping subnational disease prevalence rates.

In areas for which data are limited or unavailable, traditional direct small area estimators such as the Horvitz-Thompson or Hájek estimators can be unreliable. Indirect methods, which share information across areas using statistical models to incorporate covariate estimation or smooth estimates using random effects can produce more precise estimates. However, survey practitioners typically prefer estimates that rely on relatively few modeling assumptions, so indirect model-based small area estimators in the survey statistics literature are often specified to be robust to model misspecification and to carefully consider the design of the surveys used to collect data. However, these estimators have historically been used in countries where high-quality census data on auxiliary covariates are available and may perform suboptimally when such data are unavailable.

As an alternative to traditional model-based SAE approaches in LMIC, global health researchers commonly turn to geostatistical models which treat quantities of interest as continuous spatial surfaces for mapping key health outcomes. Geostatistical methods have been adopted to develop high-resolution pixel maps of health-related outcomes including disease prevalence (Diggle and Giorgi 2016), vaccination rates (Utazi et al. 2020), and neonatal and child mortality (Golding et al. 2017). These approaches often compensate for limited data availability by leveraging advances in spatial modeling and incorporating newly available covariate information derived from satellite imaging, but may fail to account for features of the complex surveys used to collect data, such as informative sampling or cluster effects. For example, the DHS, which conducts household surveys to collect health-related data, typically uses a multistage stratified clustered design, over-

sampling clusters in urban areas. Neglecting to consider aspects of the survey design including stratification and oversampling may result in biased or poorly calibrated estimates.

Survey-based estimators of small area quantities are typically evaluated under one of two frameworks for population inference: either the design-based framework, or the model-based framework. The design-based perspective assumes that population responses are non-random and examines performance with respect to repeated sampling. The model-based perspective assumes that population responses are generated from a superpopulation and evaluates estimators by taking expectations over both the sampling process and the population-generating process. When mapping critical health and demographic indicators such as child mortality rates, government national statistics offices often prefer estimators that satisfy design-optimality properties, such as being unbiased with respect to repeated sampling. Such estimators are generally robust to model misspecification, so traditional small area estimators from the survey statistics literature are often specified to meet these design-optimality conditions, while the design-based properties of estimators based on the geostatistical models common in global health research have not been investigated.

1.2 Motivating example

Consider the following motivating example, which introduces challenges that are characteristic of SAE problems involving health and demographic indicators in LMIC. Using survey data from the DHS Program, we wish to estimate measles vaccination coverage rates for subnational areas in Nigeria. This example was previously described by Fuglstad, Li, and Wakefield (2022)

In many LMIC, the DHS Program regularly conducts household surveys to collect health and population data. In general, the DHS Program uses a stratified two-stage cluster sampling design within each country. Countries are divided into principal administra-

tive divisions, also called Admin-1 regions. The Admin-1 regions are each partitioned into urban and rural divisions. Sampling is stratified by these subdivisions (crossing Admin-1 region with urban/rural status). Each stratum is divided into smaller collections of households called enumeration areas (EAs) or clusters. Within each stratum, a specified number of EAs is sampled with probability proportional to size (PPS), specifically meaning the number of households in the EA based on available census data. The households in each selected EA are enumerated, and a specified number of households is sampled from each. Inclusion probabilities and sampling weights are calculated and reported for each household.

An example goal is to estimate subnational vaccine coverage rates for the first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria using data from the 2018 Nigeria DHS. The 2018 DHS collected data on vaccination status for children in sampled households based on vaccination cards or caregiver recall. The sampling frame used for the 2018 DHS was based on a 2006 national census and divides Nigeria into 664,999 EAs and 74 strata. Data were successfully collected in 1389 EAs. A number of clusters were dropped due to security issues during the household listing operation. As a result, as noted in Appendix A.3 of the Nigeria DHS Final Report, estimates for the Admin-1 area of Borno may not be representative of omitted EAs. Geographic coordinates are available for almost all EAs, but have been randomly displaced by small distances to maintain privacy. Figure 1.1 provides a map of the Admin-1 boundaries and EA locations in Nigeria for which data is available. The two-stage stratified cluster design used by the DHS program complicates estimation of subnational means and totals. Urban EAs are oversampled relative to rural EAs, so estimators must account for any systematic differences between urban and rural households. In the case of measles vaccination in Nigeria, urban areas exhibit higher rates of vaccination than rural areas (Dong and Wakefield 2021). Moreover, although DHS data are often adequate for computing reliable direct estimators of indicators at the Admin-1 level, data may not be available to

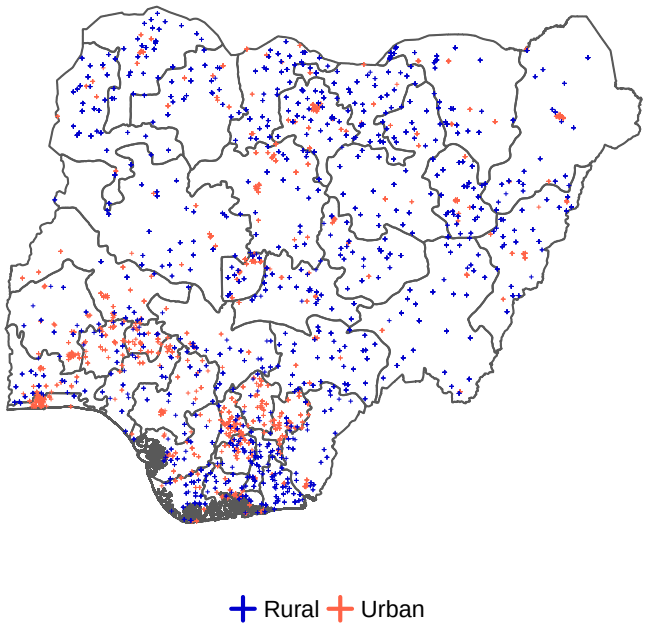

Figure 1.1: Map of Nigeria with Admin-1 level boundaries. Points indicate enumeration area locations (randomly displaced for privacy) for which data on measles vaccination is available.

obtain direct estimates of subregional rates with sufficient precision. Statistical modeling can be used to generate estimates with reduced variability using covariate information and smoothing via random effects.

In this thesis, we propose a suite of methods for using complex survey data to estimating subnational health and demographic indicators. These methods seek to bridge the SAE and geostatistical modeling literatures, borrowing ideas from both. The goal of this hybrid approach is to develop estimators that have favorable characteristics in both the model-based and design-based senses, leveraging statistical modeling to share information across areas while acknowledging the design of surveys used for data collection. These methods adopt a fully Bayesian approach to inference, enabling the production of point estimates as well as estimates of uncertainty. Finally, these methods take advantage of recent advances in computation for Bayesian inference, and we provide code to implement the models described. One primary goal of this work is to develop simple, interpretable estimators that can be viewed as practical alternatives to the often computationally intensive methods popularly adopted in this context. The remainder of this manuscript details these proposed methods as outlined below.

Chapter 2 provides an review of definitions, methods, and theoretical results related to analysis of complex survey data and SAE. In addition, we review geostatistical modeling approaches commonly adopted for estimating subnational health and demographic indicators in LMIC.

Chapter 3 proposes an areal model that extends the commonly used Fay-Herriot model for estimating demographic rates. While the Fay-Herriot model generally assumes the availability of a set of direct estimators and treats the associated design-based variances as known, we propose a model that jointly models the direct estimators and associated variance estimators, inducing spatial smoothing of both the estimated rates and variances.

Chapter 4 proposes a second extension to area level modeling for SAE that incorporates unit level covariate information and smoothing via random effects to produce a smoothed

model-assisted estimator that accounts for survey design using sampling weights.

Chapter 5 covers the use of survey weights when estimating unit level models for SAE Small Area Estimation with Random Forests and the LASSO. Typical unit level approaches may not yield design-unbiased or design-consistent estimators, so we propose a pseudo-Bayesian approach that uses a survey-weighted pseudo-posterior distribution to generate small area estimates that are robust to potential clustering and informative sampling effects.

We conclude with a discussion of our proposed approaches and a description of future challenges and research directions in Chapter 6.

Chapter 2

LITERATURE REVIEW

This chapter reviews existing definitions and methods in survey statistics and SAE to contexualize our proposed approaches. First, we provide key definitions and notation for describing estimation problems using complex survey data. We continue by introducing three frameworks for conducting population inference using survey data and evaluating sample-based estimators: a design-based approach, a model-based inference, and a combined model- and design-based approach. Finally, we review and differentiate between three classes of methods used for SAE: direct, model-assisted, and model-based. Relevant references are provided below but much of the material reviewed here has previously been presented in texts by Särndal, Swensson, and Wretman (2003) and Rao and Molina (2015).

2.1 Notation and definitions

Let $U = \{1, \ldots, N\}$ denote a set of indices for a finite population of size N. Each index is associated with a sample unit in our population; in general, we use the term unit to refer to an individual person, but in some cases it may refer to a household or a cluster of households. We assume U is partitioned into m disjoint administrative areas, $U = U(1) \cup \cdots \cup U(m)$, with $U(i)$ being the set of $N(i)$ indices corresponding to units in area i. For all $j \in U$, we use y_j to denote the value of a variable of interest for unit j and $\mathbf{z}_j$ to denote a vector of auxiliary variables.

We use $S = \{j_1, \ldots, j_n\} \subset U$ to denote a random set of n sampled indices, letting $S = S(1) \cup \cdots \cup S(m)$ be the corresponding partition by administrative area. For all $j \in U$, we

define δ_j to be the inclusion indicator for unit j, so $\delta_j = 1$ if $j \in S$ and $\delta_j = 0$ otherwise. We assume a probability sampling scheme, where S is random, and for all $j \in U$, we let π_j denote the probability that $j \in S$, also called the inclusion probability of unit j, which may depend on $\mathbf{z}_j$. Finally, we let $w_j = 1/\pi_j$ denote the sampling or design weight for unit j (defined as the inverse inclusion probability).

Since our interest is in domain estimation, it will also be helpful to define notation for referring to units within areas. Following Rao and Molina (2015), we let $y_{ij} = y_j$ if $j \in U(i)$ and $y_{ij} = 0$ otherwise. We define δ_{ij}, π_{ij}, w_{ij}, and $\mathbf{z}_{ij}$ analogously. We define $\mathbf{Z}$ to be the matrix of auxiliary variables.

Example 1. We focus on estimation of area-specific proportions such as rates of vaccination coverage. In our vaccination coverage example, $y_{ij} \in \{0, 1\}$ are binary variables with a value of 1 indicating vaccination. Our targets of estimation are the area-specific means $\overline{Y}_i$:

$$\overline{Y}_i = \frac{1}{N(i)} \sum_{j \in U(i)} y_{ij} \tag{2.1}$$

2.2 Sampling design

We focus on data collected via probability sampling designs, under which each unit has a known and positive selection probability. We formally define a sampling design as a probability distribution over all possible samples, defined conditionally on design variables included in $\mathbf{Z}$.

Definition 1. If $\mathcal{S}$ is the set of subsets of the population U, a **sampling design** is a probability distribution, denoted $\mathbb{P}_D$, over $\mathcal{S}$: $\mathbb{P}_D : \mathcal{S} \to [0, 1]$. The sample space can alternatively be defined by the sample space of the random vector of inclusion indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$.

Typically, a sampling procedure is specified which in turn implicitly defines a probability distribution $\mathbb{P}_D$ for a given population. Below, we primarily consider the following sampling designs.

Example 2. (**Common sampling designs**) For further discussion of common sampling designs, see Lohr (2019).

- **Simple random sampling without replacement**: We consider fixed size simple random sampling without replacement (SRSWOR). An SRS of fixed sample size n assigns the same probability to each possible sample of size n. Under SRS, each unit has an equal probability of appearing in the sample.
- **Stratified random sampling without replacement**: For stratified random sampling, the population is first partitioned into subgroups, or strata. Within each stratum, a sample of fixed size is taken via simple random sampling, without replacement. The strata may correspond to subgroups of interest or subgroups that facilitate efficient sampling.
- **Cluster sampling**: Under cluster sampling, individuals in the population are organized into groups called clusters. These clusters become the **primary sampling units** (PSUs) and sampling proceeds by selecting a subset of PSUs in the population at random. In stratified cluster sampling, a fixed number of clusters may be sampled from each stratum. In one-stage cluster sampling, all units within a selected cluster are sampled.
- **Two-stage stratified cluster sampling**: Under this design, sampling proceeds in two stages. First, from each stratum, a pre-specified number of clusters or PSUs is sampled. Next, from each sampled cluster, a fixed number of individuals, or **secondary sampling units** (SSUs) is sampled at random. In most of the countries in the DHS program, the DHS adopts a two-stage stratified cluster sampling design.

2.3 Inferential frameworks

We consider a number of competing frameworks for conducting population inference based on randomly sampled survey data. We are primarily concerned with three perspectives commonly adopted for SAE: design-based inference, model-based inference, and a combined model- and design-based approach. The design-based perspective assumes a fixed finite population of sampling units and conducts inference based on the randomization distribution implicitly defined by the sampling design. The model-based perspective assumes the finite population responses are drawn from some data-generating model and that the survey design is ignorable with respect to the specified model. The model- and design-based view combines the two perspectives: the finite population responses are drawn from some superpopulation model, but the survey design is generally not treated as ignorable.

In essence, to conduct inference, we must construct some hypothetical reference set and corresponding probability distribution for our actually observed data. Should we compare our sample to all the samples that could have arisen under our sampling procedure (the randomization distribution) or to all possible samples that could have been drawn from a population generating model? These varying inferential strategies may not always produce the same conclusions, leading to disagreement over the correct approach. Methods that are optimal in a model-based sense may be biased in a design-based sense; design-optimal methods may be inefficient under certain assumed models. This debate is somewhat particular to survey statistics, as noted by Little (2004), who wrote, "Finite population sampling is perhaps the only area of statistics in which the primary mode of analysis is based on the randomization distribution, rather than on statistical models for the measured variable."

Our approach is motivated by pragmatism rather than strict adherence to one of the perspectives. Given sufficient data, design-based inferences may be preferable given that

they do not depend on proper specification of a data-generating model (though proofs of design-optimality properties like design consistency and design unbiasedness do make assumptions about properties like the behavior of finite population moments). However, design-based asymptotic theory is generally based on sequences of populations and sampling designs where both the population size and sample size approach infinity. For SAE, when data are limited, the asymptotic behavior of an estimator under such a design-based paradigm may be less relevant. Our goal is to outline an approach which bridges the gap between fully model-based methods and design-based estimators like the GREG estimator, which typically are guaranteed to be design-consistent. The model-based approach facilitates asymptotic analysis based on assumed superpopulation models. By making assumptions about population structure via some assumed data-generating model, it may be possible to develop model-optimal estimators. Nevertheless, for government officials and survey practitioners, optimality in a design-based sense may be a precondition for adoption of small area estimates, especially in critical health and demographic contexts. The combined model- and design-based perspective can be viewed as attempting to bridge the gap between fully model-based or design-based approaches and has been used to motivate estimators derived from statistical models that are potentially robust to informative sampling and clustering.

To further illustrate these three approaches, we consider the following example of the Horvitz-Thompson estimator of a subpopulation mean (Horvitz and Thompson 1952):

Example 3. (Horvitz-Thompson) The Horvitz-Thompson estimator of a small area mean is defined by:

$$\widehat{\overline{Y}}_i^{HT} = \frac{1}{N(i)} \sum_{j \in S(i)} w_{ij} y_{ij} \tag{2.2}$$

2.3.1 *Design-based inference*

As outlined in Definition 1, conditional on relevant design variables $\mathbf{Z}$, a sampling design may be viewed as a probability distribution (or randomization distribution) over the space of all possible samples of the finite population. Consider that a sample-based estimator of a numeric mean $\widehat{\overline{Y}}_i$ can be viewed as a function $\widehat{\overline{Y}}_i : \mathcal{S} \to \mathbb{R}$. In general, the design-based expectation of a random variable $\widehat{\mu}$ with respect to the sampling distribution can be defined as follows, treating $\widehat{\mu}$ as a real valued function on $\mathcal{S}$.

Definition 2. (Design-based expectation)

$$E_D(\widehat{\mu}) = \sum_{s \in \mathcal{S}} \mathbb{P}_D(s)\widehat{\mu}(s) \tag{2.3}$$

The design-based variance can thus be similarly defined:

Definition 3. (Design-based variance)

$$\mathrm{Var}_D(\widehat{\mu}) = E_D[(\widehat{\mu} - E_D(\widehat{\mu}))^2] \tag{2.4}$$

As is common in the survey statistics literature, we define survey asymptotics in terms of sequences of nested samples and populations, both increasing in size (Breidt and Opsomer 2017; Särndal et al. 2003). Let $U_\infty = 1, 2, \ldots$ be an infinite sequence of elements with associated y values $y_1, y_2, \ldots$ and $U_1, U_2, \ldots$ be a sequence of populations where U_ν contains the first N_ν elements of U_∞ and $U_1 \subset U_2 \subset \cdots$. For each U_ν, conditional on design variables $\mathbf{Z}^{(\nu)}$, let $\mathbb{P}_{D,\nu}(\cdot)$ be a sampling design that assigns probabilities to each possible sample S. We let $E_{D,\nu}$ denote the design-based expectation for estimators based on $\mathbb{P}_{D,\nu}$. Assume sample size n_ν is fixed and $n_1 < n_2 < \cdots$. Thus $\nu \to \infty$ implies $n_\nu \to \infty$ and $N_\nu \to \infty$. Let μ_ν be a function of the elements of U_ν and let $\widehat{\mu}_\nu$ be an estimator of μ_ν based on the sample S.

Given this framework, we can define notions of asymptotic design-based unbiasedness and consistency:

Definition 4. An estimator $\widehat{\mu}_\nu$ is asymptotically design unbiased for μ_ν if

$$\lim_{\nu\to\infty}[E_{D,\nu}(\widehat{\mu}_\nu) - \mu_\nu] = 0 \tag{2.5}$$

Definition 5. Moreover, $\widehat{\mu}_\nu$ is design-consistent if for any fixed $\epsilon > 0$,

$$\lim_{\nu\to\infty}\mathbb{P}_{D,\nu}(|\widehat{\mu}_\nu - \mu_\nu| > \epsilon) = 0 \tag{2.6}$$

Särndal et al. (2003) note that whether these properties hold for $\widehat{\mu}_\nu$ will generally depend on the specification of the estimators, population values $\{y_j\}$, and designs $\{\mathbb{P}_{D,\nu}\}$. In particular, conditions on the limiting behavior of the finite population values and inclusion probabilities are typically needed to ensure consistency of estimators.

Example 3 (continued)**.** Proof that the HT estimator is asymptotically design-unbiased and design-consistent can be found in multiple places, including Breidt and Opsomer (2017). Their proof relies on conditions that are satisfied for many sampling designs, including simple random sampling and stratified random sampling, with and without replacement.

Breidt and Opsomer note that the HT estimator is often assumed to be asymptotically normal, but conditions that ensure asymptotical normality for one design may not hold in general.

2.3.2 Model-based inference

The model-based (also called prediction-based) approach, was proposed by Royall (1970) and further developed by Royall and Herson (1973). Model-based inference proceeds by first specifying a population generating model that holds for both observed and unobserved responses. In practice, one typically proposes a model for the population responses, uses the model to derive an "optimal" estimator for a quantity of interest such as a subpopulation mean or total, and then conducts inference with respect to the assumed model. If the assumed model is misspecified, as may be the case if key aspects

of the design are not incorporated appropriately into the model, the resulting inferences can be incorrect. Valliant, Dorfman, and Royall (2000) review finite population inference using survey data from a model-based perspective.

Rather than basing inference upon a sampling design $\mathbb{P}_D$, we assume that for all j in our population, the response value y_j is drawn from some parametric population generating model $\mathbb{P}_{M,\theta_0}$ dependent on parameters θ_0. Generally inference relies upon the assumption that inclusion is independent of the response after conditioning on covariates included in the model. The model-based expectation of a random variable μ with respect to the population generating model can be defined as follows:

Definition 6. (Model-based expectation)

$$E_{\theta_0}(\widehat{\mu}) = \int \widehat{\mu}(\mathbf{y}_S) d\mathbb{P}_{M,\theta_0}(\mathbf{y}_S) \tag{2.7}$$

where now we treat $\widehat{\mu}$ as a function of the observed response vector $\mathbf{y}_S$. Note that model-based variance, unbiasedness and consistency are not always straightforward to define. In the design-based case, typically the target estimand is some finite population parameter, but under the model-based paradigm, superpopulation parameters or random effects may also be of interest. As such, it is common to discuss mean squared prediction error as opposed to variance:

Definition 7. (Model-based mean-squared prediction error)

$$\mathrm{MSE}_{\theta_0}(\widehat{\mu}) = E_M[(\widehat{\mu} - \mu)^2] \tag{2.8}$$

Under this definition, both $\widehat{\mu}$ and μ may be random under the model-based framework. Note that this is equivalent to the variance of the estimation error when $\widehat{\mu}$ is unbiased for μ. Consistency and unbiasedness can be shown by analyzing the behavior of the estimation error $\widehat{\mu} - \mu$ as the sample size increases.

Proponents of the model-based approach have noted that inference based on an assumed model is standard in other applications of statistical analysis. Moreover, design-based

estimates of uncertainty depend on the sampling design and in effect, on all possible samples that could have been sampled, rather than solely upon the actually observed sample (Royall 1992) The debate between the design-based and model-based paradigms is longstanding and, in some respects, rooted in philosophical disagreements that may be tricky to resolve; Smith (1976, 1994) and Brewer (1999) review some of the history of this discourse and offer some perspectives on reconciling the two perspectives. We adopt a pragmatic stance that acknowledges the benefits of both perspectives; given sufficient data, a design-based approach may be more robust to model misspecification.

2.3.3 Combined model- and design-based inference

Combined model- and design-based inference bridges the gap between inference based on a randomization distribution and inference based on a population generating model. Rubin-Bleuer and Kratina (2005) formalize the ideas of Hartley and Sielken (1975) by establishing a framework for joint model- and design-based inference by defining a product probability space as the product of the sampling design space and the population generating model space. This approach is further explored by Williams and Savitsky (2021) and Han and Wellner (2021), who study the asymptotic behavior of survey sampling estimators under a combined model- and design-based probability measure. Below, we review their approach to inference.

As with the design-based approach, we consider a sequence of sampling designs and populations indexed by ν. with $U_\nu = \{1, \ldots, N_\nu\}$ indexing a finite population of size N_ν, where N_ν increases in ν. We let $\mathcal{S}_\nu$ denote the collection of all possible subsets of U_ν.

We define $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ and $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ to be measurable spaces for the response and auxiliary variables, respectively. Next, we assume that $\{(Y_j, \mathbf{Z}_j) \in \mathcal{Y} \times \mathcal{Z}\}_{j=1}^{N_\nu}$ are independent and identically distributed random vectors drawn from a superpopulation model on the probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P}_{M,\theta_0}) \equiv (\mathcal{Y} \times \mathcal{Z}, \mathcal{B}_\mathcal{Y} \times \mathcal{B}_\mathcal{Z}, \mathbb{P}_{M,\theta_0})$ where $\mathbb{P}_{M,\theta_0}$ denotes the superpopulation measure based on the data generating model and θ_0 denotes a vector of

hyperparameters.

Conditionally on the auxiliary variables $\mathbf{Z}^{(\nu)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_\nu})$, we can define a sampling design $\mathbb{P}_{D,\nu}$ which is a probability distribution over the space of possible samples $S \in \mathcal{S}_\nu$. The sample data are denoted $\{\mathbf{Y}^{(\nu)}, \mathbf{Z}^{(\nu)}, \boldsymbol{\delta}^{(\nu)}\}$ where $\boldsymbol{\delta}^{(\nu)}$ denotes the vector of sample inclusion indicators. As outlined by Han and Wellner (2021), we can construct a measurable space $(\mathcal{S}_\nu \times \mathcal{X}, \sigma(\mathcal{S}_N) \times \mathcal{A}, \mathbb{P})$ where $\sigma(\mathcal{S}_N)$ is the σ-algebra generated by $\mathcal{S}_N$.

We seek to understand the asymptotic behavior of our estimators as $\nu \to 0$ under the combined probability measure $\mathbb{P}$. Similarly, when evaluating estimators, we will generally consider average error metrics taking expectations with respect to both $\mathbb{P}_{M,\theta_0}$ and $\mathbb{P}_{D,\nu}$. Under informative sampling, we consider a combined model-and-design based mean squared error for evaluating point estimators:

Definition 8. (Model- and design-based mean-squared prediction error)

$$\mathrm{MSE}_{\theta_0,\nu}(\widehat{\mu}_i) = \mathbb{E}_{\theta_0,\nu}[(\widehat{\mu}_i - \mu_i)^2] \tag{2.9}$$

where $\mathbb{E}_{\theta_0,\nu}(\cdot) \equiv \mathbb{E}_{D,\nu}[\mathbb{E}_{M,\theta_0}(\cdot)]$.

Definitions of variance, unbiasedness, and consistency are generally similar as to under the model-based paradigm, except that we now take expectations with respect to both the population generating model and the sampling design.

2.4 Existing approaches for small area estimation

In this section, we review SAE approaches with a focus on methods commonly used to estimate health and demographic indicators. Broadly speaking, we consider three kinds of approaches: direct weighted estimators, model-assisted estimators, and model-based estimators.

2.4.1 *Direct estimators*

Direct weighted estimators, such as the Horvitz-Thompson estimator discussed in Example 3, estimate an area-specific quantity using only data from the area in question. Generally, this ensures that direct estimators are design-unbiased and and make few assumptions about population structure ,so they may be preferable when sufficient data are available. We will frequently make use of the Hájek estimator of a mean (Hájek 1971), which extends the HT estimator using survey weights to approximate the totals in Equation (2.1):

Example 4. (Hájek) The Hájek estimator of a small area mean is defined by:

$$\widehat{\overline{Y}}_i^H = \frac{1}{\widehat{N}(i)} \sum_{j \in S(i)} w_{ij} y_{ij} = \frac{1}{\sum_{j \in S(i)} w_{ij}} \sum_{j \in S(i)} w_{ij} y_{ij} \tag{2.10}$$

For many sequences of designs and populations, the numerator and denominator are design consistent for $\sum_{j \in U(i)} y_{ij}$ and $N(i)$ respectively. The estimator $\widehat{\overline{Y}}_i^H$ is simple to compute and may even be preferable to the HT estimator (Särndal et al. 2003). In general, existing software computes estimators of the design variance of $\widehat{\overline{Y}}_i^H$ via linearization or resampling; we use the `survey` package in `R` for computation (Lumley 2004, 2011).

2.4.2 *Model-assisted estimators*

Model-assisted estimators are motivated by working superpopulation models but are specified to ensure design consistency and unbiasedness even when the working model is incorrect. For a review of model-assisted methods see Särndal et al. (2003) and Breidt and Opsomer (2017). A characteristic model-assisted approach is given by the difference estimator:

Example 5. (Difference estimator) A difference estimator of a small area mean is defined by:

$$\widehat{\overline{Y}}_i^{DIF} = \frac{1}{N(i)} \sum_{j \in U(i)} \widehat{y}_{ij} + \sum_{j \in S(i)} w_{ij} (y_{ij} - \widehat{y}_{ij}) \tag{2.11}$$

where $\widehat{y}_{ij}$ represents a working model prediction for unit j in area i.

The difference estimator combines model-based predictions from a working model with a direct estimator of the mean of the residuals in the area based upon the sample. Breidt and Opsomer (2017) show that under appropriate asymptotic conditions, the difference estimator is design consistent. In particular, the direct estimator for the residual mean must be design consistent and the difference between predictions from the working model estimated on sample data and predictions from the working model estimated on the full population must be asymptotically negligible. The generalized regression estimator (GREG) can be framed as an example of a difference estimator (Särndal et al. 2003). Model-assisted estimators of this form are closely related to augmented inverse-probability weighted estimators (Robins et al. 1994) with estimated weights, as noted by Lumley et al. (2011). In the case of binary response data, a logistic regression working model can be used, leading to the logistic generalized regression estimator (LGREG) (Lehtonen and Veijanen 1998) . Model-assisted estimators are typically asymptotically design unbiased and design consistent, but quantification of uncertainty can be difficult. Variance estimators based on linearization approximations generally do not account for uncertainty in the first sum on the right of Equation (2.11), which results from model estimation. The working model should be carefully selected as overfitting can also result in underestimation of uncertainty.

2.4.3 Model-based estimators

When data are limited, direct estimators can become unreliable and model-based methods are used to leverage smoothing and auxiliary information. Area-level models are typically applied to smooth direct estimates, treating them as noisy observations of the true area-specific quantities. If survey microdata are available, unit level models relate individual survey responses to unit level auxiliary covariates and can explicitly account for spatial dependence and between-area variation using random effects.

Both the area level and unit level models used for small area estimation can be classified as hierarchical models that incorporate random intercept parameters. Typically, area-specific random effects are used, but in some cases, additional levels of random effects can be included (Marhuenda et al. 2013). Jiang (2017) and Sugasawa and Kubokawa (2020) review hierarchical models in the context of SAE, while Gelman and Hill (2007) and Hodges (2016) provide more general reviews.

Traditionally, random effects have been used to model sources of variation which affect groups of units and which induce dependence between responses from the same group. From this perspective, the random effects themselves are typically not of interest, but can be used to model the correlation structure of the response data. This interpretation is common in association studies, where random effects can be used to separate group level correlations from unit level associations, which may be of primary interest (Gelman 2006).

In SAE, however, random effects are typically used to represent area level differences not explained by covariates, so the values of the random effects themselves are of interest. Early model-based approaches to SAE such as the Fay-Herriot model (Fay and Herriot 1979) and nested error regression model (Battese et al. 1988) drew upon research on frequentist prediction of random effects for linear mixed models (Harville 1976; Kackar and Harville 1984; Laird and Ware 1982) as well as research on Bayesian linear models with random coefficients (Lindley and Smith 1972; Smith 1973). Both frequentist and Bayesian perspectives on prediction of random effects are reviewed by Robinson (1991) and Jiang (2007). Practically, the mixed modeling approach enables flexible models with parameters for each small area of interest and reduces the risk of overfitting by placing a shrinkage-inducing prior on random effects. Outside of SAE, prediction of individual random effects has been used to assess the effectiveness of schools (Raudenbush and Willms 1995), health authorities (Goldstein and Spiegelhalter 1996), and baseball players (Efron and Morris 1975).

Using hierarchical models imposes additional computational difficulties compared with

fixed effects models, as additional parameters such as variance components must be estimated. Hierarchical models may be studied from both frequentist and Bayesian perspectives, but the Bayesian approach has become popular as software such as Stan (Carpenter et al. 2017) and INLA (Lindgren and Rue 2015) have enabled fast approximate inference for hierarchical models, especially for nonlinear models for which likelihood functions may not be available in closed form.

In the below, we review hierarchical models commonly used for SAE, including both area level and unit level models.

Area level models

The standard Fay-Herriot model combines a sampling model for the direct estimators with a linking model for the true finite population means μ_i (Fay and Herriot 1979). We can specify the models as follows, using $\widehat{\mu}_i$ to denote the direct estimator of the mean for area i:

Example 6. (Fay-Herriot model)

$$\widehat{\mu}_i = \mu_i + \epsilon_i \tag{2.12}$$

$$\mu_i = \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta} + u_i \tag{2.13}$$

where for all i, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ are independently and identically distributed area-specific random effects and $\epsilon_i \sim N(0, V_i)$ are independent sampling errors. We use V_i to denote the sampling variance of $\widehat{\mu}_i$. Finally, $\widetilde{\mathbf{x}}_i$ represents a vector of area-specific covariates and $\boldsymbol{\beta}$ denotes the corresponding vector of coefficients.

The basic Fay-Herriot model can be viewed as a linear mixed model, but unlike most mixed models, it features two random parameters u_i and ϵ_i at the area level, where u_i represents true area level variation in the finite population μ_i parameters and ϵ_i represents sampling error. As such, V_i is assumed to be known to ensure identifiability. In practice,

V_i is estimated using sample data. This basic area level Fay-Herriot model can be used to generate model-based estimates either by taking a frequentist approach and computing the empirical best linear unbiased predictor (EBLUP) or by using a Bayesian approach to compute the posterior distribution of μ_i; for more details see Chapters 6 and 9 of Rao and Molina (2015).

Assuming design consistency of the direct estimator and a sequence of designs and populations such that $V_i \to 0$, the EBLUP is also a design consistent estimator of μ_i. A fully Bayesian approach can also be used and can produce design consistent estimators as well. The basic area level model assumes that area random effects are iid, but this model has been extended to allow for random effects with spatial and spatiotemporal correlation structures (Chung and Datta 2020; Ghosh et al. 1998; Pratesi and Salvati 2008). Mercer et al. (2015) use a Fay-Herriot type model of logit-transformed direct estimators with spatiotemporal random effects to estimate child mortality rates.

The basic area level model assumes that area random effects are independently distributed, but this model has been extended to spatial and spatiotemporal correlation structures (Ghosh et al. 1998; Petrucci and Salvati 2006; Pratesi and Salvati 2008). Chung and Datta (2020) found that a spatial area level model can improve estimation when there is spatial structure in the direct estimators not explained by observed covariates, as may be the case if covariate data are limited. Mercer et al. (2015) use a Fay-Herriot type model with spatiotemporal random effects to estimate child mortality rates. Alternatively, Porter et al. (2014) extend the Fay-Herriot model to include functional covariates based on readily available sources such as satellite imagery.

Unit level models

When more detailed covariate information is available, modeling unit level responses can aid SAE. For continuous responses, the nested error regression model, also called the basic unit level model, was proposed by Battese, Harter, and Fuller (1988):

Example 7. (Battese-Harter-Fuller nested error regression model)

$$y_{ij} = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i + \varepsilon_{ij} \tag{2.14}$$

Above, β_0 denotes an intercept, $\mathbf{x}_{ij}$ denotes covariate values for individual j, and $\boldsymbol{\beta}_1$ denotes the corresponding coefficients. The area level effects are denoted $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$. Finally, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ represents random and independent measurement error.

Under this model, $\overline{Y}_i = \beta_0 + \overline{\mathbf{x}}_i^T \boldsymbol{\beta}_1 + u_i + \overline{\varepsilon}_i$ where $\overline{\mathbf{x}}_i$ and $\overline{\varepsilon}_i$ denotes the area means of $\mathbf{x}_{ij}$ and ε_{ij}, respectively. By the law of large numbers, $\overline{\varepsilon}_i$ converges in probability to $E(\varepsilon_{ij}) = 0$ as $N(i) \to \infty$, so instead of estimating $\overline{Y}_i$, it is standard in the SAE literature to focus on estimation of

$$\mu_i = E(\overline{Y}_i \mid \overline{\mathbf{x}}_i, u_i) = \beta_0 + \overline{\mathbf{x}}_i^T \boldsymbol{\beta}_1 + u_i.$$

The nested error regression model is a linear mixed model with one level of area specific random effects and is closely related to other commonly studied linear mixed models. If covariates are omitted, the nested error regression model is equivalent to a one-way random effects model or one-way analysis of variance (ANOVA) model (Faraway 2014). If we expand the nested error regression model to allow random area specific coefficients for the covariates, it becomes equivalent to the two-level model (Goldstein 2010) or two-stage models for repeated measurements (Laird and Ware 1982). The above model uses one level of random effects for each area, but for multistage designs, unit level models could also include random effects for each stage of sampling, as suggested by Marhuenda et al. (2013). Corral et al. (2021) review a number of unit level models used for poverty mapping.

For a binary response, a logistic unit level model can be specified:

Example 8. (Logistic nested error regression model)

$$P(y_{ij} = 1 \mid \mathbf{x}_{ij}, \beta_0, \boldsymbol{\beta}_1, u_i) = q_{ij} \tag{2.15}$$

$$\text{logit}(q_{ij}) = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i \tag{2.16}$$

Above β_0 denotes an intercept, $\mathbf{x}_{ij}$ denotes covariate values for individual j, and $\boldsymbol{\beta}_1$ denotes the corresponding coefficients. Again, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ represents an area level random effect.

Other likelihoods can be used for other types of response data. Such models can be viewed as generalized linear mixed models (GLMM) (Jiang 2007). Skrondal and Rabe-Hesketh (2009) review prediction for GLMM, which is more challenging than in the linear mixed model case as closed form expressions generally do not exist for prediction of random effects.

As discusssed in Chapters 4 and 5, incorporating sampling weights for unit level data when estimating mixed model parameters requires careful consideration of weight scaling and the dependence structure of the response data.

Model-based geostatistics

Approaches developed in the SAE literature usually acknowledge the importance of design-based optimality conditions and use simpler models for which design-based analysis is tractable. Model-based geostatistical approaches tend to utilize more complex models, leveraging recent advances in spatial modeling and computation for hierarchical models. Recent research has produced maps of health indicators at resolutions as fine as 1 km by 1 km using geostatistical models (Diggle and Giorgi 2016; Utazi et al. 2020). When creating such high-resolution maps, it is common to model the risk of experiencing an outcome such as vaccination or disease using a continuous spatial process. These continuous risk surfaces are often modeled using Gaussian processes. When the number of prediction locations is high, using Gaussian process models can be time-consuming, but approximate methods can speed up computation. The integrated nested Laplace approximation-stochastic partial differential equation (INLA-SPDE) approach is popular

for approximate Bayesian inference with spatial and spatiotemporal Gaussian process models (Lindgren and Rue 2015; Rue et al. 2017). Although the continuous spatial modeling approach allows for prediction at any location, interpretation of the continuous surface is complicated: the surface is assumed to exist even at locations where no individual is present. For this reason, it may be preferable to model actual prevalences among groups of individuals in the finite population.

Such models can also include unstructured cluster level random effects to account for clustering; however, without complete census frame information, it is not obvious how to aggregate cluster effects when generating predictions. Furthermore, such models typically do not explicitly account for urban/rural stratification when using data from the DHS and other surveys. Paige et al. (2022a) and Dong and Wakefield (2021) have shown that urbanicity can be associated with health outcomes, leading to bias if the stratification is not incorporated into the model. An additional complication results from changing levels of urbanization over time. From a design-based perspective, it is important to account for the urban/rural stratification used at the time of sampling as a unit's inclusion probability depends on its sampling stratum. However, in many LMICs, increasing urbanization means that clusters in rural strata may change over time to resemble urban clusters more closely. From a model-based perspective, it may make sense to treat these clusters as urban if their exposure to potential outcomes such as disease or vaccination are affected by urbanization. Geostatistical models often incorporate covariates like intensity of night time lights or population density that could be viewed as surrogates for urbanicity, but such covariates are only surrogate and will not align with the original partition used to define sampling strata.

Parameter estimation

For mixed effects models such as the nested error regression model, a number of approaches have been proposed for estimating model parameters and generating point and

interval estimates of small area means μ_i, including frequentist empirical best linear unbiased prediction (EBLUP), empirical Bayesian (EB) and hierarchical or fully Bayesian (HB) strategies, which are reviewed by Rao and Molina (2015). These strategies lead to estimators that can be studied under the model-based or combined model- and design-based frameworks outlined above. Below we briefly discuss these three approaches, assuming that the sampling design is ignorable.

- **Empirical best linear unbiased prediction**: For linear mixed effects models, when variance components are known, best linear unbiased predictors (BLUP) for μ_i minimize the mean squared error (among linear unbiased predictors). Empirical BLUP $\widehat{\mu}_i^{EBLUP}$ can be obtained by estimating the variance components using sample data, for example via maximum likelihood (ML), restricted maximum likelihood (REML), or Henderson's method of "fitting constants," and then plugging in the estimated variance parameters into the BLUP formula. Prediction intervals can be constructed by studying the asymptotic behavior of the prediction error $\widehat{\mu}_i^{EBLUP} - \mu_i$.

- **Empirical Bayes**: For generalized linear mixed models with non-Gaussian likelihoods, the BLUP may not be available in closed form. The empirical Bayes approach proceeds by computing the posterior density for μ_i given response values $\mathbf{y}$ and model parameters θ. The model parameters θ are then estimated from the sample data, (for example, via maximum likelihood) and then the estimated posterior density is used to generate sample predictions of μ_i. Generally estimates are not averaged over a distribution for θ and priors are not explicitly required, so this approach is frequentist in nature and estimators are evaluated with respect to their frequentist properties.

- **Hierarchical Bayes**: A fully Bayesian or hierarchical Bayesian approach to inference can be adopted by assuming priors on the model parameters θ and then computing or sampling from the posterior $p(\mu_i \mid \mathbf{y})$, integrating over uncertainty in θ. Prediction

intervals can be obtained by computing quantiles of the posterior $p(\mu_i \mid \mathbf{y})$. The fully Bayesian approach automatically quantifies uncertainty via the posterior distribution instead of requiring estimation of mean squared predictive error. However, it also requires the use of prior distributions for θ. Priors for θ require careful specification and may be viewed negatively by analysts who view informative priors as placing undesirable assumptions on the model space.

Informative sampling

When using unit level models for SAE, it is crucial to account for features of the survey design including unequal sampling probabilities, stratification, and clustering. Model-based approaches often treat sampling design as ignorable, assuming that the distribution of responses in the sample is identical to that of the population. Under such an assumption, a model for sampled responses can be used directly to make inferences about the population. However, if the sampling design is not ignorable, one may need to account for potential differences between sampled and non-sampled units. One possible solution is to incorporate design variables, such as variables defining sampling strata or clusters, as model predictors. Ideally, after conditioning on all relevant design variables, the responses will be independent of sample inclusion indicators. If we can identify such a model, we can say that sampling is uninformative with respect to the model.

In practice, however, we may only observe a subset of design variables or the functional form of the relationship between the design variables and responses may be unknown, making it difficult to specify a model for which sampling is uninformative. We will say that sampling is informative with respect to the model if the model does not apply to both sampled and non-sampled units. In such a setting, we can address the effects of informative sampling using other sources of information about the survey design, such as sampling weights.

While estimators based on basic area level models are often design consistent, unit level

models may not generally produce design consistent estimators. Model-based estimators may be biased if the design is not ignorable with respect to the model, as is often the case under informative sampling or unequal inclusion probabilities. Below, we review existing strategies used to address the design when using unit level models; see Parker et al. (2020) for another overview.

Pseudo-likelihood methods, as introduced by Binder (1983) and Skinner (1989), incorporate survey weights into model estimation. This approach is commonly used for estimation of linear and generalized linear models, as reviewed by Lumley and Scott (2017) and has been extended to estimation for both linear mixed models (Pfeffermann et al. 1998) and generalized linear mixed models (Asparouhov 2006; Rabe-Hesketh and Skrondal 2006) with multiple levels of random effects for multistage sampling designs. Using weights with multilevel models is often complicated by the need for weights corresponding to each level of random effects. For a model with cluster level effects using the pseudo-likelihoods proposed by Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006), separate weights are needed to account for cluster level effects and unit level effects, but many surveys only provide one set of sampling weights corresponding to the final inclusion probabilities of each unit. The DHS data we use only provides scaled final design weights. Even if sampling weights are available at all stages, however, using unscaled weights can lead to bias in estimation of variance parameters and subsequently small area means (Asparouhov 2006; Korn and Graubard 2003; Pfeffermann et al. 1998). As a result, such pseudo-likelihood methods can depend on approximating and rescaling sampling weights. Savitsky and Williams (2022) outline a pseudo-Bayesian approach that achieves approximately unbiased estimation for mixed model parameters under multistage sampling by proposing a modified pseudo-likelihood and Slud (2020) suggests a expectation maximation algorithm based upon the same pseudo-likelihood.

If stage-specific sampling weights are unavailable but joint sampling weights are available or can be estimated, pairwise likelihood approaches can be used for fitting mixed

models (Rao et al. 2013; Yi et al. 2016); Huang (2019)]. However, such methods do not automatically generate predictions for random effects. Alternatively, Pfeffermann and Sverchkov (2007) directly model the sampling weights in order to account for informative sampling.

All of these methods acknowledge the design, but may be sensitive to scaling of weights or rely on availability of higher-order or pairwise sampling weights. Unbiased estimation of unit level models with random effects can thus depend on correct model specification or ad hoc weight rescaling. Resulting small area mean estimators may accordingly be affected by bias or incorrect variance estimation.

Chapter 3

A VARIANCE SMOOTHING AREAL MODEL FOR ESTIMATING PROPORTIONS

3.1 Introduction

Area level models for small area estimation treat direct weighted estimates as noisy observations, using area level covariates and random effects to improve upon the precision of the direct estimators. As outlined in Section 2.4.3, the Fay-Herriot area level model (Fay and Herriot 1979) assumes that for each area, the direct estimator is available and can be modeled using the Gaussian distribution centered around the true parameter of interest. Such models can be directly applied for estimating proportions. For ease of exposition, we use $\widehat{p}_i$ to denote the direct estimator and p_i to denote the true rate of interest:

$$\widehat{p}_i = p_i + \epsilon_i \tag{3.1}$$

where $\epsilon_i \sim N(0, V_i)$ are independent sampling errors, with V_i denoting sampling variances typically assumed to be known. This provides a model for the sampling variability of $\widehat{p}_i$. The Fay-Herriot model combines this sampling model with a linking model for the parameters of interest p_i:

$$p_i = \widetilde{\mathbf{x}}_i^T \boldsymbol{\beta} + u_i \tag{3.2}$$

where for all i, $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ are independently and identically distributed area-specific random effects. Finally, $\widetilde{\mathbf{x}}_i$ represents a vector of area-specific covariates and $\boldsymbol{\beta}$ denotes the corresponding vector of coefficients. The linking model variance σ_u^2 controls the magnitude of deviations from the mean model. Since the direct estimators $\widehat{p}_i$ account for the survey design via the use of survey weights, area level methods are less sensitive to the

effects of informative sampling and other design features than unit level methods. Under certain regularity conditions, the resulting estimators are design consistent; for a review, see Rao and Molina (2015). In practice, the variances V_i are usually estimated using sample-based estimators $\widehat{V}_i$, but the standard Fay-Herriot model does not account for uncertainty in $\widehat{V}_i$. This is a well known problem in the SAE literature (Arora and Lahiri 1997; Bell 2008; Rivest and Vandal 2002) and has motivated a number of proposed extensions of Fay-Herriot that incorporate variance modeling. However, existing approaches are not well-suited for estimation of rates of key health outcomes such as vaccination in a low- and middle- income countries (LMIC) context. In particular, existing methods often rely on the availability of informative area level covariates for modeling sampling variances and such auxiliary information may not be available if census data is limited or unreliable. Additionally, approaches that model the sampling variances often do not account for uncertainty in the modeled variance estimates, simply treating them as known in the standard Fay-Herriot model. Finally, although the Gaussian approximation for the direct estimator's sampling distribution may be effective when estimating means of continuous-valued responses, it may be less appropriate when the target estimand is a proportion of binary responses. For estimating a proportion, the direct weighted estimator is typically a weighted sum of binary valued random variables, so its mean and sampling variance can be strongly related. Correctly modeling this mean-variance relationship is critical as it may explain a substantial part of the heterogeneity in sampling variances for a set of estimators.

To address these issues, we propose a fully Bayesian area level model for small area proportions that jointly models the direct estimators and sampling variance estimators. We use spatially structured area level random effects to induce spatial smoothing of both means and variances. In simulations, we find that our proposed method produces interval estimates with improved empirical coverage rates compared with those produced based on the standard Fay-Herriot approach. Below we expand upon our proposed

model and review variance-smoothing area level models for small area estimation. Section 3.2 reviews existing area level models and discusses recent efforts to incorporate variance smoothing for small area estimation. In Section 3.3, we outline our spatial variance-smoothing area level model for estimation of small area proportions. We compare our approach with other area level methods via simulation in Section 3.4 and by application to data from the Demographic and Health Surveys (DHS) in Section 3.5. Finally, we compare our method with existing approaches and discuss potential directions for future research in Section 3.6.

To address these issues, we propose a fully Bayesian area level model for small area proportions that jointly models the direct estimators and sampling variance estimators (Gao and Wakefield 2022). We use spatially structured area level random effects to induce spatial smoothing of both means and variances. In simulations, we find that our proposed method produces interval estimates with improved empirical coverage rates compared with those produced based on the standard Fay-Herriot approach.

We consider two motivating examples of estimating subnational demographic rates using data from the DHS Program. In the first, we consider the vaccination coverage rates example outlined in Section 1.2 using 2018 Nigeria DHS data to estimate regional vaccination coverage rates for the first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months (National Population Commission - NPC/Nigeria and ICF. 2019). In the second, we use 2015–16 Malawi DHS data to estimate HIV prevalence among women aged 15–49 (NSO/Malawi and ICF 2017). Figure 3.1 provides maps of direct survey-weighted estimators for both indicators. The measles vaccination example represents an estimation problem where the estimated area level proportions have a large spread and are generally located away from zero or one; in the HIV prevalence example, the direct estimates exhibit less variability and are on average closer to zero. In Nigeria and Malawi, the DHS Program uses a stratified two-stage cluster sampling design. Countries are divided into administrative regions which are further partitioned into urban and rural ar-

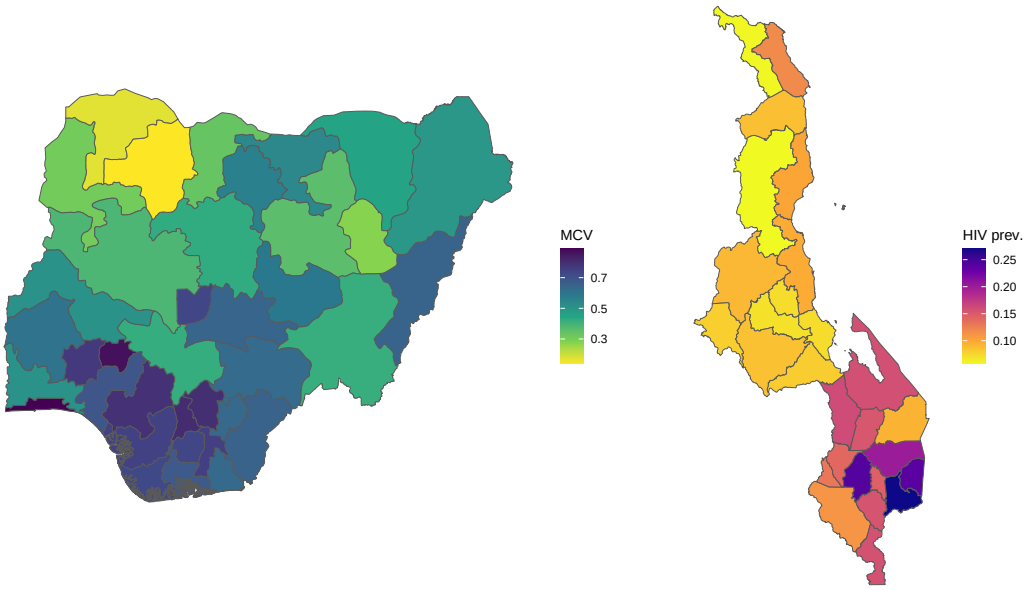

Figure 3.1: Direct weighted estimates of vaccination coverage rate for first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria, 2018 (left) and HIV prevalence rate for women aged 15–49 in Malawi, 2015–2016 (right).

eas. The sampling strata are defined by crossing these regions with urban/rural status. In Nigeria, the divisions used for defining strata are called Admin-1 regions; in Malawi, they are called Admin-2 regions. Each stratum is divided into collections of households called enumeration areas (EAs) or clusters. The first stage of sampling selects a pre-specified number of EAs in each stratum with probability proportional to the number of households in the EA. The second stage of sampling selects a fixed number of households in each sampled EA.

The 2015-16 Malawi DHS used voluntary finger prick blood sampling to collect data on HIV prevalence. We desire estimates of HIV prevalence for each of Malawi's 28 districts, also referred to as Admin-2 areas. For this survey, the sampling frame was obtained from a 2008 census which identified 12,558 EAs distributed between 56 strata. Ultimately, data were collected from 827 EAs, from which a total of 8,497 women aged 15–49 were eligible for HIV testing. Ultimately, 93% of eligible women were tested, but the HIV test results were anonymized, with volunteers not informed of their results and instead receiving access to educational materials and free counseling and testing (NSO/Malawi and ICF 2017). For additional background on the Nigeria example, see Section 1.2.s

For both Nigeria and Malawi, the DHS provides GPS coordinates for nearly all EAs, but the locations have been adjusted to maintain privacy by adding small distances at random. Figure 3.2 provides maps of the small area boundaries and sampled EA locations in Malawi. Since the island region of Likoma is disconnected from the mainland and has a very small population, we omit its data from our analysis.

3.2 Existing approaches

Since the standard Fay-Herriot model treats V_i as known, a number of extensions and alternative approaches have been proposed to relax the assumption of known sampling variances. In this section, we review area level Fay-Herriot type models that account for unknown sampling variances with both continuous and binary response data. We also

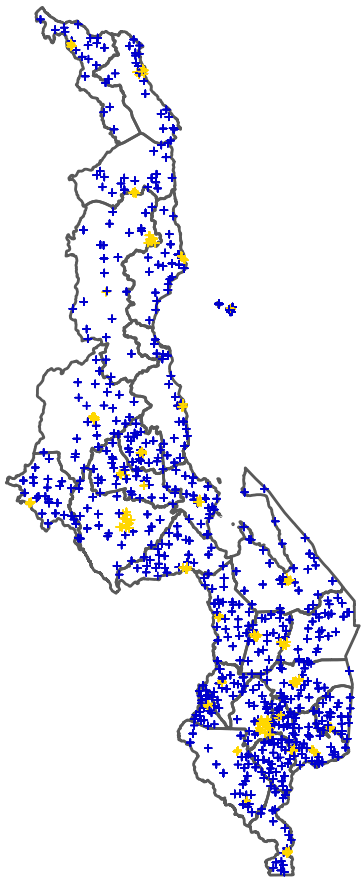

Figure 3.2: Small area boundaries and sampled enumeration area locations for 2015–16 Malawi DHS.

discuss other extensions to the Fay-Herriot model that influence our proposed model.

3.2.1 Variance Smoothing for Continuous Response

The basic Fay-Herriot model has been extended to account for unknown V_i in a number of ways for continuous response data. Some research has focused on adjusting estimates of mean squared error for Fay-Herriot estimators to account for uncertainty in sampling variances (Kleffe and Rao 1992; Rivest and Vandal 2002; Wang and Fuller 2003). The estimated sampling variances $\widehat{V}_i$ are typically assumed to be independent of the direct estimators $\widehat{p}_i$.

Building upon this research, other papers have proposed to incorporate modeling of sampling variances into the Fay-Herriot model. The most common strategy is to combine the Fay-Herriot model with a model for the sample-based variance estimators $\widehat{V}_i$. As an example, You and Chapman (2006) assume the following sampling model for the variance estimators $\widehat{V}_i$:

$$\widehat{V}_i \mid V_i \sim \frac{V_i}{d_i}\chi^2_{d_i} \tag{3.3}$$

where d_i denotes the degrees of freedom for area i. In addition, $\widehat{V}_i$ are assumed to be independent of the mean estimators p_i. If the response values for area i were independently and identically distributed Gaussian random variables, the above model (3.3) would hold for the variance estimator $\widehat{V}_i = s_i^2/n_i$ with $d_i = n_i - 1$, where n_i denotes the sample size for area i. When responses are sampled at random with replacement within areas, such an assumption may be appropriate, but for complicated sampling schemes, different values of d_i or even alternative models may be necessary. You and Chapman (2006) adopted a hierarchical Bayesian approach and placed inverse Gamma priors on the variance parameters $\sigma_u^2 \sim IG(r_0, s_0)$ and $V_i \sim IG(r_i, s_i)$, with r_i, s_i chosen to be small for all areas $i = 1, \ldots m$. Notably, they allow the prior for V_i to vary across areas, which makes the sampling variances V_i independent across areas.

The You and Chapman model builds upon a similar model proposed by Arora and Lahiri (1997) and is representative of similar approaches using scaled-chi squared distributions to model the sampling variance estimators $\widehat{V}_i$. These approaches typically differ in the priors placed on the true sampling variances V_i or choices of degrees of freedom d_i. Bell (2008) reviews and compares a number of sampling variance modeling approaches.

Maiti et al. (2014) assume the same variance sampling model (3.3), but adopt an empirical Bayes approach, setting the prior $\sigma_i^2 \sim IG(r, s)$ and estimating $\{r, s, \sigma_u^2\}$ via maximum likelihood. In addition to modeling $\widehat{V}_i$, Hwang et al. (2009) and Dass et al. (2012) noted that assuming a common prior for V_i for all areas i could induce shrinkage in the resulting variance estimates and produce improved interval estimates for the parameters of interest. In this vein, Sugasawa et al. (2017) explore different priors for the sampling variances V_i, adopting a fully Bayesian approach to estimation. Alternatively, Polettini (2017) induces shrinkage for the sampling variance estimates using a semiparametric Dirichlet process model with random variances.

3.2.2 Variance Smoothing for Binary Response

When response values are binary and the target of estimation is a small area proportion, it may be helpful to account for the mean-variance relationship observed in binary response data. Generalized variance functions (GVFs), which model the functional relationship between the expectation and variance of a survey estimator, can be used as an alternative to linearization-based approximations or resampling methods for estimating V_i. If the model used is appropriate, the resulting modeled variance estimates could improve upon the direct variance estimates in terms of precision. An introduction to GVFs is provided in Chapter 7 of Wolter (2007).

For small area estimation of proportions, several GVF-like approaches to variance estimation have been previously proposed based on treating the responses like binomial data.

Liu et al. (2014) assume the following model for V_i:

$$V_i = \frac{p_i(1-p_i)}{n(i)} DEFF_i \tag{3.4}$$

where $DEFF_i$ denotes the design effect, defined as the ratio of the variance of p_i under the implemented survey design to the the variance of p_i under simple random sampling. As described in their paper, Liu et al. estimate design effects using available information on sample sizes and survey weights and treat them as known. Model-based estimates of V_i can be produced by replacing the unknown p_i values above with their direct estimators. Hawala and Lahiri (2018) propose a similar GVF for count data. Maples (2016) similarly proposes a GVF for producing variance estimates based on estimating the design effect using additional information about any unequal weighting or clustering in the sampling procedure. Franco and Bell (2013) adopt a different strategy using a GVF to compute an effective sample size for each area of interest, which they use to fit a binomial model.

Mohadjer et al. (2012) similarly use a GVF to produce variance estimates for use in an area level model, assuming the following model

$$\log(V_i/p_i^2) = \gamma_0 + \gamma_1 \log(\widetilde{p}_i) + \gamma_2 \log(1-\widetilde{p}_i) + \gamma_3 \log(n(i)) + \varepsilon_i \tag{3.5}$$

where $\epsilon(i) \sim N(0, \sigma_\epsilon^2)$ and $\widetilde{p}_i$ denotes a predictor of p_i based on a model dependent solely on auxiliary covariate information and not explicitly on any direct estimates.

The GVF approaches described thus far treat the resulting variance estimates as known, so the resulting Fay-Herriot estimates do not account for uncertainty in the variance model. Maples et al. (Maples et al. 2009) address this by combining a GVF with a sampling model for the direct variance estimates. In particular, they assume Model (3.3) holds for the direct variance estimates $\widehat{V}_i$ and then propose the following linking model for V_i:

$$V_i \mid \alpha, \boldsymbol{\gamma} \sim IG(\alpha+1, \alpha \exp(\mathbf{z}_i^T \boldsymbol{\gamma})) \tag{3.6}$$

where α controls the precision of the variance linking model, $\mathbf{z}_i$ are area level covariates and $\boldsymbol{\gamma}$ are corresponding coefficients estimated using an empirical Bayes approach.

Maples et al. outline a procedure for using bootstrap sampling to estimate effective sample size for each area, which informs their choice for the degrees of freedom d_i in the variance sampling model. They show that this model produces smoothed variance estimates that could help to correct underestimation by the direct variance estimators.

As an alternative to modeling heterogeneity in the sampling variances, Hirose et al. (2023) propose a variance-stabilizing transformation of the direct estimates $\widehat{p}_i$. This approach is similarly based on assuming a mean-variance relationship and designing a particular transformation to remove the effect of the mean-variance relationship. However, Hirose et al. (2023) apply the same variance-stabilizing transformation for all areas, which prevents any variability in the mean-variance relationship and which could be misspecified. We propose an approach more similar to that of Mohadjer et al. (2012), which allows for some variability in the mean-variance relationship through the estimated γ parameters and random ε terms.

3.2.3 Alternative sampling and linking models

The linking and sampling mean models in the Fay-Herriot approach assume responses are continuous, but since p_i are bounded between 0 and 1, it may be inappropriate to treat $\widehat{p}_i$ as Gaussian, especially when p_i is close to 0 or 1 and when V_i is large. In the health and demography setting, Mercer et al. (2015) apply a logit transformation to direct estimates of mortality rates before fitting a Fay-Herriot-type model. More generally, You and Rao (2002) and Sugasawa et al. (2018) propose unmatched sampling and linking models, combining the sampling model given by Equation (3.1) with an alternative linking model that transforms the finite population parameters of interest p_i to make a Gaussian approximation more appropriate. As an example, Liu et al. (2014) considered the following logit-normal linking model:

$$\text{logit}(p_i) \mid \boldsymbol{\beta}, \sigma_u^2 \sim N(\widetilde{\mathbf{x}}_i^T \boldsymbol{\beta}, \sigma_u^2) \tag{3.7}$$

Mohadjer et al. (2012) apply this model to estimation of adult literacy rates. Liu et al. also consider alternative models including a beta-logistic model combining a beta sampling model with the above logistic linking model, which accounts for the limited range of $\widehat{p}_i$ but will not reflect its true sampling distribution. Franco and Bell (2013) and Chen et al. (2014) consider binomial sampling models, treating observed area level counts as being drawn from a binomial distribution with size parameter given by some measure of effective sample size. As an alternative to unmatched sampling and linking models, Mercer et al. (2015) describe an approach that uses Gaussian sampling and linking models to model $\text{logit}(\widehat{p}_i)$ and $\text{logit}(p_i)$.

3.3 A joint mean- and variance-smoothing model

We assume that for all $i = 1, \ldots, m$, we have direct estimates of area level proportions $\widehat{p}_i$ and corresponding variance estimates $\widehat{V}_i$. We propose a Bayesian joint model for the full data $(\widehat{\mathbf{p}}, \widehat{\mathbf{V}})$ that induces spatial smoothing for both the proportion and variance estimates. Our approach uses two sets of unmatched models, one for the estimated proportions $\widehat{\mathbf{p}}$ and one for the variance estimates $\widehat{\mathbf{V}}$, with these models being linked through the use of a generalized variance function. We use a spatial linking model for the proportions that induces spatial smoothing for both the proportions and the estimated variances.

3.3.1 Mean model

For modeling the direct estimates $\widehat{p}_i$, we use unmatched sampling and linking models, combining a Gaussian sampling model with a spatial logit-Gaussian linking model:

$$\widehat{p}_i \mid p_i, V_i \overset{ind}{\sim} N(p_i, V_i). \tag{3.8}$$

$$\text{logit}(\mathbf{p}) \mid \boldsymbol{\beta}, \sigma_u^2, \phi \sim \boldsymbol{\mathcal{N}}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{BYM2}(\sigma_u^2, \phi)). \tag{3.9}$$

In the above, p_i denotes the finite population area-specific proportion and V_i denotes the sampling variance of the direct estimator $\widehat{p}_i$. We use the shorthand $\text{logit}(\mathbf{p})$ to denote the

vector $(\text{logit}(p_1), \ldots, \text{logit}(p_i))^T$, which we assume is drawn from a multivariate Gaussian distribution with mean $\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $m \times (p+1)$ design matrix containing covariate information and $\boldsymbol{\beta}$ is a $(p+1)$-vector containing the intercept and corresponding coefficients. Finally $\boldsymbol{\Sigma}_{BYM2}(\sigma_u^2, \phi)$ denotes a spatial covariance matrix dependent on marginal variance parameter σ_u^2 and spatial correlation parameter ϕ. We use the BYM2 model, a reparametrization of the Besag-York-Mollié (Besag et al. 1991) model proposed by Riebler et al. (2016) which determines the structure of $\boldsymbol{\Sigma}_{BYM2}(\sigma_u^2, \phi)$. Below, we review the BYM2 model, rewriting the mean linking model as follows for clarity:

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3.10}$$

$$\mathbf{u} \mid \boldsymbol{\sigma}_u^2, \phi \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma}_{BYM2}(\sigma_u^2, \phi)). \tag{3.11}$$

Under the BYM2 model, we assume $\mathbf{u}$ can be partitioned into an unstructured component $\widetilde{\mathbf{u}}_1$ and a structured spatial component $\widetilde{\mathbf{u}}_{2*}$:

$$\mathbf{u} = \sigma_u \left(\sqrt{1-\phi}\widetilde{\mathbf{u}}_1 + \sqrt{\phi}\widetilde{\mathbf{u}}_{2*} \right) \tag{3.12}$$

We assume $\widetilde{\mathbf{u}}_1 \sim N(\mathbf{0}, \mathbf{I})$ is a vector of iid Gaussian random area effects and assume an intrinsic conditional autoregressive (ICAR) Gaussian prior for $\widetilde{\mathbf{u}}_{2*}$. The ICAR prior, as proposed by Besag et al. (1991) (Besag et al. 1991), assumes that spatial components $\widetilde{u}_{2i*}$ and $\widetilde{u}_{2j*}$, representing the values of $\widetilde{\mathbf{u}}_{2*}$ for areas i and j, are correlated if areas i and j are defined to be neighbors. Under an ICAR prior, we assume that for a particular area i, the mean of $\widetilde{u}_{2i*}$ is equal to the mean of all neighboring effects and the precision of $\widetilde{u}_{2i*}$ is proportional to the number of neighbors. Using this parameterization, σ_u^2 denotes the marginal variance of $\mathbf{u}$ and ϕ represents the proportion of variation assigned to the spatial component.

Under a BYM2 model, $\mathbf{u}$ has the covariance matrix

$$\boldsymbol{\Sigma}_{BYM2}(\sigma_u^2, \phi) = \sigma_u^2((1-\phi)\mathbf{I} + \phi\mathbf{Q}_*^-) \tag{3.13}$$

Here, $\mathbf{Q}_*$ denotes the precision matrix of $\widetilde{\mathbf{u}}_{2*}$ and $\mathbf{Q}_*^-$ is its generalized inverse. Note that the precision matrix implied by the ICAR prior, $\mathbf{Q}_*$, is singular, yielding an improper prior. To ensure identifiability, we must place a sum-to-zero constraint on $\mathbf{u}$. In order to make the marginal variance parameter σ_u interpretable, we scale $\mathbf{Q}_*$ to make the geometric mean of the marginal variances equal to one, as recommended by Riebler et al. (2016)

3.3.2 Variance model

We similarly use unmatched models for the corresponding variance estimates $\widehat{\mathbf{V}}$, using a chi-squared sampling model with a log-normal linking model. We use the chi-squared sampling model described in Equation (3.3), assuming that for all i, the variance estimate $\widehat{V}_i$ is an unbiased estimator of V_i. The linking model assumes the true $\log(V_i)$ values are Gaussian distributed with expected values given by a generalized variance function $f(p_i, \mathbf{z}_i; \boldsymbol{\gamma})$ whose inputs are the area proportion p_i, other area level predictors $\mathbf{z}_i$, and parameters $\boldsymbol{\gamma}$. We can write down the unmatched models as follows:

$$\frac{d_i \widehat{V}_i}{V_i} \mid d_i, V_i \overset{ind}{\sim} \chi^2_{d_i} \tag{3.14}$$

$$\log(V_i) \mid p_i, \mathbf{z}_i, \boldsymbol{\gamma}, \sigma^2_\tau \overset{ind}{\sim} N\left(f(p_i, \mathbf{z}_i; \boldsymbol{\gamma}), \sigma^2_\tau\right) \tag{3.15}$$

Here, d_i denotes the degrees of freedom parameter for area i, which we determine based on the survey design and sample size as discussed below. We use σ^2_τ to denote the variance of the linking model errors which allow for area-specific deviations from the linking model.

We define the generalized variance function as follows:

$$f(p_i, \mathbf{z}_i; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \log(p_i(1 - p_i)) + \gamma_2 \log(n(i)) \tag{3.16}$$

where $n(i)$ denotes the sample size for area i. Note that if we set $\gamma_0 = 0, \gamma_1 = 1, \gamma_2 = -1$, the right hand side resembles the logarithm of the binomial variance. As such, this GVF

can be viewed as a generalized version of the binomial variance. The GVF used here could also be altered to introduce additional covariates or different functional relationships between p_i and V_i. We can view the variance linking model (3.15) as a prior that shrinks the estimate $\widehat{V}_i$ towards a model-based prediction dependent on the binomial mean-variance relationship.

As described above, the mean linking model induces spatial smoothing for estimates $\widehat{p}_i$. By combining the mean and variance models and incorporating the means p_i into the GVF, we induce spatial correlation into the resulting samples of V_i, potentially aiding estimation in areas with fewer samples. When ϕ is zero, no spatial correlation is induced among the sampling variances reflecting an assumption that the mean-variance relationship for binary responses contributes more to the sampling variance than any residual spatial dependence between the variance parameters after accounting for the underlying prevalence.

We treat the degrees of freedom parameter d_i as known for all areas i. The appropriate choice for d_i depends on the sampling design. As mentioned above, if the data for a given area were iid Gaussian (for example, reflecting simple random sampling with replacement), the typical variance estimator would follow a χ^2 distribution with $d_i = n(i) - 1$ degrees of freedom. However, for sampling without replacement and cluster sampling designs, other choices of d_i may be more appropriate depending on how $\widehat{V}_i$ is computed for each area. Maples et al. (2009) outline a resampling procedure for estimating degrees of freedom for their variance sampling model. When computing variance estimates from DHS data, we use a simplified variance estimator based on the with-replacement variance estimator for multistage designs presented in Equation (4.6.2) of Särndal et al. (2003), which is computed as a sum over clusters:

$$\widehat{V}_i = \frac{1}{n_c(i)(n_c(i)-1)} \sum_{j \in S(i)} \left(\frac{\widehat{t}_{ij}}{\pi_j} - \widehat{t}_i \right)^2 \tag{3.17}$$

where $S_c(i)$ denotes the set of indices of sampled clusters for area i, $n_c(i)$ denotes the

number of sampled clusters, and π_j denotes the probability of sampling cluster j. Finally $\widehat{t}_{ij}$ denotes the direct estimator for the total for cluster j in area i and $\widehat{t}_i$ denotes the direct estimator for the total of area i. Since this is a sum of squared error terms over $n_c(i)$ clusters, we set d_i to be equal to $n_c(i) - 1$.

3.3.3 Estimation

We adopt a fully Bayesian approach to estimation by placing priors on the following hyperparameters:

$$\{\boldsymbol{\beta}, \sigma_u^2, \phi, \boldsymbol{\gamma}, \sigma_\tau^2\} \sim \Pi(\boldsymbol{\theta}) \tag{3.18}$$

where $\boldsymbol{\theta}$ denotes any parameters used to specify the priors. Details on the priors used in each example are provided in Appendix B.2. We compute approximate posterior distributions for p_i for all areas i using Markov chain Monte Carlo sampling as implemented in the Stan programming language (Carpenter et al. 2017). Functions for fitting the models described above have been collected in an R package called `VSALM` available at `https://github.com/peteragao/VSALM` and the code for the below simulations and analysis is available at the associated repository `https://github.com/peteragao/VSALM-paper`. In this context, the Bayesian approach offers a number of potential benefits. In particular, we are able to sample from the joint posterior distributions for the proportions of interest for all areas, giving a natural way to quantify uncertainty and also enabling comparisons between areas. Moreover, the sampling approach implemented in Stan is fast and flexible, enabling users to fit and compare potential models quickly.

3.4 Simulation study

3.4.1 Population generating model

We use simulations to evaluate our spatial variance smoothing estimator, comparing its performance with that of the direct weighted Hájek estimator and an estimator derived from a model without variance smoothing. For our simulations, we generate an artificial population that mimics data from the 2018 Nigeria DHS. First, we generate synthetic cluster locations across Nigeria using a pixel grid of estimated population counts for Nigeria in 2006 (mimicking the sampling frame used for the DHS survey) (WorldPop and Center for International Earth Science Information Network (CIESIN), Columbia University 2006). For each of the 73 strata used for the DHS, we sample 300 pixels without replacement with probability proportional to population. These sampled pixels represent enumeration areas or clusters. For each cluster location, we randomly generate cluster sizes $N_c \sim \text{Poisson}(10)$, yielding a population of $N = \sum_c N_c$ individuals.

For each individual i in our population, we generate data using a population generating model motivated by the models used by Corral et al. (2021) (see Section 7.2) and Gao and Wakefield (2023+). For each cluster c, we simulate cluster level covariate information as follows:

1. The covariate $x_{1,c}$ is the realized value of a binary random variable $X_{1,c}$ with $P(X_{1,c} = 1) = 0.5$;
2. The covariate $x_{2,c}$ is the realized value of a binary random variable $X_{2,c}$ with $P(X_{2,c} = 1) = 0.3 + 0.5\frac{a(c)}{37}$, where $a(c)$ is the index of the area containing cluster c;
3. The covariate $x_{3,c} = x_{3,a(c)}$ is obtained from a 37×1 ICAR random vector with marginal variance 1 for the Admin-1 areas.
4. The covariate $x_{4,c} = x_{4,a(c)}$ is obtained from a 774×1 ICAR random vector with marginal variance 1 for the Admin-2 areas.

5. The covariate $x_{5,c}$ is obtained from a random vector generated using a stochastic partial differential equation (SPDE)-based approximation (Lindgren et al. 2011) to a Gaussian process with Matérn covariance with smoothness parameter 1 and marginal variance of 1.

Maps of these covariates are provided in Appendix A.2.1. Based on these covariates, we simulate a cluster level risk parameter q_c for each cluster from the following model:

$$\text{logit}(q_c) = \text{logit}(\mu) + 0.25x_{1,c} - 0.25x_{2,c} + 0.5x_{3,c} + 0.25x_{4,c} + 0.25x_{5,c} + u_i + v_c \quad (3.19)$$

where $u_i \overset{iid}{\sim} N(0, 0.25^2)$ are independent and identically distributed area level random effects (for the area i containing cluster c), and $v_c \overset{iid}{\sim} N(0, 0.5^2)$ represents represent independent and identically distributed cluster level effects. In the above, μ denotes the global superpopulation prevalence. The covariates are held constant for all simulations, but the response variables and random area and cluster effects are resampled for each new simulation. We repeatedly generate $Y_j \mid q_{c(j)} \sim \text{Bernoulli}(q_{c(j)})$, where $c(j)$ is the cluster of individual j. For each simulation, we can thus compute true population Admin-1 area level proportions p_i. We induce spatial dependence in the responses via via the covariates $x_{3,c}, x_{4,c}$, and $x_{5,c}$, which are each simulated from multivariate Gaussian models with different spatial correlation structures. To obtain our simulated samples, we use a cluster sampling design. In each simulation, we sample eight clusters from each stratum, keeping all individuals in sampled clusters. We compute sampling weights w_i for each individual from the corresponding inverse inclusion probabilities.

We compare our unmatched joint smoothing model-based estimator with the direct **Hájek** estimator and a number of alternative model-based estimates. First, we consider spatial joint sampling (**Spatial Unmatched JS**) and non-spatial joint sampling (**Unmatched JS**) models, where the non-spatial version is obtained by replacing the BYM2 prior for the area effects $\mathbf{u}$ with an iid multivariate Gaussian prior. We also consider an estimator produced using a model that omits the variance smoothing model entirely, which we

	$\mu = 0.1$				$\mu = 0.5$			
Method	**RMSE**	**MAE**	**90% Cov.**	**MIL**	**RMSE**	**MAE**	**90% Cov.**	**MIL**
Direct (Hájek)	3.44	2.63	83	10.42	6.16	4.92	85	19.55
Unmatched MS	3.53	2.55	82	9.10	5.80	4.62	85	17.48
Spatial Unmatched MS	3.28	2.39	83	8.60	5.61	4.46	85	17.20
Unmatched JS	3.24	2.46	89	9.78	6.44	5.06	89	20.20
Spatial Unmatched JS	3.03	2.28	90	9.34	6.19	4.87	90	19.72

Table 3.1: RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and mean interval length ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling. The reduced model omits one of the spatial covariates in the full model.

refer to as the **mean smoothing (MS)** model-based estimator. This model is specified using the unmatched models (3.8) and (3.9) but treating $\widehat{V}_i$ as known for all i. We consider both spatial (**Spatial Unmatched MS**) and non-spatial versions (**Unmatched MS**). For all model-based estimators, we adopt a fully Bayesian approach to inference as described in Section 3.3. We obtain point estimates $\widehat{p}_i$ and 90% interval estimates $(\widehat{p}_i^-, \widehat{p}_i^+)$. by sampling from these approximate posterior distributions. Further details on the estimation procedure are provided in the Appendix.

$$\text{RMSE}(\widehat{\mathbf{p}}) = \sqrt{\frac{1}{m}\sum_i (p_i - \widehat{p}_i)^2} \tag{3.20}$$

$$\text{MAE}(\widehat{\mathbf{p}}) = \frac{1}{m}\sum_i |p_i - \widehat{p}_i| \tag{3.21}$$

$$\text{Cov}_{90}(\widehat{\mathbf{p}}) = \frac{1}{m}\sum_i \mathbf{1}\{p_i \in (\widehat{p}_i^-, \widehat{p}_i^+)\} \tag{3.22}$$

$$\text{MIL}_{90}(\widehat{\mathbf{p}}) = \frac{1}{m}\sum_i (\widehat{p}_i^+ - \widehat{p}_i^-) \tag{3.23}$$

We consider two sets of simulations with differing global prevalence rates. We let $\mu = 0.1$ for the first set, which is similar to the overall HIV positivity rate in the Malawi data.

For the second set, we let $\mu = 0.5$, which is similar to the national MCV-1 vaccination rate in the Nigeria data. Table 3.1 summarizes results for our two sets of simulations. In each setting, the results represent the average values of the metrics (3.20)-(3.23) across 1,000 simulated populations. We observe that in the low prevalence examples, the spatial unmatched joint model-based estimates perform best in terms of RMSE and MAE. Moreover, prediction intervals constructed based on the direct estimator and the mean-only smoothing model-based estimators tend to exhibit undercoverage, whereas the joint model-based intervals achieve closer to nominal coverage. In the moderate prevalence examples, the joint model-based estimates perform slightly worse than the mean-only model-based estimates in terms of the RMSE and MAE; however, the Hájek and mean-only model intervals show slight undercoverage since uncertainty in $\widehat{V}_i$ is not acknowledged. The joint modeling approach thus yields slightly more conservative prediction intervals which may be desirable for decision making.

3.5 Applications

We apply our joint smoothing model-based estimator to two examples involving DHS data, demonstrating its use for a low prevalence indicator (Malawi HIV prevalence rates) and for a moderate prevalence indicator (Nigeria measles vaccination rates). We show that our method induces spatial smoothing of estimated variances and produces more conservative interval estimates than an approach using an area level model that only smooths means.

For both examples, we compare direct weighted estimation with the model-based smoothing methods described above. We first fit both the spatial mean smoothing unmatched model (**Spatial Unmatched MS**), which omits the variance model, as well as the full spatial joint smoothing unmatched model (**Spatial Unmatched JS**). For all models, we use no covariates and we compute approximate posterior distributions for all area level proportions p_i and obtain corresponding point and interval estimates by sampling

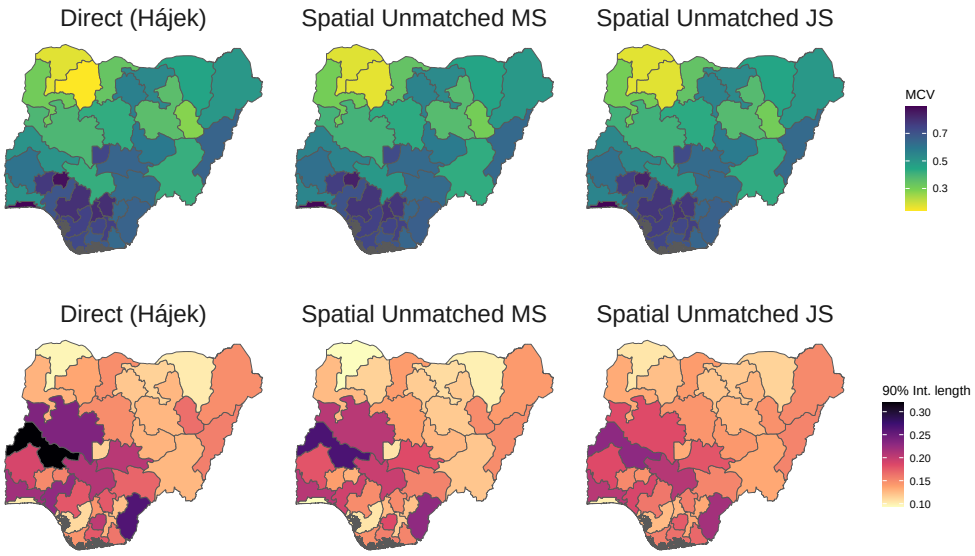

Figure 3.3: Direct and model-based point estimates (top) and length of corresponding 90% interval estimates (bottom) of vaccination coverage rate for first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria, 2018.

from these posteriors. Figure 3.3 compares point estimates of MCV-1 coverage rates (top) and the length of interval estimates (bottom) for Admin-1 areas among children aged 12–23 months in Nigeria in 2018. Figure 3.4 similarly provides point estimates of HIV prevalence rates (top) and the length of interval estimates (bottom) for Admin-1 areas for women aged 15–49 in Malawi, 2015–2016. For both examples, the bottom set of maps illustrates the estimated uncertainty of the direct and model-base estimates using the length of 90% credible intervals. In general, we observe that the point estimates agree well for all three methods. For the Malawi HIV example, the posterior median estimate of the spatial dependence parameter phi for the joint smoothing model is 0.88 with 90% credible interval (0.36, 1.00) while for the Nigeria vaccination example, posterior median of ϕ is 0.86 with 90% credible interval (0.39, 1.00). In both cases, we observe some spatial smoothing of the interval lengths, suggesting that the joint smoothing model induces

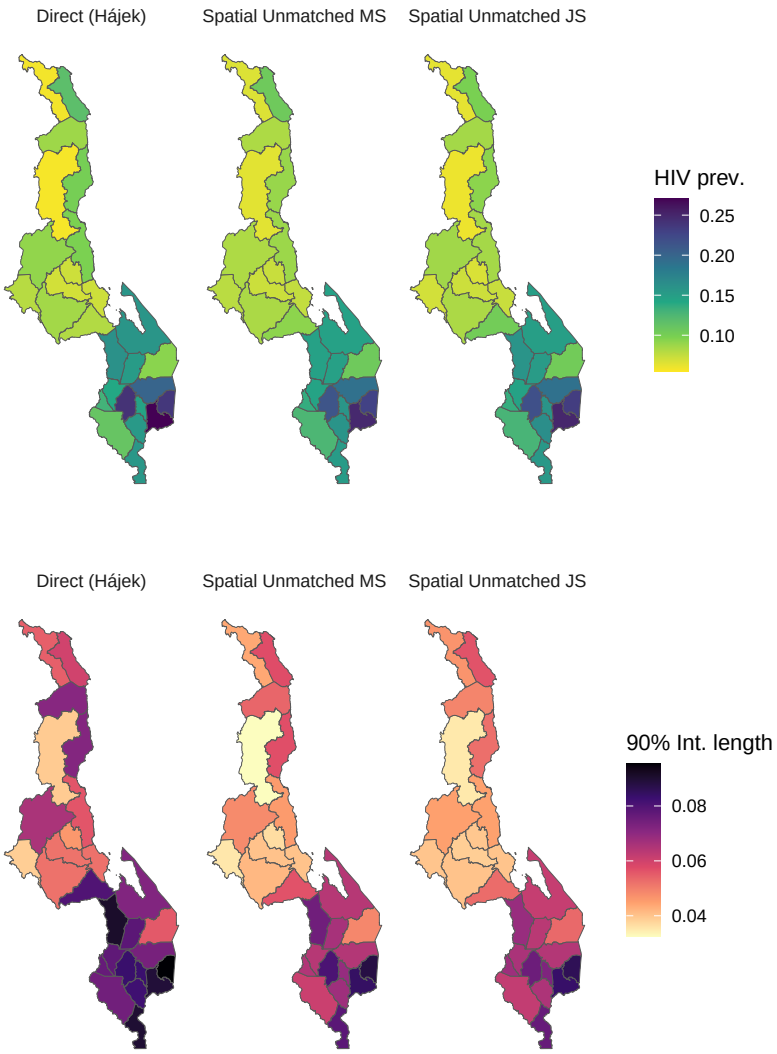

Figure 3.4: Direct and model-based point estimates (top) and length of corresponding 90% interval estimates (bottom) of HIV prevalence rate for women aged 15–49 in Malawi, 2015–2016.

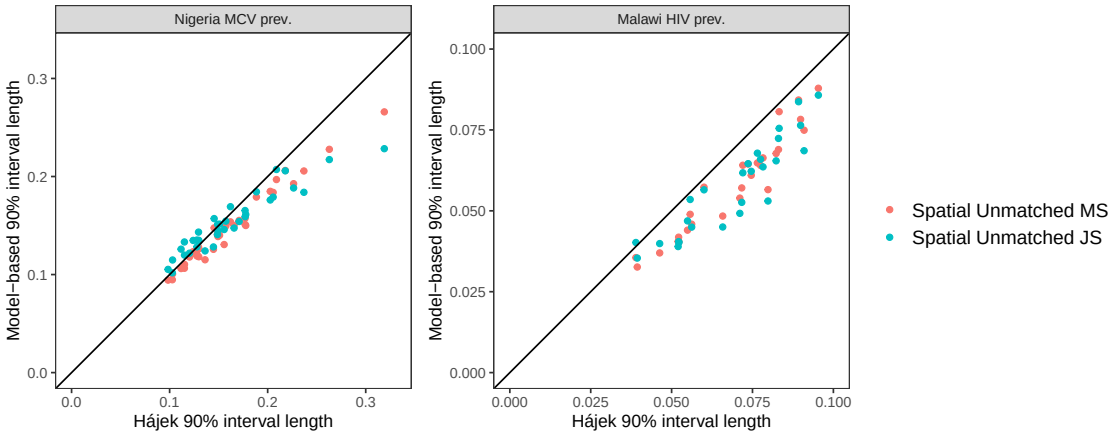

Figure 3.5: Comparison of model-based 90% credible interval lengths with Hájek 90% confidence interval lengths for Malawi HIV example (left) and Nigeria MCV example (right).

spatial smoothing of the direct estimator variances V_i. Figure 3.5 provides a scatter plot comparing the model-based 90% credible interval lengths produced by the mean smoothing and joint smoothing models with the design-based 90% confidence interval length associated with the Hájek estimator. For both examples, the joint smoothing intervals are more conservative when the Hájek intervals are short and narrower when the Hájek intervals are long. In addition to the spatial smoothing of interval lengths seen in Figures 3.3 and 3.4, this suggests the joint model can help smooth variance estimates globally.

3.6 Discussion

Our proposed model-based estimator for small area proportions combines several proposed extensions of the basic Fay-Herriot area level model, including using unmatched linking and sampling models to address the non-Gaussian response data (2002), incorporating spatial smoothing via correlated random effects in the mean linking model (2008)

and introducing a variance smoothing model so that the resulting estimators exhibit both smoothed means and variances (Maiti et al. 2014; Sugasawa et al. 2017; You and Chapman 2006). We propose a spatial joint smoothing model and adopt a fully Bayesian approach to estimation, which facilitates quick computation of point and interval estimates. Through simulation and application, we have shown that inferences based on our model can improve upon those based on a model that only incorporates smoothing of means. Interval estimates obtained from our model can correct for the undercoverage seen in models that only smooth means, suggesting our model may more accurately account for uncertainty in estimated variances of direct weighted estimators.

For our clustered binary response data, the variance smoothing model we have adopted may help address undercoverage of interval estimates caused by treating variances of direct weighted estimators as known. However, for other designs and contexts, such a model may be inappropriate. In general, the choice of variance sampling and linking models should depend on a number of factors including any clustering and stratification in the design as well as the distribution and presumed mean-variance relationship of the response variables.

Moreover, we acknowledge that our variance smoothing model is a simplification of the true distribution of design-based variances V_i. in particular, the use of a chi-squared distribution for the variance sampling model relies upon the assumption that the direct estimator of variance $\widehat{V}_i$ for a particular area i is computed as the sum of several squared Gaussian terms. Since our data is non-Gaussian, this assumption may be violated and other sampling models for $\widehat{V}_i$, such as a Gaussian model, could be explored in future work. Moreover, within each area, we have assumed that the variance for each stratum is equal, but this assumption may be inappropriate for our data since each area of interest is divided into urban and rural subregions, which may be qualitatively different from one another. Finally, the appropriate number of degrees of freedom d_i depends on the specific design used; using resampling methods like those explored by Maples et al. (2009) to

choose d_i may help improve the fit of the variance sampling model.

Although we have presented one approach applied to two different types of problems, in practice, for decision making purposes, different estimation problems may have varying priorities. For example, when implementing targeted vaccination programs, it is important to identify communities with especially low vaccination rates, whereas designing policy for providing resources associated with HIV involves identifying communities with high rates of positivity. Given that the variance of a direct prevalence rate estimator may depend on its expected value, various modeling decisions such as choosing to apply a transformation for $\widehat{p}_i$ may lead to different results depending on the expected value of $\widehat{p}_i$. As such, it is crucial to carefully consider the distribution of direct estimators before selecting a model. In the above, we have used unmatched sampling and linking models for the area level proportions, but we also considered first computing the logit-transformed direct estimators and then applying matched sampling and linking models treating both $\widehat{p}_i$ and p_i as Gaussian random variables. In our simulations and application, this approach did not outperform the unmatched models we adopted, but future research could help illustrate when such an approach could be useful.

When mapping subnational health and demographic indicators in LMIC, unit level models, and in particular geostatistical models using spatial Gaussian processes, are often used as they allow estimates to be generated at arbitrary resolutions and can incorporate unit level covariate information. However, such approaches may often struggle to account for design effects such as those caused by clustering and informative sampling. While unit level models may be able to generate prevalence estimates at the individual cluster level, aggregating those cluster level estimates upwards to produce area level estimates may introduce additional errors and lead to improperly calibrated interval estimates (Fuglstad et al. 2022; Paige et al. 2022a; b). Area level models are specified to generate estimates for a preselected set of regions. Moreover, area level models are often simpler and faster to implement than unit level models. For these reasons, we have ex-

plored the feasibility of using area level models to generate maps of health indicators such as vaccination rates and disease prevalence rates in LMIC. Our method, like many area level methods, directly accounts for survey design by incorporating available sampling weight information. By incorporating a spatial variance smoothing model and using unmatched sampling and linking models, we are able to address some of the difficulties related to applying area level models for use in this specific context.

Chapter 4

COMBINING AREA LEVEL AND UNIT LEVEL MODELING FOR SMALL AREA ESTIMATION OF PROPORTIONS

4.1 Introduction

In SAE, model-based approaches are typically based on either area level models or unit level models. Area level models treat direct weighted estimators as response data, incorporating shrinkage and area level covariate information to improve the precision of estimates. Since the direct estimators are often design consistent, the resulting model-based estimators inherit favorable design optimality properties. Moreover, since only aggregate quantities are modeled, fewer distributional assumptions about response data are needed.

Conversely, unit level models incorporate higher resolution covariate information and can directly model binary responses or count data. When using unit level models, neglecting to acknowledge the survey design can result in biased or poorly calibrated estimators. There are two potentially intertwined issues: informative sampling and clustering. Under informative sampling, where the sample response is correlated with the inclusion probability even after conditioning on model covariates, unit level model-based estimators may be biased unless the estimation procedure is adjusted to account for this dependence (Parker et al. 2020; Pfeffermann and Sverchkov 2007). Similarly, when cluster sampling is used, failing to account for within-cluster correlation may reduce the accuracy of point estimates and result in improperly calibrated interval estimates.

Drawing upon the motivating example presented in Section 2.4, we focus on estimating

area-specific proportions, where we use p_i to denote the rate of interest for area i:

$$p_i = \frac{1}{N(i)} \sum_{j \in U(i)} y_{ij} \tag{4.1}$$

Estimating small area proportions as opposed to means of continuous variables introduces a number of additional challenges. Traditional area level models such as the Fay-Herriot model they typically assume that estimators $\hat{p}_i$ are normally distributed and centered on p_i, but may be inappropriate as p_i is bounded between 0 and 1. For this reason, unit level models specific to binary response data may be preferable if survey microdata are available, as long as the modeling and estimation process accounts for the survey design.

To address these challenges, we propose a two-stage smoothed model-assisted estimator of small area proportions that draws from both area level and unit level methods. Our approach first uses a working unit level model that leverages unit level auxiliary information to generate model-assisted estimates. We treat these model-assisted estimates as data for a second-stage area level model that incorporates spatial smoothing and models proportions on the logit-transformed scale. As long as the model-assisted estimators are design-consistent and their design variances converge to zero, our smoothed model-assisted estimators will also be design-consistent. Our method can thus be viewed as a bridge between classical SAE approaches and the geostatistical unit level models commonly used in global health research. Our proposed model incorporates unit level covariates and spatial random effects while also explicitly accounting for the sampling design.

4.1.1 Stage One: Model-Assisted Estimation

While the unit level models described in Section 2.4 may produce biased estimators under model misspecification, model-assisted estimators are motivated by working superpopulation models but are specified to ensure design consistency and unbiasedness even when

the working model is incorrect. For a review of model-assisted methods see Särndal et al. (2003) or Breidt and Opsomer (2017) . A characteristic model-assisted approach is given by the difference estimator

$$\widehat{p}_i^{DIF} = \frac{1}{N(i)} \left\{ \sum_{j \in U(i)} \widehat{y}_{ij} + \sum_{j \in S(i)} w_{ij}(y_{ij} - \widehat{y}_{ij}) \right\} \tag{4.2}$$

where $\widehat{y}_{ij}$ represents the working model prediction for unit j in area i. The difference estimator for area i combines model-based predictions from the working model with a direct estimator of the mean of the residuals in the area based upon the sample. Breidt and Opsomer (2017) show that under certain regularity conditions, the difference estimator is design-consistent. In particular, they assume the direct estimator for the residual mean is design-consistent and that predictions from the working model estimated on sample data and predictions from the working model estimated on the full population are asymptotically equivalent. The popular generalized regression estimator (GREG) can be framed as an example of a difference estimator using a working linear regression model to generate predictions (Särndal et al. 2003). In the case of binary response data, Lehtonen and Veijanen (1998) previously proposed the use of a working logistic regression model to compute a logistic generalized regression (LGREG) estimator. Kennel and Valliant (2010) extended the LGREG for use with cluster sample and Myrskylä (2007) compared the LGREG and GREG with binary responses, finding that when the model fit is strong, the LGREG is preferable.

For the first stage of our smoothed model-assisted approach for estimating small area proportions, we compute a model-assisted estimator using a working logistic regression model of the form:

$$P(y_{ij} = 1 \mid \mathbf{x}_{ij}, \boldsymbol{\beta}) = q_{ij} \tag{4.3}$$

$$\text{logit}(q_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} \tag{4.4}$$

Since we use one working model for all areas, our approach resembles the “modified

GREG" or survey regression estimator described by Rao and Molina (2015), except applied for a logistic generalized regression estimator. We estimate model parameters by maximizing a sampling-weighted log-likelihood function. This parameter estimation strategy is also called the pseudo-likelihood approach and has been reviewed by Lumley and Scott (2017). Based on the parameter estimates, we generate working predictions $\widehat{y}_{ij} = \text{expit}(\mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}})$ for all i and all $j \in U(i)$.

Based on these predictions, we construct the model-assisted estimator as a difference estimator:

$$\widehat{p}_i^{MA} = \frac{1}{\widehat{N}(i)} \left(\sum_{j \in U(i)} \widehat{y}_i + \sum_{j \in S(i)} w_{ij}(y_{ij} - \widehat{y}_{ij}) \right) \tag{4.5}$$

where $\widehat{N}(i) = \sum_{j \in S(i)} w_{ij}$, yielding a Hájek-like estimator. Under certain regularity conditions, this estimator is design-consistent; for further details see the Appendix.

Model-assisted estimators are typically asymptotically design unbiased and design-consistent, but quantification of uncertainty can be difficult. Linearization-based variance approximations generally do not account for uncertainty in the first sum on the right of Equation (2.11) resulting from model estimation (Myrskylä 2007). The working model should be carefully selected as overfitting can also result in underestimation of uncertainty. For our model-assisted estimator, we estimate variance by modifying the with-replacement variance estimator of a total described by Kennel and Valliant (2010) for use with a mean:

$$\widehat{V}(\widehat{p}_i^{MA}) = \frac{1}{\widehat{N}(i)^2} \frac{n(i)}{(n(i)-1)} \sum_{j \in S(i)} \left(w_{ij} e_{ij} - \widehat{\overline{e}}_i \right)^2 \tag{4.6}$$

where $n(i)$ denotes sample size for area i, $\widehat{\overline{e}}_i = \frac{1}{n(i)} \sum_{j \in S(i)} w_{ij} e_{ij}$, and $e_{ij} = \widehat{y}_{ij} - y_{ij}$. This estimator is designed for unclustered sampling designs; when applying our approach to DHS data, we adapt Kennel and Valliant's cluster sampling variance estimator. Note that this variance estimator ignores variability resulting from $\widehat{N}(i)$ and estimation of the regression parameters. In practice, variance estimation may be improved via resampling methods such as the bootstrap.

4.1.2 *Stage Two: Spatial Logistic Area Level Model*

After computing the model-assisted estimators and their associated variance estimators, we use a Fay-Herriot model to smooth across areas. Since our targets of estimation p_i are bounded between 0 and 1, we incorporate a logit transformation into both the sampling and linking models. In essence, we apply a spatial area level model to logit-transformed model-assisted estimators. Our linking and sampling models can be specified as follows:

$$\text{logit}(p_i) = \mathbf{x}_i^T \boldsymbol{\beta} + U(i) \tag{4.7}$$

$$\text{logit}(\widehat{p}_i^{MA}) = \text{logit}(p_i) + \epsilon(i) \tag{4.8}$$

where for $i = 1, \ldots, m$, $\widetilde{\mathbf{x}}_i = (1, \widetilde{x}_{i1}, \ldots, \widetilde{x}_{ip})^T$ represents a length $p+1$ vector of area-specific covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denotes the vector containing the intercept and corresponding fixed effect coefficients. We use $\mathbf{u} = (u_1, \ldots, u_m)^T$ to denote random area level effects, which we assume to be spatially correlated and drawn from a multivariate Gaussian distribution, $\mathbf{u} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_u^2, \phi))$. Here, σ_u and ϕ denote parameters controlling the spatial correlation matrix $\boldsymbol{\Sigma}$. Finally, we use $\epsilon(i)$ to denote independent sampling errors $\epsilon(i) \sim N(0, V_i)$, where $V_i = \text{Var}(\text{logit}(\widehat{p}_i^{MA}))$, which we treat as known. In practice, we estimate V_i by first estimating $\widehat{V}(\widehat{p}_i^{MA})$ using Equation (4.6) and then applying the delta method to obtain the approximation:

$$V_i \approx \frac{\widehat{V}(\widehat{p}_i^{MA})}{(\widehat{p}_i^{MA}(1 - \widehat{p}_i^{MA}))^2} \tag{4.9}$$

We adopt a hierarchical Bayesian approach to inference by defining hyperparameter priors, yielding the following alternative representation:

$$\text{logit}(\widehat{p}_i^{MA}) \mid p_i, V_i \overset{ind}{\sim} N(\text{logit}(p_i), V_i), \qquad i = 1, \ldots m \tag{4.10}$$

$$\text{logit}(\mathbf{p}_i) \mid \boldsymbol{\beta}, \sigma_u^2, \phi \sim \boldsymbol{\mathcal{N}}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\sigma_u^2, \phi)) \tag{4.11}$$

$$\boldsymbol{\beta}, \sigma_u^2, \phi \sim \pi(\boldsymbol{\xi}) \tag{4.12}$$

where we use $\text{logit}(\mathbf{p}_i)$ to denote the vector $(\text{logit}(p_1), \ldots, \text{logit}(p_m))^T$, and $\widetilde{\mathbf{X}}$ to denote the $m \times (p+1)$ matrix of area level covariates. Finally, $\pi(\boldsymbol{\xi})$ denotes the hyperparameter

priors and ξ represents corresponding parameters, which should be specified based on the specific application at hand.

By specifying different structures for $\mathbf{\Sigma}$, we can obtain varying models for the spatial dependence in $\text{logit}(\mathbf{p}_i)$. Typically, we specify an $m \times m$ adjacency matrix representing adjacency relationships between the areas. We model the area level random effects $\mathbf{u}$ using the BYM2 model, a reparametrization of the Besag-York-Mollié (Besag et al. 1991) model proposed by Riebler et al. (2016) for the area level random effects vector $\mathbf{u}$:

$$\mathbf{u} = \sigma_u \left(\sqrt{1-\phi}\tilde{\mathbf{u}}_1 + \sqrt{\phi}\tilde{\mathbf{u}}_{2*} \right) \tag{4.13}$$

Here, we assume $\tilde{\mathbf{u}}_1 \sim N(\mathbf{0}, \mathbf{I})$ is an random area effect with no spatial structure. We use $\tilde{\mathbf{u}}_{2*}$ to denote a structured spatial component which follows an intrinsic conditional autoregressive (ICAR) model (intuitively, the mean of $\tilde{u}_{2i*}$ is set to the mean of all neighboring effects and the precision is specified to be proportional to the number of neighbors). As such, σ_u controls the marginal variance of $\mathbf{u}$ and ϕ controls the proportion of variation assigned to the structured component. Under this model, $\mathbf{u}$ has the covariance matrix

$$\text{Var}(\mathbf{u} \mid \sigma_u, \phi) = \sigma_u((1-\phi)\mathbf{I} + \phi\mathbf{Q}_*^-) \tag{4.14}$$

Above, $\mathbf{Q}_*^-$ is the generalized inverse of $\mathbf{Q}_*$, which denotes the precision matrix of $\tilde{\mathbf{u}}_{2*}$. As discussed by Riebler et al. (2016), $\mathbf{Q}_*$ is singular, making the ICAR prior for the random effects improper, so we place a sum to zero constraint on the elements of $\mathbf{u}$ to ensure identifiability. Moreover, the marginal variance of each effect $\mathbf{u}_i$ depends on its number of neighbors, so to make the overall variance parameter σ_u interpretable, $\mathbf{Q}_*$ is scaled (following the procedure described in Section 3.2 of Riebler et al. (2016)) to make the geometric mean of the marginal variances is equal to one.

Following Riebler et al. (2016) and Simpson et al. (2017), we place penalized complexity (PC) priors on σ_u and ϕ. These priors penalize the Kullback-Leibler distance of a full model from a simpler base model and shrink ϕ and σ_u to zero. We place a flat prior on $\boldsymbol{\beta}$, so that $\pi(\boldsymbol{\beta}) \propto 1$. To fit our spatial logistic area level model, we use the R package `INLA`

(Rue et al. 2017) which is commonly used to conduct approximate Bayesian inference for hierarchical models and is popular for mapping health indicators Utazi et al. (2020).

The approach presented here is specialized for spatially structured binary response data, but the overall strategy of using area level models to smooth model-assisted estimators can be adapted for other types of data. If the response is continuous, rather than binary, the same approach can be applied using a working linear regression model in the first stage and using a similar second stage model, but without applying the logit transformations, as proposed by Fay (2018). Other models for spatial random effects could be used, including, for example, those discussed in Section 2.4.3

4.2 Simulation study

4.2.1 Population generating model

Below, we use simulations to compare our smoothed model-assisted estimator with existing direct, model-assisted, and model-based estimators. The set up is motivated by simulations used by Corral et al. (2021). Using the WorldPop 100m population counts grid for Nigeria corresponding to the 2006 census (2006), we sample 300 pixels without replacement with probability proportional to population in each of 73 strata defined by crossing the 37 Admin-1 areas with urban/rural status (one area corresponding to Lagos is entirely urban). Each sampled pixel represents a simulated cluster location. We then randomly generate cluster sizes for each simulated cluster so that the size of cluster c is given by $n_c \sim \text{Poisson}(15)$. For each cluster c in area i, we simulate a cluster level risk q_{ic} using the model:

$$\begin{aligned}\text{logit}(q_{ic}) = x_{1,ic} - x_{2,ic} + 0.5x_{3,ic} + 0.25x_{4,ic} + 0.25x_{5,ic} \\ + 1.5x_{6,ic} + 0.1x_{7,ic} + 0.1x_{8,ic} + u_i + v_{ic} \quad (4.15)\end{aligned}$$

where $u_i \overset{iid}{\sim} N(0, 0.1^2)$ are independent and identically distributed area level random effects, and $v_{ic} \overset{iid}{\sim} N(0, 0.5^2)$ represents random and independent and identically distributed

cluster level effects. The covariates are specified as follows:

1. The covariate $x_{1,ic}$ is the realized value of a binary random variable $X_{1,ic}$ with $P(X_{1,ic} = 1) = 0.5$;
2. The covariate $x_{2,ic}$ is the realized value of a binary random variable $X_{2,ic}$ with $P(X_{2,ic} = 1) = 0.3 + 0.5\frac{a(c)}{37}$;
3. The covariate $x_{3,ic} = x_{3,i}$ is obtained from a 37×1 random vector modeled as an ICAR random effect with marginal variance 1 for the Admin-1 areas.
4. The covariate $x_{4,ic} = x_{4,i}$ is obtained from a 774×1 random vector modeled as an ICAR random effect with marginal variance 1 for the Admin-2 areas.
5. The covariate $x_{5,ic}$ is obtained from a random vector generated using a stochastic partial differential equation (SPDE) -based approximation to a Gaussian process with Matérn covariance (smoothness 1) and marginal variance of 1 (Lindgren et al. 2011).
6. The covariate $x_{6,ic}$ is obtained from a random vector generated using an SPDE-based approximation to a Gaussian process 7. The covariate $x_{7,ic}$ denotes estimated travel times to cities in 2015 (Weiss et al. 2018).
7. The covariate $x_{8,ic}$ denotes proportion of people per grid square living in poverty in 2010 (Tatem et al. 2017).

The covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ represent informative non-spatial covariates, while $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$ exhibit spatial correlation. The covariates $\mathbf{x}_7$ and $\mathbf{x}_8$ are based on real covariates commonly used for modeling health outcomes in LMIC. Based on the above cluster level risks, we generate responses $Y_{icj} \sim \text{Bernoulli}(q_{ic})$ where Y_{icj} denotes unit j in cluster c in area i. As described above, our population consists of 300 clusters of varying sizes. From this population, we repeatedly sample ten clusters from each stratum, using all response values from each sampled cluster. We use an informative sampling scheme in which we oversample clusters with large values for $x_{6,ic}$: clusters with values of $x_{6,ic}$ in the top quartile are three times as likely to be sampled as clusters in the bottom three quartiles.

Since the values of $\mathbf{x}_6$ are spatially correlated, this may induce spatial structure in the model residuals if this oversampling is not addressed when estimating model parameters. Based on this design, we compute sampling probabilities and design weights w_{ij} for each individual. In practice, we generate the covariate values and cluster sizes once and then sample a list of indices identifying the sampled clusters. These indices and cluster characteristics are held constant across simulations but the response variables, area effects, and cluster effects are repeatedly regenerated.

4.2.2 Estimation procedure

For each simulation, we compute true population Admin-1 area level proportions p_a and compare with several estimators computed from the sampled data. For all estimators that rely on covariate modeling, we consider two potential models, a reduced model and a full model. The full model includes all covariates except $\mathbf{x}_4$. We remove the area-specific covariate $\mathbf{x}_4$ in order to induce spatial correlation in the model residuals. The reduced model includes all covariates except $\mathbf{x}_4$ and $\mathbf{x}_6$ so it does not account for the effect of oversampling the stratum defined by $\{x_{6,ic} > \text{median}(x_{6,ic})\}$, meaning the design is not ignorable after conditioning on model covariates. Conversely, the full model partially accounts for this by including $\mathbf{x}_6$ as a covariate. Furthermore, for all model-based approaches incorporating smoothing via random effects, we consider both non-spatial smoothing using iid Gaussian area level random effects and spatial smoothing using the BYM2 model for area level random effects.

Below, we describe the estimators used for comparison. First, we compute the direct weighted **Hájek** estimator. We also compute model-assisted estimators (**MA**) using both the full and reduced models. Next, we compute several area level model-based estimators. Applying the spatial logistic area level model described in Section 4.1.2 to the Hájek estimator yields a spatial smoothed Hájek (**SH**) estimator. Similarly, by applying the same model to the model-assisted estimator, we obtain our proposed spatial smoothed

model-assisted estimator (**SMA**). For comparison, we also compute non-spatial versions of smoothed estimators by assuming independent and identically distributed Gaussian random effects in the logistic area level linking model given in Equation (4.7).

Finally, we compute a number of unit level model-based estimators, using a **Binomial** model as well as two models designed to account for effects of clustering: a **Betabinomial** model and a lognormal-binomial **Lono-Binomial** model (Dong and Wakefield 2021). These particular likelihoods, as used in SAE, are discussed in further detail by Dong and Wakefield (2021), but we briefly outline their use here. First, we implement the binomial unit level model specified in Equation (2.15). The betabinomial model accounts for overdispersion in our response data potentially related to clustering and can be specified as follows. We assume each unit in a given cluster c in area i has the same risk q_{ic}:

$$Y_{icj} \mid q_{icj} \sim \text{Bernoulli}(q_{ic}) \tag{4.16}$$

$$q_{ic} \mid \mu_{ic}, d \sim \text{Beta}(\mu_{ic}, d) \tag{4.17}$$

$$\text{logit}(q_{ic}) = \mathbf{x}_{ic}^T\boldsymbol{\beta} + U(i) \tag{4.18}$$

where we parameterize the beta distribution via

$$E(q_{ic} \mid \mu_{ic}, d) = \mu_{ic} \tag{4.19}$$

$$\text{Var}(q_{ic} \mid \mu_{ic}, d) = \frac{\mu_{ic}(1-\mu_{ic})}{d+1} \tag{4.20}$$

Above, d denotes a dispersion parameter. The lognormal-binomial model (referred to by Dong and Wakefield as the Lono-Binomial Overdispersion model) instead assumes that

$$y_{icj} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}, u_i \sim \text{Bernoulli}(r_{ic}) \tag{4.21}$$

$$\text{logit}(r_{ic}) = q_{ic} + v_{ic} = \mathbf{x}_{ic}^T\boldsymbol{\beta} + u_i + v_{ic} \tag{4.22}$$

where for all areas i and clusters c, r_{ic} denotes a cluster level parameter defined as the sum of the cluster level prevalence q_{ic} and iid Gaussian cluster level error v_{ic}. For all of these models, we implement both non-spatial iid and spatial BYM2 models for the area

level random effects $\mathbf{u}$. For all unit level models, area level estimates are made by making predictions of q_{ic} for all clusters in the population and then aggregating upwards to the area level.

Additional information on the estimation procedures, including information on software used and priors for model hyperparameters, can be found in the Appendix. Code for the simulations (and for the application detailed below) can be found on GitHub.

4.2.3 Results

For each method, we compute point estimates $\widehat{p}_i$ as well as 90% interval estimates $(\widehat{p_i}^-, \widehat{p_i}^+)$. For each vector of estimates $\widehat{p}_i$, we compute root mean squared error (RMSE) and mean absolute error (MAE). We also compute the coverage of the 90% interval estimates and the mean interval lengths (MIL) across all areas. For definitions of these metrics, see Section 3.4. We summarize these error metrics by averaging their values across all 1,000 simulated populations. For all estimators incorporating covariate information, we provide comparisons for both reduced (Table 4.1) and full models (Table 4.2) across the 1,000 generated response vectors. Note that the Hájek and SH estimators do not make use of any covariates. In general, introducing covariate information reduces the error of point estimates and the methods using the full set of covariates achieve the lowest error. For the reduced models in Table 4.1, the non-spatial SMA estimator does not improve on the MA estimator, suggesting that the smoothing via independent random effects is not particularly beneficial. However, the spatial SMA estimator performs best, suggesting that when there are relevant design variables that are left out of the fitted model, the SMA approach may improve upon direct or unit level approaches. The non-spatial SMA estimator, which uses independent area effects may be less able to capture the differences between areas due to the spatial covariates omitted from the reduced model.

In Table 4.2, we observe that the spatial unit level models perform best, suggesting that when all relevant design variables are included in the model and the design is truly ignor-

	Method	**RMSE**	**MAE**	**90% Cov.**	**MIL**
	Direct (Hájek)	4.44	3.28	86	13.99
	MA	3.70	2.81	87	11.82
Non-spatial	SH	4.84	3.44	89	14.66
	SMA	3.71	2.79	87	12.27
	Binomial	4.22	3.11	75	8.24
	Betabinomial	4.13	3.05	83	10.10
	Lono-Binomial	4.42	3.23	81	9.95
Spatial	SH	4.24	3.13	91	13.40
	SMA	***3.47***	***2.65***	87	11.49
	Binomial	4.14	3.02	76	8.06
	Betabinomial	3.96	2.88	85	9.84
	Lono-Binomial	4.38	3.18	81	9.72

Table 4.1: Averaged RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and MIL ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling for methods using no covariates or only the reduced set of covariates (omitting one of the spatial covariates used in population generation). The lowest RMSE and MAE are in ***bold italics***.

	Method	RMSE	MAE	90% Cov.	MIL
	MA	3.17	2.41	87	9.68
Non-spatial	SMA	3.18	2.41	87	9.92
	Binomial	2.68	2.02	88	7.99
	Betabinomial	2.69	2.04	89	8.43
	Lono-Binomial	4.42	3.23	81	9.95
Spatial	SMA	3.03	2.28	88	9.47
	Binomial	***2.62***	***1.96***	88	7.76
	Betabinomial	***2.62***	***1.96***	90	8.21
	Lobo-Binomial	***2.62***	***1.96***	89	8.10

Table 4.2: Averaged RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and MIL ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on sample data obtained via informative sampling for methods using the full set of covariates. The lowest RMSE and MAE are in ***bold italics***.

able, unit level approaches may be preferred. However, the Spatial SMA approach still offers benefits over the Hájek and MA approaches. In terms of calibration, The Hájek, model-assisted, and area level model-based estimators have coverage rates close to the nominal 90% rate. However, some of the unit level model-based interval estimates exhibit undercoverage, especially when only the reduced set of covariates is used. For the full set of covariates, the estimators based on the binomial display undercoverage, while the betabinomial and lognormal-binomial interval estimates achieve close-to-nominal coverage.

4.3 Application: Vaccination coverage in Nigeria

We apply our smoothed model-assisted estimator to generate Admin-1 level estimates of measles vaccination rates using the 2018 Nigeria DHS data. In this case, y_{ij} represents observed vaccination status of child j in area i. We use two main unit level covariates obtained from grid-based estimates of travel times to cities in 2015 (Weiss et al. 2018) and the proportion of people per grid square living in poverty in 2010 (Tatem et al. 2017). The associated fixed effect estimates were significantly different from zero in a survey-weighted logistic regression with measles vaccination as outcome; however, these covariates are themselves estimated using geostatistical models, so any associations should be interpreted with caution. We also use a map of estimated population density (WorldPop) to derive a binary covariate that classifies each pixel as either urban or rural.

When using unit level covariates to predict binary response variables, covariate information on the entire population is required to generate estimates. In our setting, when recent and reliable population data may not be available, satellite imagery can provide covariates on a pixel grid spanning the domain. Instead of predicting each child separately, we generate predictions for each pixel and average over the pixel level predictions for a given area. When averaging, we weight each pixel's prediction by the estimated number of children aged 1-5 in the pixel using maps created by WorldPop. We harmonize

the covariate rasters and population density rasters to a common pixel grid, the 1km by 1km grid provided by WorldPop. We also use the map of estimated population density to derive a binary covariate that classifies each pixel as either urban or rural assigning the highest density pixels in a given area to be urban so that the total proportion of population classified as urban in each area matches the proportion reported in the 2018 Nigeria DHS report.

Using this data, we compare a number of the estimation methods outlined above. We first consider the Hájek estimator $\widehat{p}_i^H$ and the model-assisted estimator $\widehat{p}_i^{MA}$ where the working model is a logistic regression model. We then consider smoothed Hájek estimators and smoothed model-assisted estimators obtained by fitting the model specified in (4.7) and (4.8) for $\widehat{p}_i^H$ and $\widehat{p}_i^{MA}$, respectively. For each, we consider both iid and BYM2 models for the area level random effects $\mathbf{u}$, yielding four estimators: smoothed Hájek with iid area effects (**SH**) and BYM2 area effects (**Spatial SH**) as well as smoothed model-assisted with iid area effects (**SMA**) and BYM2 area effects (**Spatial SMA**).

Finally, we consider a geostatistical model; to account for clustering, we use the **Spatial Betabinomial** model described above with a BYM2 prior for the area level random effects $\mathbf{u}$. Since pixels do not necessarily coincide with clusters, we cannot use the Lono-Binomial model, which requires us to identify the sampling frame of clusters in order to aggregate estimates appropriately. Figure 4.1 compares point estimates of measles vaccination rates (left) and the length of interval estimates (right) for Admin-1 areas among children aged 12-23 months in Nigeria in 2018. Point and interval estimates for all methods are provided in the Appendix. We omit results for the non-spatial mixed models as their results are similar to those of the spatial models. On the right side, we quantify uncertainty using the length of 90% credible intervals (for the smoothed and unit level models) and design-based confidence intervals (for the Hájek and model-assisted estimators) for Admin-1 areas in Nigeria in 2018. The interpretation of uncertainty estimates requires some care since the intervals for the Hájek and model-assisted estimators only estimate design-based un-

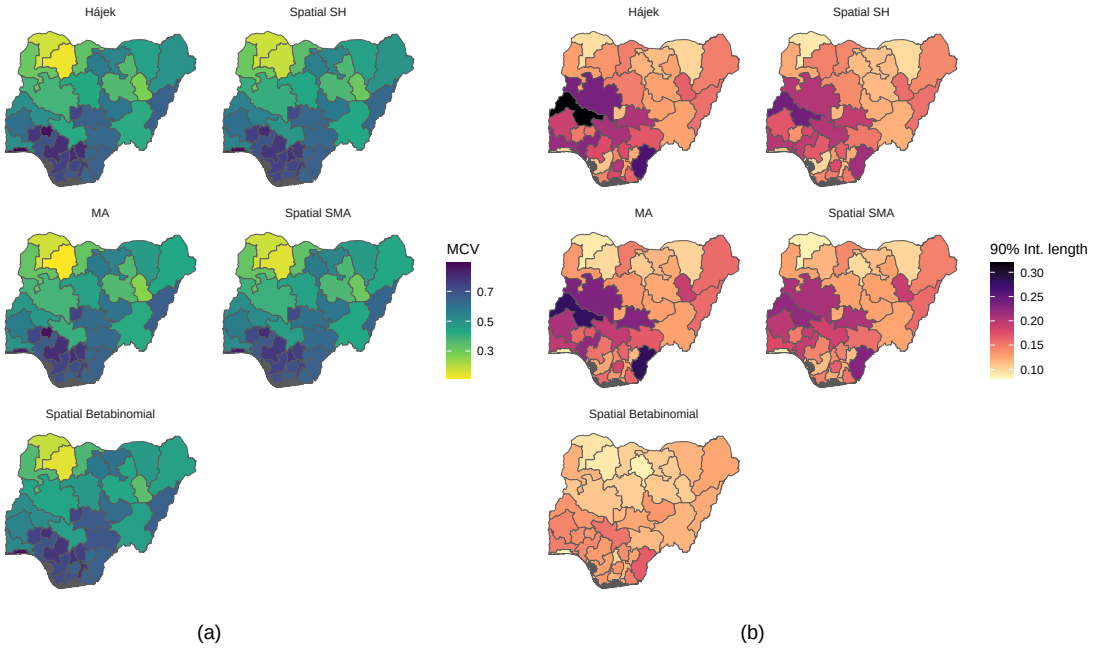

Figure 4.1: Estimated measles vaccination rates (left) and 90% prediction interval lengths for estimated measles vaccination rates (right) among children aged 12-23 months for Admin-1 areas in Nigeria in 2018.

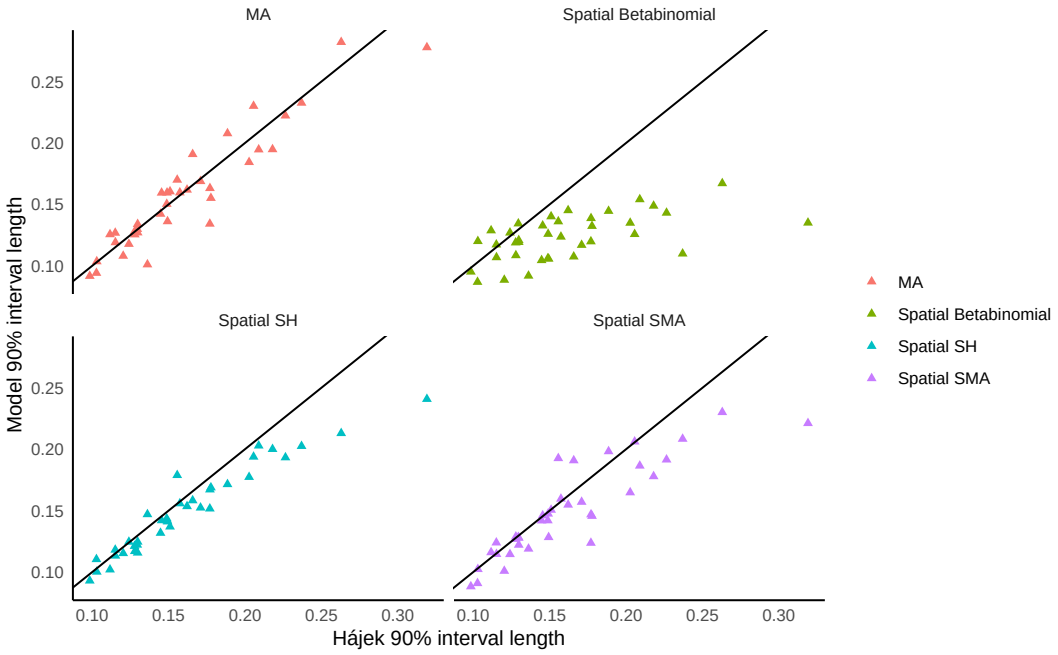

Figure 4.2: Comparison of reference Hájek design-based 90% confidence interval length (x-axis) with model-assisted and model-based interval estimate lengths (y-axis) for four methods for measles vaccination rates.

certainty, while the smoothed and betabinomial intervals are drawn from posterior distributions which also account for model parameter uncertainty. The point estimates for all the methods are similar but the interval estimate lengths vary considerably. In particular, incorporating unit level covariates shrinks the interval estimates as seen when comparing the Hájek and model-assisted estimators. The interval lengths are shortest for the unit level model, while the smoothed Hájek and smoothed model-assisted intervals are more conservative. Figure 4.2 compares the interval lengths for the Hájek estimates with the lengths of interval estimates produced by the other methods, illustrating that the unit level model (betabinomial) intervals are considerably shorter than those produced by the rest of the methods. In particular, our smoothed model-assisted intervals are more conservative; as our simulations show, when relevant design variables are omitted from unit

level models, resulting prediction intervals can exhibit undercoverage and corresponding smoothed model-assisted intervals may be better calibrated. In DHS surveys, clusters are sampled with probability proportional to size, but the cluster sizes are generally not published. As such, cluster size may be a relevant design variable that we are unable to incorporate into unit level models. The smoothed model-assisted point estimates and interval estimates may thus be preferable to the unit level model estimates. Among the unit level models, we recommend the use of the betabinomial estimates which account for potential clustering effects.

4.4 Discussion

Our proposed smoothed model-assisted estimator for small area means incorporates unit level covariate information and smoothing via random effects while retaining favorable design optimality properties. Our method seeks to bridge the SAE and model-based geostatistics literatures, drawing from and offering benefits to both perspectives.

The basic question of how best to estimate area specific quantities given limited data arises in many settings; in a sense, any subpopulation with limited sample data may be considered a small area. For example, multilevel regression and poststratification has been used to generate local estimates of opinion using survey data with high nonresponse or nonprobability sampling (Si et al. 2020). Our smoothed model-assisted approach is particularly tailored for estimating subnational health and demographic indicators. In this context, properties like asymptotic design unbiasedness and design consistency are high priorities for national statistics offices that create and distribute estimates. Using spatial and spatio-temporal smoothing and unit level covariate information in small area estimation may offer large benefits in areas with limited data. Finally, the household surveys used in this setting typically have high response rates, informative sampling weights, and geographic information.

Although the above simulations and application illustrate potential benefits of our ap-

proach, in some settings, the new method may offer limited improvement. When data is not available in every area for which estimates are desired, it is still computationally possible to sample from the smoothing model posterior for unsampled areas. However, as they do not incorporate actual observations from the area in question, such estimates cannot meaningfully be called design-consistent. Typical unit level models enable predictions to be made for unsampled areas and when sampling is not informative, such predictions may be preferable to those that would result from our approach. Another limitation is that our estimator requires careful specification of the working model. When the model is overly flexible, typical approximations that ignore variability from model estimation will underestimate the variance of model-assisted estimators. Resulting smoothed model-assisted estimators may thus be over confident. Finally, when reliable population information is unavailable, as may be the case when estimating health and demographic indicators in LMIC, it is common to use satellite-derived covariate rasters and aggregate pixel level predictions to compute area level estimates. This aggregation process is affected by measurement error in the covariates, misalignment between population density maps and household locations, and the resolution of the pixel grid. The effects of aggregation on the resulting area level estimates are not well understood; Paige et al. (2022a) consider potential implications.

When unit level covariates are strongly associated with a variable of interest, using covariate modeling in SAE offers accuracy and efficiency gains. However, unit level models do not generally produce design-consistent estimators. Various solutions have been proposed, including pseudo-likelihood and specifically pseudo-Bayesian methods, pairwise likelihood methods, and direct modeling of the sample distribution. Pseudo-likelihood methods are sensitive to scaling and pairwise likelihood estimation requires knowledge of pairwise sampling probabilities, while uncertainty quantification for pseudo-Bayesian approaches relies on applying ad hoc corrections, which we discuss further in the next chapter. Direct modeling of the sample distribution may necessitate undesirable model

assumptions. As such, more work is needed to understand how best to use unit level covariate modeling in a setting where design optimality properties are prioritized.

Chapter 5

UNIT LEVEL MODELING FOR SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING

5.1 Introduction

When survey data are limited, direct weighted estimators of small area means such as the Horvitz-Thompson (Horvitz and Thompson 1952) or Hájek (Hájek 1971) estimators can be imprecise or unreliable. Statistical models can generate improved estimates by incorporating auxiliary covariate information, explicitly accounting for between area variability using random effects, and leveraging spatial dependence to smooth across nearby areas. Unit level models, especially those incorporating spatial random effects, are commonly used in global health research for mapping subnational health and demographic indicators in low- and middle-income countries (LMIC). These models often account for spatial variation using Gaussian processes that are continuous in space allowing estimates to be generated at arbitrary resolutions. Accounting for survey design features such as unequal sampling probabilities and clustering is crucial and nontrivial when using unit level models to map demographic indicators using complex survey data. Model-based approaches to small area estimation (SAE) often assume that the sampling design is ignorable, meaning that the distribution of sampled responses will be identical to that of non-sampled responses, so a model for the sampled responses can be used to conduct population inference directly. When the sampling design is not ignorable, it is crucial to account for potential differences between sampled and non-sampled units. One approach is to include all variables used to specify the sampling design, such as stratification variables, as predictors in a model, so that a particular unit's response will be independent

of whether it is sampled after conditioning on relevant design variables. If we can specify such a model, we can say that sampling is uninformative with respect to the model. For example, if a particular survey design involves sampling clusters with probability proportional to size, cluster size may be a relevant variable to include in the model.

In practice, we may only observe a subset of relevant design variables or the functional form of the relationship between the design variables and responses may be unknown, making it difficult to specify a model for which sampling is uninformative. In this chapter, we will say that sampling is informative with respect to the model if the model does not apply to both sampled and non-sampled units. We aim to address the effects of informative sampling by leveraging other sources of information about the sampling design such as sampling weights.

Sampling weights are commonly used to compute direct weighted estimators of small area quantities such as the Horvitz-Thompson (1952) or Hájek (1971) estimators. Area level models commonly used throughout SAE like the Fay-Herriot model (1979) implicitly account for the survey design by approximating the sampling distributions of these direct weighted estimators. However, when estimating unit level models, addressing informative sampling with sampling weights is less straightforward.

A number of modified approaches have been proposed that incorporate sampling weights when fitting specific unit level models. Rao and Molina (2015) and Parker et al. (2020) review some of these methods for the basic unit level model. These modifications account for some possible design features such as unequal sampling probabilities, but may not explicitly address informative sampling and must be extended for use with non-Gaussian response variables.

More generically, for inference using parametric models with complex survey data, pseudo-likelihood methods (Binder 1983) incorporate sampling weights into the likelihood and can achieve design-consistent estimation of model parameters under certain asymptotic assumptions. Analogously, pseudo-likelihoods can be used in place of true

likelihoods to conduct approximate Bayesian inference using pseudo-posterior distributions (Savitsky and Toth 2016). These pseudo-likelihood methods have been extended to mixed effects models (Asparouhov 2006; Pfeffermann et al. 1998; Rabe-Hesketh and Skrondal 2006), but frequentist estimators for model parameters resulting from these pseudo-likelihood approaches can be sensitive to weight scaling (Savitsky and Williams 2022; Slud 2020). Moreover, for pseudo-Bayesian methods in particular, credible sets for model parameters based on pseudo-posterior distributions do not generally achieve valid frequentist coverage rates, even asymptotically (Han and Wellner 2021; León-Novelo and Savitsky 2019; Williams and Savitsky 2021). In addition, this body of research has generally focused on estimation of fixed effects, treating the random effects as nuisance parameters. In the context of SAE, prediction of the area level random effects is of principal importance. Although previous research has applied pseudo-Bayesian approaches for small area estimation (Parker et al. 2022), the issues of weight scaling and miscalibrated interval estimates have not been explored extensively in the context of SAE.

In this chapter, we outline a strategy for conducting pseudo-Bayesian inference for small area means using unit level models. As pseudo-Bayesian credible sets for model parameters may not converge on valid frequentist confidence sets due to dependence between units and informative sampling, we adapt a post-processing method proposed by Williams and Savitsky (2021) to rescale our credible sets for small area means. In simulations that we report, the rescaled interval estimates achieve close to nominal empirical coverage rates. We apply our strategy for estimating small area means of both continuous and binary response variables.

The rest of this chapter is organized as follows. In Section 5.2, we outline our notation and describe the combined model- and design-based inferential framework we use to assess our estimators. Section 5.3 reviews standard estimation approaches for unit level models using sampling weights and Section 5.4 details our pseudo-Bayesian approach (and post-

processing method) for generating point and interval estimates of small area means. In Section 5.5, we evaluate the performance of our approach in simulation, and in Section 5.6, we apply our method to estimate vaccination rates using data from the Demographic and Health Surveys. Finally, in Section 5.7, we discuss our method in context and outline directions for future research.

5.2 Background and inferential framework

5.2.1 Notation

Let $U = \{1, \ldots, N\}$ index a finite population of size N. For all $j \in U$, we let y_j denote the response value of interest for unit j and $\mathbf{z}_j$ denote a vector of auxiliary variables. We assume U can be partitioned into m disjoint administrative areas, $U = U(1) \cup \cdots \cup U(m)$, where $U(i)$ denotes the $N(i)$ indices corresponding to units in area i. Let $S = \{j_1, \ldots, j_n\} \subset U$ denote a random set of n sampled indices, where $S = S_1 \cup \cdots \cup S_m$ is the corresponding partition by administrative area.

We assume a probability sampling scheme where for all $j \in U$, π_j denotes the probability that $j \in S$, also called the inclusion probability of unit j, which may depend on $\mathbf{z}_j$. For all $j \in U$, we define δ_j to be the inclusion indicator for unit j. In other words, $\delta_j = 1$ if $j \in S$ and $\delta_j = 0$ otherwise. We let $w_j = 1/\pi_j$ denote the sampling weight for unit j (defined as the inverse inclusion probability).

Following Rao and Molina (2015), we let $y_{ij} = y_j$ if $j \in U(i)$ and $y_{ij} = 0$ otherwise. We define δ_{ij}, π_{ij}, w_{ij}, and $\mathbf{z}_{ij}$ analogously. We define $\mathbf{Z}$ to be the matrix of auxiliary variables. The small area means $\overline{\mathbf{Y}} = \{\overline{Y}_1, \ldots, \overline{Y}_m\}$ can be defined such that for each i,

$$\overline{Y}_i = \frac{1}{N(i)} \sum_{j \in U(i)} y_{ij}.$$

5.2.2 *Unit level modeling for small area estimation*

Unit level modeling approaches to SAE relate individual survey responses y_{ij} to unit-specific auxiliary information and borrow strength from similar or nearby areas when estimating a small area quantity. For continuous responses, Battese, Harter, and Fuller (1988) introduced the nested error regression model (also called the basic unit level model by Rao and Molina (2015)):

$$y_{ij} = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i + \varepsilon_{ij} \tag{5.1}$$

where β_0 denotes an intercept term, $\mathbf{x}_{ij}$ denotes observed covariate values, and $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)$ denotes the corresponding coefficients. We assume that $\mathbf{x}_{ij}$ corresponds to a subset of the variables included in $\mathbf{z}_{ij}$, allowing for the possibility that not all relevant variables used to design the survey are observed. The area level effects are denoted $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ represent random and independent unit level effects. For binary or count responses, other models can be used for y_{ij}.

Under this model, $\overline{Y}_i = \beta_0 + \overline{\mathbf{x}}_i^T \boldsymbol{\beta}_1 + u_i + \overline{\varepsilon}_i$ where $\overline{\mathbf{x}}_i$ and $\overline{\varepsilon}_i$ denotes the area means of $\mathbf{x}_{ij}$ and ε_{ij}, respectively. From a model-based perspective, if we view ε_{ij} as representing noise or measurement error added to the true quantity of interest for individual j, then a more appropriate target estimand is

$$\mu_i = E(\overline{Y}_i \mid \overline{\mathbf{x}}_i, u_i) = \beta_0 + \overline{\mathbf{x}}_i^T \boldsymbol{\beta}_1 + u_i.$$

Another justification for using μ_i instead of $\overline{Y}_i$ as the target estimand is that even if we view y_{ij} as being measured without error, by the law of large numbers, $\overline{\varepsilon}_i$ converges in probability to $E(\varepsilon_{ij}) = 0$ as $N(i) \to \infty$, so it is standard in the SAE literature to focus on estimation of μ_i.

5.2.3 *Interpretation of the basic unit level model*

Practically speaking, treating the area specific intercepts u_i as Gaussian random effects explicitly models variability between areas and shrinks small area mean estimates towards

$\beta_0 + \bar{\mathbf{x}}_i^T \boldsymbol{\beta}_1$. Traditionally, the observed values of a random effect such as u_i are viewed as draws from some population, but only population characteristics (i.e. averaged over u_i), and not the draws themselves, are of interest. Hodges (2016) calls random effects interpreted in this way "old-style" to distinguish them from "new-style" random effects, which may represent the entire population of interest or may represent draws from some distribution from which additional draws cannot be obtained. Under the model (5.1), u_i represent area-specific deviations from the global mean that are of interest and in small area estimation, we are typically interested in estimates for a fixed set of m areas, so we interpret the random effects in the nested error regression model as "new-style" effects. From this perspective, incorporating area-specific random effects produces a flexible model constrained by the Gaussian assumption on u_i, preventing overfitting when data are limited in some or all of the areas.

When specifying a population level model, however, it may be undesirable to use a model of the form (5.1) including random effects. Generally we are interested in between-area differences that are stable in an asymptotic sense instead of random deviations u_i that may vary across populations. Note that we can rewrite the nested error regression model as follows:

$$y_{ij} = \beta_{0i} + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + \varepsilon_{ij} \tag{5.2}$$

where area specific intercepts $\beta_{0i} = \beta_0 + u_i$ are independent $N(\beta_0, \sigma_u^2)$ variables and now $\mu_i = \beta_{0i} + \bar{\mathbf{x}}_i^T \boldsymbol{\beta}_1$. Under a frequentist approach, instead of treating the β_{0i} as draws from a Gaussian distribution, we could view them as fixed area specific intercepts, which would make μ_i fixed across populations after conditioning on auxiliary variables $\mathbf{X}$. Given sufficient data in each area, it could be sensible to use a model with only fixed effects to avoid shrinkage. From this perspective, β_{0i} account for stable population level differences between areas that are not explained by differences in the available predictors $\mathbf{x}_{ij}$. From a Bayesian hierarchical modeling perspective, the difference between using "fixed" effects versus random effects is less salient: the β_{0i} parameters are always treated as random and

the shift only involves a change in the prior on β_{0i}.

5.2.4 Joint model-and-design based inference

When conducting inference for model parameters such as μ_i, it is common to assume that sampling is uninformative with respect to the model in question. In other words, the model is assumed to hold for both the sample and population. In practice, this assumption is difficult to verify, as the model must accurately describe the functional form of the relationship between $\mathbf{x}_{ij}$ and y_{ij}, including possible interaction terms.

If the model is misspecified for the sample data, then model-based estimators need to be adjusted. For example, Pfefferman and Sverchkov directly address this by modeling sample inclusion mechanism (Pfeffermann and Sverchkov 2007). Another approach is to use the population level model to define "census" model-based estimators for model parameters that could be computed given complete population data (Binder 1983; Lumley and Scott 2017). Traditional design-based sample estimators that utilize sampling weights can subsequently approximate these census estimators. This inferential approach accounts for model-based variability in the census estimators and design-based variability in the sample estimators.

We study point and interval estimators of μ_i under this combined model- and design-based framework, as developed by Rubin-Bleuer and Kratina (2005) and also examined by Williams and Savitsky (2021) and Han and Wellner (Han and Wellner 2021). We consider a sequence of sampling designs and populations indexed by ν. Let $U_\nu = \{1, \ldots, N_\nu\}$ index a finite population of size N_ν, where N_ν increases in ν. Let $\mathcal{S}_\nu$ be the collection of all possible subsets of U_ν.

Let $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ and $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ be measurable spaces for the response and auxiliary variables. Han and Wellner assume $\{(Y_j, \mathbf{Z}_j) \in \mathcal{Y} \times \mathcal{Z}\}_{j=1}^{N_\nu}$ are independent and identically distributed random vectors drawn from a superpopulation model on the probability space

$(\Omega, \mathcal{F}, \mathbb{P}_{M_0}) \equiv (\mathcal{Y} \times \mathcal{Z}, \mathcal{B}_{\mathcal{Y}} \times \mathcal{B}_{\mathcal{Z}}, \mathbb{P}_{M_0})$ where $\mathbb{P}_{M_0}$ denotes the superpopulation measure and θ^* denotes a vector of hyperparameters.

For cluster and multistage designs, we may wish to consider more complicated dependence structures for the superpopulation model. As an example, Rubin-Bleuer and Kratina outline a two-stage super-population model under which the population is partitioned into primary sampling units (PSU), within which responses and auxiliary variables may be dependent. The above notation may be adapted to reflect this dependence structure where $\{(Y_j, \mathbf{Z}_j)\}_{j=1}^{N_\nu}$ are organized into groups of final-stage sampling units that are independent from one another. In order to simplify the exposition, we continue the discussion treating $\{(Y_j, \mathbf{Z}_j)\}_{j=1}^{N_\nu}$ as independent.

Conditionally on $\mathbf{Z}^{(\nu)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{N_\nu})$, we can define a sampling design $\mathbb{P}_{D_\nu}$, which we view as a probability distribution over the space of possible samples $S \in \mathcal{S}_\nu$. We let $\{\mathbf{Y}^{(\nu)}, \mathbf{Z}^{(\nu)}, \boldsymbol{\delta}^{(\nu)}\}$ denote the data for the νth finite population where $\boldsymbol{\delta}^{(\nu)}$ denotes the vector of sample inclusion indicators. As outlined by Rubin-Bleuer and Kratina, we can construct a product measurable space $(\mathcal{S}_\nu \times \Omega, \sigma(\mathcal{S}_N) \times \mathcal{F}, \mathbb{P})$ where $\sigma(\mathcal{S}_N)$ is the σ-algebra generated by $\mathcal{S}_N$. For convenience, we use P_0 to denote the marginal distribution of Y.

We seek to understand the asymptotic behavior of our estimators as $\nu \to 0$ under the combined probability measure $\mathbb{P}$. Similarly, when evaluating estimators, we will generally consider average error metrics taking expectations with respect to both $\mathbb{P}_{M_0}$ and $\mathbb{P}_{D_\nu}$. Under informative sampling, we consider a combined model-and-design based mean squared error for evaluating point estimators:

$$\mathrm{MSE}(\widehat{\mu}_i) = \mathbb{E}_{M_0, D_\nu}[(\widehat{\mu}_i - \mu_i)^2]$$

where $\mathbb{E}_{M_0, D_\nu}(\cdot) \equiv \mathbb{E}_{D_\nu}[\mathbb{E}_{M_0}(\cdot)]$. We are also interested in identifying interval estimates $(\widehat{\mu}_i^-, \widehat{\mu}_i^+)$ such that

$$\mathbb{P}\left(\mu_i \in (\widehat{\mu}_i^-, \widehat{\mu}_i^+)\right) = 1 - \alpha$$

for some pre-specified level α, where $\mathbb{P}$ indicates the joint probability measure.

This framework assumes a true population-generating model determining $\mathbb{P}_{M_0}$ and P_{θ^*}. However, the population-generating model does not need to match the model used by the data analyst to motivate the estimator, which may not even correctly describe the dependence between responses. We primarily consider estimators based on nested error regression models of the form specified in Equation (5.1). In the Appendix, we discuss the impact of model misspecification, finding that even when the model is misspecified (for example, if relevant covariates are missing), including area-specific mean parameters in the model can yield reasonable small area estimators.

5.3 Standard estimation approaches

In this section, we review parameter estimation approaches for the nested error regression model, beginning with approaches which assume ignorability of the sampling design. We proceed to review pseudo-likelihood and pseudo-Bayesian approaches that incorporate sampling weights.

5.3.1 Parameter estimation assuming ignorability

First, we consider the model (5.2) treating β_{0i} as random, under the assumption that sampling is uninformative with respect to the model. As detailed by Rao and Molina (2015), the frequentist approach proceeds by estimating variance components $\{\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\}$ via restricted maximum likelihood or a method of moments. Based on these estimates,the empirical best linear unbiased predictor (EBLUP) $\widehat{\mu}_i^{EBLUP}$ can be computed for all i. Either linearization-based approximation or resampling methods can be used to estimate the MSE of $\widehat{\mu}_i^{EBLUP}$. Prediction intervals can be constructed around the EBLUP based on the asymptotic distribution of $\widehat{\mu}_i^{EBLUP} - \mu_i$.

The model (5.1) can be reframed as a Bayesian hierarchical model by placing a known

prior distribution g (which may depend on hyperparameters τ) on the parameters θ:

$$\begin{aligned} y_{ij} \mid \beta_0, \boldsymbol{\beta}_1, u_i, \sigma_\varepsilon^2 &\sim N(\beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i, \sigma_\varepsilon^2) \\ u_i \mid \sigma_u^2 &\sim N(\mathbf{0}, \sigma_u^2) \\ \theta = (\beta_0, \boldsymbol{\beta}_1, \sigma_u^2, \sigma_\varepsilon^2) &\sim g(\theta). \end{aligned} \tag{5.3}$$

Under this model, our targets are the posterior distributions $p(\mu_i \mid \mathbf{Y}^{(\nu)}, \mathbf{X}^{(\nu)}, \boldsymbol{\delta}^{(\nu)})$. Using samples from this posterior, where we denote the k-th sample using $\widehat{\mu}_i^{(k)}$, we can compute posterior summary statistics and credible intervals for μ_i.

5.3.2 Parameter estimation under informative sampling

Unless sampling probabilities are constant within areas, the EBLUP for the model (5.2) is not generally design-consistent (Rao and Molina 2015). To address this, You and Rao (2002) propose a pseudo-EBLUP method for unequal probability sampling designs that incorporates sampling weights when estimating regression coefficients $\boldsymbol{\beta}_1$. This approach is not intended to address general informative sampling, though it can do so in many cases. As for the standard EBLUP, the variance components σ_u^2 and σ_ε^2 can be estimated via restricted maximum likelihood. Subsequently the fixed effects parameters are obtained by solving weighted estimating equations.

The resulting parameter estimates are used to predict u_i and compute a pseudo-EBLUP $\widehat{\mu}_i^{psEBLUP}$. The MSE of $\widehat{\mu}_i^{psEBLUP}$ can be estimated via linearization-based approximations (You and Rao 2002) or resampling (Torabi and Rao 2010).

More generically, given a parametric model, pseudo-likelihood methods incorporate survey weights to construct a sample weighted log-likelihood that approximates the full population log-likelihood (Binder 1983; Skinner 1989). Instead of attempting to incorporate all relevant design features when specifying a model likelihood p_θ, pseudo-likelihood methods propose a particular superpopulation model of interest that could be fit given

full population data. The pseudo-likelihood is subsequently used to approximate complete population inference for superpopulation parameters using sampling weights. If the weights contain information about informative sampling that cannot otherwise be easily incorporated into a regression model or prediction algorithm, then these approaches may yield estimates with reduced bias.

If the population were fully observed, the census log-likelihood would take the form:

$$\ell(\theta; \mathbf{Y}) = \sum_{j=1}^{N_\nu} \log p_\theta(y_j) \tag{5.4}$$

where p_θ denotes the likelihood of y_{ij} given parameters θ. The census log-likelihood may be approximated via a sample weighted pseudo-log-likelihood (Binder 1983):

$$\ell^\pi(\theta; \mathbf{Y}) = \sum_{j=1}^{N_\nu} \frac{\delta_{\nu j}}{\pi_{\nu j}} \log p_\theta(y_j) \tag{5.5}$$

where $\delta_{\nu j}$ and $\pi_{\nu j}$ denote the inclusion indicator and inclusion probability for unit j for the ν-th population. The census log-likelihood and pseudo-log-likelihood may be used to derive census estimating equations and analogously, sample weighted estimating equations. The weighted estimating equations can be used to derive maximum pseudo-likelihood estimates of θ. More generally, similar estimating equations can be used to estimate any finite population parameter of interest that can be specified as a solution to a system of census estimating equations, even without some motivating superpopulation model.

The pseudo-log-likelihood implies a pseudo-likelihood of the form

$$\prod_{j=1}^{N_\nu} p_\theta(y_j)^{w_{\nu j}} = \prod_{j=1}^{N_\nu} p_\theta(y_j)^{\delta_{\nu j}/\pi_{\nu j}}. \tag{5.6}$$

The pseudo-likelihood is not a true likelihood due to the introduction of the weights, but by treating it as such, pseudo-Bayesian inference can be conducted for θ, as introduced by Savitsky and Toth (2016). Parker et al. (2020) provide an example of pseudo-Bayesian inference applied for SAE, but do not explicitly address uncertainty quantification, yielding credible intervals for parameters that exhibit undercoverage.

If the entire population were observed, the population posterior for θ could be defined as follows, for all measurable subsets $B \subset \Theta$:

$$\Pi_\nu(B \mid \mathbf{Y}^{(\nu)}) = \frac{\int_B \prod_{j=1}^{N_\nu} p_\theta(y_j) g_\nu(\theta) d\theta}{\int \prod_{j=1}^{N_\nu} p_\theta(y_j) g_\nu(\theta) d\theta} = \frac{\int_B \exp(N_\nu \mathbb{P}_\nu \log p_\theta) g_\nu(\theta) d\theta}{\int \exp(N_\nu \mathbb{P}_\nu \log p_\theta) g_\nu(\theta) d\theta} \tag{5.7}$$

where $g_\nu(\theta)$ denotes a prior on the hyperparameters θ and $\mathbb{P}_\nu$ denotes the empirical measure based on the ν-th population:

$$\mathbb{P}_\nu(t) = \frac{1}{N_\nu} \sum_{j=1}^{N_\nu} t(Y_j) \tag{5.8}$$

where t denotes a measurable real-valued function. When only a sample of size n_ν is observed, the population posterior distribution can be approximated by a pseudo-posterior distribution replacing the population likelihood with the pseudo-likelihood:

$$\Pi^\pi_\nu(B \mid \mathbf{Y}^{(\nu)}, \boldsymbol{\delta}^{(\nu)}) = \frac{\int_B \prod_{j=1}^{N_\nu} p_\theta(y_j)^{\delta_{\nu j}/\pi_{\nu j}} g_\nu(\theta) d\theta}{\int \prod_{j=1}^{N_\nu} p_\theta(y_j)^{\delta_{\nu j}/\pi_{\nu j}} g_\nu(\theta) d\theta} \tag{5.9}$$

$$= \frac{\int_B \exp(N_\nu \mathbb{P}^\pi_\nu \log p_\theta) g_\nu(\theta) d\theta}{\int \exp(N_\nu \mathbb{P}^\pi_\nu \log p_\theta) g_\nu(\theta) d\theta} \tag{5.10}$$

where $\mathbb{P}^\pi_\nu$ is the sample weighted empirical measure for measurable t:

$$\mathbb{P}^\pi_\nu(t) = \frac{1}{N_\nu} \sum_{j=1}^{N_\nu} \frac{\delta_{\nu j}}{\pi_{\nu j}} t(Y_j). \tag{5.11}$$

Note that Π^π_ν is not a standard posterior distribution due to the introduction of sampling weights, but is scaled to integrate to one. In this sense, inference based on the pseudo-posterior distribution can be viewed as approximating inference based on the population posterior. As with the unweighted posterior, we can draw samples from the pseudo-posterior for θ and accordingly obtain estimates of μ_i.

Intuitively, pseudo-posterior credible sets for a superpopulation parameter can be viewed as approximations of the corresponding population posterior credible sets, which would

be based on the full population of size N_ν. In general, credible sets based on pseudo-posterior samples will thus be too conservative. León-Novelo and Savitsky (2019) observe undercoverage of the credible sets based on pseudo-posterior samples. Various solutions have been proposed for this problem, including rescaling of weights and post-processing of pseudo-posterior samples.

For pseudo-likelihood based approaches, sampling weights are often rescaled so that the weights w_{ij} sum to the sample size or to the "effective" sample size, defined as the sample size for a simple random sample achieving the same variance for an estimator as with the existing design (Kish 1965). Rescaling methods are discussed for a frequentist multilevel model (Pfeffermann et al. 1998) and in the pseudo-Bayesian setting (Savitsky and Toth 2016).

5.4 Proposed approach

In this section, we describe a general pseudo-Bayesian approach for small area estimation using unit level models. We apply the post-processing rescaling method described by Williams and Savitsky (2021) to correct the coverage of credible sets for $\boldsymbol{\mu}$ based on pseudo-posterior samples. Although pseudo-Bayesian approaches have previously been adopted for SAE, they have generally been applied on an ad hoc basis and do not explicitly address miscalibration of the pseudo-posterior credible sets. We assume a sequence of sampling designs such that as $\nu \to \infty$, $n(i) \to \infty$ for all areas i. Under this asymptotic framework, as $\nu \to \infty$, direct estimators of small area means become more reliable.

Han and Wellner (2021) and Williams and Savitsky (2021) establish results on the asymptotic behavior of the pseudo-posterior distribtion and the pseudo-maximum likelihood estimator (pseudo-MLE) under certain regularity conditions. In particular, they establish Bernstein-von Mises type results for the pseudo-posterior distribution and derive the asymptotic sampling distribution of the pseudo-MLE. Both of these distributions are asymptotically normal and concentrate on θ^*, the parameter vector minimizing the

Kullback-Leibler divergence $\theta \mapsto P_0 \log(p_0/p_\theta)$. However, their asymptotic covariances do not agree, so credible intervals based on pseudo-posterior distributions will not generally converge on valid frequentist confidence intervals. Note that neither Han and Wellner nor Williams and Savitsky explicitly addresses misspecification of the superpopulation model but both rely upon results of Kleijn and van der Vaart (2012), which establishes a Bernstein-von Mises result for misspecified Bayesian models that illustrates the posterior's concentration on θ^*. In the Appendix, we adapt the conditions of Han and Wellner and describe how their results can be applied for this small area estimation context.

5.4.1 Computing a pseudo-posterior

We describe our approach to pseudo-Bayesian inference for the hierarchical model (5.3) before applying our strategy to other models. Under the hierarchical model, our parameters of interest are $\theta = (\beta_0, \boldsymbol{\beta}_1, \sigma_u^2, \sigma_\varepsilon^2)$ and $\mathbf{u} = (u_1, \ldots, u_m)$, so our goal is to approximate the joint population posterior density:

$$p(\mathbf{u}, \theta) \propto \prod_{i=1}^{m} \prod_{j \in U(i)} p_\theta(y_{ij} \mid u_i) p_\theta(u_i) g(\theta) \tag{5.12}$$

where $p_\theta(y_{ij} \mid u_i)$ denotes the density for response y_{ij} given area effect u_i and parameter vector θ, $p_\theta(u_i)$ denotes the density for the area effect, and $g(\theta)$ denotes the prior. To approximate this population posterior, we use the following sampling-weighted pseudo-posterior density:

$$p^\pi(\mathbf{u}, \theta) \propto \prod_{i=1}^{m} \prod_{j \in U(i)} p_\theta(y_{ij} \mid u_i)^{\delta_{ij}/\pi_{ij}} p_\theta(u_i) g(\theta) \tag{5.13}$$

We can approximate this density via sampling algorithms or numerical approximation and use this pseudo-posterior to conduct inference for $\widehat{\mathbf{u}}$, $\widehat{\theta}$, and subsequently $\widehat{\mu}_i$.

5.4.2 *Post-processing adjustment*

Generalized posteriors produced by replacing a standard likelihood in a Bayesian analysis with a pseudo-likelihood are not expected to quantify parameter uncertainty accurately as pseudo-likelihoods are not generally true likelihoods (Miller 2021; Ribatet et al. 2012). In a complex survey sampling context, both Williams and Savitsky (2021) and Han and Wellner (2021) consider pseudo-Bayesian inference, noting that "vanilla" credible sets for model parameters based on a pseudo-posterior distribution do not generally converge on valid frequentist confidence intervals. The need to rescale pseudo-posterior distributions for pairwise likelihood analysis with survey data is also discussed by Thompson et al. (2022).

Given a parametric model and superpopulation with KL-divergence minimizing parameters θ^* and log-likelihood $\ell_{\theta^*} = \log p_{\theta^*}$, Williams and Savitsky (2021) observe that under certain assumptions for the sampling design and the model likelihood, the pseudo-MLE, defined as the the estimator obtained by maximizing the frequentist pseudo-likelihood and denoted $\hat{\theta}^{\pi}_{\nu}$, is asymptotically Gaussian. In particular, $\sqrt{N_\nu}(\hat{\theta}^{\pi}_{\nu} - \theta^*)$ converges asymptotically to a Gaussian random variable with mean 0 and variance $H^{-1}_{\theta^*} J^{\pi}_{\theta^*} H^{-1}_{\theta^*}$ where H_{θ^*} is the Fisher information:

$$H_{\theta^*} = -\frac{1}{N_\nu} \sum_{j \in U_\nu} \mathbb{E}_{P_{\theta^*}} \ddot{\ell}_{\theta^*}(Y_j, \mathbf{X}_j) \tag{5.14}$$

and $J^{\pi}_{\theta^*}$ is the variance matrix of the score functions under the combined measure $\mathbb{P}$:

$$J^{\pi}_{\theta^*} = \mathbb{E}_{\theta^*, \nu} \left[\mathbb{P}^{\pi}_{\nu} \dot{\ell}_{\theta^*} \dot{\ell}^T_{\theta^*} \right] \tag{5.15}$$

Moreover, under their set of regularity conditions, Williams and Savitsky derive the asymptotic distribution of the pseudo-posterior:

$$\sup_B \left| \Pi^{\pi}_{\nu}(B \mid \mathbf{Y}^{(\nu)}, \mathbf{X}^{(\nu)}, \boldsymbol{\delta}^{(\nu)}) - \mathcal{N}_{\hat{\theta}^{\pi}_{\nu}, N_\nu^{-1} H^{-1}_{\theta^*}}(B) \right| \to 0 \tag{5.16}$$

where $\hat{\theta}^{\pi}_{\nu}$ is the pseudo-MLE and H_{θ^*} is the Fisher information, as defined in (5.14). Here, $\mathcal{N}_{\hat{\theta}^{\pi}_{\nu}, N_{\nu}^{-1} H_{\theta^*}^{-1}}$ indicates the probability measure for a multivariate Gaussian distribution centered on $\hat{\theta}^{\pi}_{\nu}$ with covariance matrix $N_{\nu}^{-1} H_{\theta^*}^{-1}$.

Based on the differing forms of the covariance matrices for the pseudo-MLE and pseudo-posterior, Williams and Savitsky propose adjusting samples as follows:

$$\widehat{\theta}^{WS(k)} = \left(\widehat{\theta}^{(k)} - \overline{\theta} \right) R_2^{-1} R_1 + \overline{\theta} \tag{5.17}$$

where $R_1^T R_1 = H_{\theta^*}^{-1} J_{\theta^*}^{\pi} H_{\theta^*}^{-1}$ is the asymptotic covariance of the pseudo-MLE and $R_2^T R_2 = H_{\theta^*}^{-1}$ is the asymptotic covariance of the pseudo-posterior. Here, $\widehat{\theta}^{(k)}$ is the k-th sample from the pseudo-posterior and $\hat{\theta}^{WS(k)}$ the k-th adjusted sample. Finally, $\overline{\theta}$ is the mean of the pseudo-posterior draws, but could be replaced with the pseudo-MLE. The authors call $R_2^{-1} R_1$ a multivariate design effect adjustment, which will vanish for a SRS.

In practice, H_{θ^*} is estimated as the observed information, i.e. the negative Hessian of the weighted log-likelihood at the pseudo-MLE. Following Williams and Savitsky, we estimate $J_{\theta^*}^{\pi}$ via a resampling approach (Preston 2009) that seeks to estimate the variance of the score functions by sampling PSUs with replacement from the sample. We use numerical differentiation when estimating both H_{θ^*} and $J_{\theta^*}^{\pi}$. Further details are provided in the Appendix.

5.4.3 Rescaling small area estimates

Under the hierarchical model (5.3), this rescaling approach enables us to produce credible sets for the parameter vector $\theta = (\beta_0, \boldsymbol{\beta}_1, \sigma_u^2, \sigma_\varepsilon^2)$ that converge on asymptotically correct frequentist confidence sets for the true model parameters. However the interpretation is less clear for rescaled credible sets of area level random quantities such as u_i. For the purpose of small area estimation, we are interested in between area variations that are stable as $\nu \to \infty$. As such, we propose a strategy that rescales the pseudo-posterior

distributions for $\beta_{0i} = \beta_0 + u_i$ based on the asymptotic distributions of the pseudo-MLEs resulting from likelihood analysis of the model treating the β_{0i} parameters as fixed effects.

In practice, for $k = 1, \ldots, K$, we draw samples $\beta_0^{(k)}, \boldsymbol{\beta}_1^{(k)}, \sigma_u^{2(k)}, \sigma_\varepsilon^{2(k)}$, and $\mathbf{u}^{(k)}$ from the pseudo-posterior distribution based on the hierarchical model (5.3). We can then transform these samples to express them in terms of the parameters of the model (5.2) with fixed area-specific intercepts, yielding a sample vector $\widehat{\theta}^{(k)} = (\beta_{0i}^{(k)}, \boldsymbol{\beta}_1^{(k)}, \sigma_\varepsilon^{2(k)})$. We then estimate the rescaling matrices H_{θ^*} and $J_{\theta^*}^{\pi}$ using the likelihood arising from (5.2), treating β_{0i} as fixed parameters. In other words, the model we use for rescaling is a fixed effects model and the asymptotic distribution of the pseudo-MLE is based on a model that treats β_{0i} as stable across populations. From a Bayesian standpoint, this perspective shift is natural as the distinction between fixed and random effects is less salient: we are simply defining a hierarchical prior on the parameters of interest u_i. As $\nu \to \infty$, the effect of the prior disappears and both the pseudo-posterior and pseudo-MLE converge upon the same parameter vectors.

5.5 Simulations

To assess the performance of our pseudo-Bayesian approach we carry out a simulation study using a range of population models and sampling designs. For each choice of population model, we generate a single finite population of responses. Using each design, we repeatedly sample a subset of responses and then compute estimators of the small area means μ_i.

5.5.1 Population generating models

We carry out simulations for populations of continuous response data and binary response data generated using the models described below. For both response models, we first generate auxiliary variables for a clustered population letting $\mathbf{z}_{icj}$ denote the aux-

iliary variables for individual j in cluster c in area i. We assume that each unit belongs to one cluster and clusters are nested within areas. Finally, we assume area i contains $N_C(i)$ clusters indexed by the set $C(i) = \{c_{i_1}, \ldots, c_{i_{N_C(i)}}\}$, $j = 1, \ldots, N_{ic}$.

1. We generate $z_{1icj} \overset{ind}{\sim} N\left(\frac{i}{m}, 1\right)$, where $1 \leq i \leq m$ indexes the area, so the mean of z_{1icj} varies across areas.

2. We generate $z_{2icj} = \frac{i}{m} + z'_{2icj}$, where $z'_{2icj} \overset{iid}{\sim} \text{Exp}(1/2)$. The variable z_{2icj} represents a measure of unit size, which we will use to specify a sampling design. We define $z_{2ic\cdot} = \sum_j z_{2icj}$ to be the cluster size obtained by summing the sizes for all units in the relevant cluster. We define the scaled unit size x_{2ij} to be equal to z_{2icj} scaled to have mean zero and variance 1. Similarly, we define the scaled cluster size $\widetilde{x}_{2ij}$ to be equal to $z_{2ic\cdot}$ scaled to have mean zero and variance 1, where c is the cluster containing unit j.

Continuous responses

To generate continuous response data, we simulate data from population models of the form

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 \widetilde{x}_{2ij} + \varepsilon_{ij} \tag{5.18}$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_\varepsilon)$. As described above, $\widetilde{x}_{2ij}$ denotes the cluster size, representing a relevant design variable that is unavailable to the analyst. To estimate the area level means, we fit the following nested error regression model:

$$y_{ij} = \beta'_0 + \beta'_1 x_{1ij} + u_i + \varepsilon'_{ij} = \beta'_{0i} + \beta'_1 x_{1ij} + \varepsilon'_{ij} \tag{5.19}$$

For the purpose of estimating model parameters, we assume that $\varepsilon'_{ij} \overset{iid}{\sim} N(0, \sigma^2_\varepsilon)$. Note that for the estimation model, we use β'_0, β'_1, and ε'_{ij} to denote model parameters to emphasize that this model is misspecified and we cannot expect to obtain consistent

estimators for the true population parameters β_0 and β_1 in general. Based on this estimation model, u_i is used to capture stable between area differences induced by area level differences in the mean value of $\widetilde{x}_{2ij}$.

Binary responses

We also implement simulations for binary response data using the population generating models of the form:

$$\begin{aligned} y_{ij} \mid q_{ij} &\sim \text{Bernoulli}(q_{ij}) \\ q_{ij} &= \text{expit}(\beta_0 + \beta_1 x_{1ij} + \beta_2 \widetilde{x}_{2ij}) \end{aligned} \tag{5.20}$$

We again use $\widetilde{x}_{2ij}$ to denote the unobserved cluster size for individual j. The pseudo-Bayesian approach described above can be adapted to non-Gaussian response data by using alternative likelihoods. To estimate the area level means, we fit the following logistic regression model:

$$\begin{aligned} y_{ij} \mid q_{ij} &\sim \text{Bernoulli}(q_{ij}) \\ q_{ij} \mid u_i &= \text{expit}(\beta_0' + \beta_1' x_{1ij} + u_i) = \text{expit}(\beta_{0i}' + \beta_1' x_{1ij}) \\ u_i &\overset{iid}{\sim} N(0, \sigma_u^2) \end{aligned} \tag{5.21}$$

The areal effects u_i capture between area differences in the log-transformed odds induced by area level differences in the mean value of $\widetilde{x}_{2ij}$. Again, we can either view $\beta_{0i}' = \beta_0' + u_i$ as a fixed area specific intercept term or as a random intercept by placing a Gaussian prior on u_i. Based on this estimation model, the implied target of estimation is defined as

$$\mu_i = \mathbb{E}_{\theta_0,\nu} \text{expit}(\beta_{0i}' + \beta_1' x_{1ij}) \tag{5.22}$$

where the expectation is taken with respect to both the estimation model and design, assuming that the covariate x_{1ij} is random. In practice we compute the following estimator:

$$\widehat{\mu}_i = \sum_{j \in U(i)} \text{expit}(\beta_{0i}' + \beta_1' x_{1ij}) \tag{5.23}$$

Unlike in the continuous case, where only area level means of covariates are needed, for the logistic regression model, we need to know x_{1ij} for all units in the population. In practice, this is rare, so using more coarse covariate information (that discretizes covariate values or uses only area-level covariates) may be necessary. In general, when working with nonlinear models, aggregating unit level model predictions for individual clusters or pixels to obtain area level predictions is difficult and requires careful consideration of the sampling design, especially when cluster-level characteristics lead to heterogeneity in responses within an area (Paige et al. 2022a; b).

5.5.2 Estimation procedure

Based on our estimation models, we consider three approaches for estimating area level means. First, we consider a Bayesian approach ignoring the weights (**Unwt**) and treating sampling as ignorable. For the continuous response case, this method is equivalent to a hierarchical Bayesian version of the nested error regression model. Next, we implement our pseudo-Bayesian approach using the sampling weights, both with (**WtRscl**) and without (**Wt**) the rescaling step described in the previous section. The sampling weights used in this analysis are normalized so that their sum is equal to the observed sample size n. For all of these Bayesian estimators, we compute posterior medians and 90% credible sets. We compare these three Bayesian estimators with two design-based estimators: the **Hájek** estimator and a generalized regression estimator (**GREG**) based on a working model with fixed area specific intercepts. For continuous data, this working model takes the form $y_{ij} = \beta'_{0i} + \beta_1 x'_{1ij} + \varepsilon'_{ij}$. We compute 90% prediction intervals based on the estimated mean squared predictive error of these estimators.

Further detail on the estimation procedures, including descriptions of priors for model hyperparameters and the software used can be found in the Appendix. Code used to produce the results throughout this manuscript can be found on GitHub at https://github.com/peteragao/svyulm.

Since $\widetilde{x}_{2ij}$ is unobserved, our estimation models are misspecified. As noted by Williams and Savitsky (2021), their asymptotic results rely on correct parameterization of the dependence structure for the population. However, the unobserved $\widetilde{x}_{2ij}$ is constant within clusters, and induces cluster dependence in our observations. Simulations by Williams and Savitsky indicate that their proposed rescaling method may be robust to this misspecification. We consider an asymptotic framework in which the number of clusters sampled in each area is increasing, but we are primarily interested in the performance of these estimators in a small sample setting, which will be of practical relevance to SAE.

5.5.3 Sampling designs

We consider different sampling designs which induce dependence between observed response values, some of which are informative with respect to the analyst-specified model. For all designs, we stratify sampling by the m small areas.

1. **Stratified random sampling without replacement (SRS)** Within each area i, we sample $n(i)$ individuals at random without replacement. Under this design, assuming the sampling fraction is small, the design effect is expected to be small, making the effect of incorporating sampling weights during estimation negligible.

2. **Single stage informative sampling (PPS1)** For this design, within each area, we sample $n(i)$ units without replacement, with probability proportional to size $s_{ij} = x_{2ij} - \min(x_{2ij}) + 1$ using Midzuno's method as implemented in the R package `sampling` (Tillé and Matei 2021). This yields an single stage design with unequal sampling probabilities (PPS1) that is informative with respect to the analyst-specified model since s_{ij} is correlated with the unobserved cluster size $\widetilde{x}_{2ij}$.

3. **Two stage informative sampling (PPS2)** Within each area i, we sample $n_C(i)$ clusters without replacement with probability proportional to size $\widetilde{x}_{2ij} - \min(\widetilde{x}_{2ij}) + 1$.

Within each sampled cluster, we sample $n(i,c)$ units with probability proportional to size $s_{ij} = x_{2ij} - \min(x_{2ij}) + 1$. This yields a two stage design with unequal sampling probabilities (PPS2) that is informative with respect to the model since $\widetilde{x}_{2ij}$ is unobserved.

5.5.4 Results

For each data generating model, we generate auxiliary variables for a finite population consisting of $N = 90000$ individuals divided evenly between $m = 20$ areas. Each area is divided into $N_C = 150$ clusters of thirty individuals. Based on this fixed auxiliary data, we repeatedly simulate response data and sample from the resulting population for each sampling design for a total of 1000 simulations. For the continuous response case, we simulate data from the following model:

$$y_{ij} = x_{1ij} + 2\widetilde{x}_{2ij} + \varepsilon_{ij} \tag{5.24}$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0,1)$. For the binary response case, we simulate population data from the following model:

$$\begin{aligned} y_{ij} \mid q_{ij} &\sim \text{Bernoulli}(q_{ij}) \\ q_{ij} &= \text{expit}(x_{1ij} + 2\widetilde{x}_{2ij}) \end{aligned} \tag{5.25}$$

Based on the simulated response data, we compute the finite population area level mean $\overline{Y}_i$ for each area i. In each simulation, we compute point estimates $\widehat{\mu}_i$ as well as 90% interval estimates $(\widehat{\mu}_i^-, \widehat{\mu}_i^+)$ for every Bayesian approach and design-based estimator. For each method, we compute root mean squared error (RMSE) and mean absolute error (MAE). We also compute the emprical coverage of the 90% interval estimates and the mean inter-

val lengths (MIL) across all areas, averaged across all simulations.

$$\text{RMSE}(\widehat{\boldsymbol{\mu}}) = \sqrt{\frac{1}{m}\sum_{a}(\overline{Y}_i - \widehat{\mu}_i)^2} \tag{5.26}$$

$$\text{MAE}(\widehat{\boldsymbol{\mu}}) = \frac{1}{m}\sum_{i}|\overline{Y}_i - \widehat{\mu}_i| \tag{5.27}$$

$$\text{Cov}_{90}(\widehat{\boldsymbol{\mu}}) = \frac{1}{m}\sum_{i}\mathbf{1}\{\overline{Y}_i \in (\widehat{\mu}_i^-, \widehat{\mu}_i^+)\} \tag{5.28}$$

$$\text{MIL}_{90}(\widehat{\boldsymbol{\mu}}) = \frac{1}{m}\sum_{i}(\widehat{\mu}_i^+ - \widehat{\mu}_i^-) \tag{5.29}$$

First, we consider simulations with continuous response data. Table 5.1 provides a summary of simulation results for a small sample setting where $n(i) = 30$ and Table 5.2 provides an analogous summary for a larger sample setting where $n(i) = 100$. For the PPS2 design, we assume that $n(i)/5$ clusters are sampled and five individuals are sampled within each cluster.

Under stratified random sampling, the model-based (Unwt, Wt, and WtRscl) approaches perform similarly and produce point estimates with the lowest RMSE and MAE, and achieve close to nominal interval coverage rates. For the large sample simulations, the model-based estimates perform similarly to the GREG estimator, indicating the reduced shrinkage compared with the small sample case.

Under the informative single stage (PPS1) sampling, in the small sample simulations, the weighted model-based methods again perform best in terms of point estimates, with the weighted and rescaled estimates (WtRscl) yielding slightly wider interval estimates on average. For the large sample simulations, the unweighted and weighted but not rescaled interval estimates exhibit undercoverage while the weighted and rescaled intervals are better calibrated.

Finally, for the informative two stage (PPS2) sampling design, the weighted and rescaled method achieves the best performance in terms of RMSE and MAE as well as calibrated

Table 5.1: Averaged evaluation metrics of estimators of area level means across 1,000 continuous response simulations for SRS, PPS1, and PPS2 designs for a sample size of 30 units per area.

Design	Method	RMSE (x 100)	MAE (x 100)	MIL (x 100)	90% Int. Cov.
SRS	Hájek	38.19	30.44	125.09	89
	GREG	33.54	26.70	109.80	89
	Unwt	32.60	25.99	107.70	90
	Wt	32.59	26.00	107.67	90
	WtRscl	32.59	26.00	104.72	88
PPS1	Hájek	41.39	33.04	133.91	88
	GREG	36.66	29.19	117.37	88
	Unwt	35.85	28.68	109.11	87
	Wt	35.59	28.38	107.50	87
	WtRscl	35.59	28.38	111.98	87
PPS2	Hájek	70.63	56.14	217.43	84
	GREG	67.59	53.74	208.13	84
	Unwt	72.12	57.32	105.10	54
	Wt	64.96	51.63	103.04	58
	WtRscl	64.96	51.63	200.35	84

Table 5.2: Averaged evaluation metrics of estimators of area level means across 1,000 continuous response simulations for SRS, PPS1, and PPS2 designs for a sample size of 100 units per area.

Design	Method	RMSE (x 100)	MAE (x 100)	MIL (x 100)	90% Int. Cov.
SRS	Hájek	20.84	16.66	69.00	90
	GREG	18.26	14.56	60.56	90
	Unwt	18.14	14.45	60.11	90
	Wt	18.14	14.46	60.11	90
	WtRscl	18.14	14.46	59.47	90
PPS1	Hájek	22.72	18.10	74.56	90
	GREG	19.95	15.90	65.39	90
	Unwt	23.16	18.75	60.91	81
	Wt	19.75	15.75	60.10	87
	WtRscl	19.75	15.75	64.25	89
PPS2	Hájek	37.04	29.41	126.05	90
	GREG	35.40	28.13	120.71	90
	Unwt	47.51	38.95	60.47	44
	Wt	34.88	27.72	59.37	61
	WtRscl	34.88	27.72	118.69	89

Table 5.3: Averaged evaluation metrics of estimators of area level means across 1,000 binary response simulations for SRS, PPS1, and PPS2 designs for a sample size of 30 units per area.

Design	Method	RMSE (x 100)	MAE (x 100)	MIL (x 100)	90% Int. Cov.
SRS	Hájek	8.05	6.42	26.41	89
	GREG	7.68	6.11	25.06	88
	Unwt	7.23	5.77	23.10	89
	Wt	7.23	5.77	23.10	89
	WtRscl	7.23	5.78	23.22	89
PPS1	Hájek	8.86	7.03	28.47	87
	GREG	8.45	6.70	26.99	87
	Unwt	7.56	5.97	23.00	88
	Wt	7.81	6.19	23.12	87
	WtRscl	7.83	6.22	25.10	89
PPS2	Hájek	11.79	9.38	35.53	80
	GREG	11.49	9.15	34.28	80
	Unwt	10.75	8.37	22.64	74
	Wt	10.33	8.10	22.85	75
	WtRscl	10.32	8.12	31.57	85

interval coverage. Under this design, the other model-based methods exhibit large undercoverage.

Table 5.3 provides a summary of simulation results for binary response data with sample size $n(i) = 30$ and Table 5.4 provides an analogous summary with sample size $n(i) = 100$. The results from these simulations are similar to those from the continuous case under the SRS and PPS2 designs, with the weighted and rescaled estimates generally producing the best point estimates and interval estimates with close to nominal coverage. Under the PPS1 design, the benefits of rescaling are less clear, but the weighted and rescaled method

Table 5.4: Averaged evaluation metrics of estimators of area level means across 1,000 binary response simulations for SRS, PPS1, and PPS2 designs for a sample size of 100 units per area.

Design	Method	RMSE (x 100)	MAE (x 100)	MIL (x 100)	90% Int. Cov.
SRS	Hájek	4.39	3.48	14.53	90
	GREG	4.18	3.31	13.82	90
	Unwt	4.11	3.26	13.43	90
	Wt	4.11	3.26	13.42	90
	WtRscl	4.11	3.26	13.41	90
PPS1	Hájek	4.87	3.86	15.82	89
	GREG	4.65	3.69	15.04	89
	Unwt	4.57	3.61	13.36	87
	Wt	4.53	3.58	13.42	87
	WtRscl	4.53	3.58	14.59	89
PPS2	Hájek	6.32	5.01	20.78	88
	GREG	6.16	4.89	20.17	88
	Unwt	7.14	5.63	13.25	66
	Wt	5.94	4.70	13.35	75
	WtRscl	5.94	4.70	19.50	89

does not perform significantly worse than the other model-based approaches.

5.6 Application: Vaccination coverage in Guinea

We apply the pseudo-Bayesian approach to estimate measles vaccination coverage for prefectures in Guinea in 2018 based on survey data collected by the Demographic and Health Surveys (DHS) Program. The DHS Program conducts surveys in many LMIC and typically implement a stratified two-stage cluster sampling design. Each country is first divided by its principal administrative regions, usually called Admin-1 regions. Each of these regions is partitioned into urban and rural components. Sampling is stratified by crossing the Admin-1 regions with urban/rural labels. For the first stage of sampling, each stratum is divided into clusters, or enumeration areas (EAs). Within each stratum, a pre-specified number of clusters is sampled with probability proportional to size. In the second stage, a pre-specified number of households is sampled from each selected cluster. Under this sampling design, cluster size is a relevant design variable that is typically not made public but which could be associated with a response of interest.

We desire estimates of subnational vaccination rates for the first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Guinea using data from the 2018 Guinea DHS (2019). This survey interviewed mothers in each selected household and collected vaccination data for their children based on vaccination cards or caregiver recall. We intent to produce estimates for each prefecture, which correspond to subdivisions of Guinea's eight Admin-1 regions. We will refer to these prefectures as secondary administrative divisions, Admin-2 regions. We rely on the boundaries published by Database of Global Administrative Areas (GADM) (2022).

The design for the 2018 DHS was based on a sampling frame created using data from a census conducted in 2017 which identified 9679 clusters, or enumeration areas, divided into 15 strata (from splitting eight Admin-1 areas into urban/rural components minus the entirely urban zone of Conakry). Data were collected from 401 clusters. Unlike in many

other countries, the sampled number of urban clusters is roughly proportional to the total number of urban clusters in the population. Within each cluster, twenty households were sampled at random. DHS Program publishes coordinates for all selected clusters after displacing their locations by small distances to protect privacy. Figure 5.1 provides the Admin-1 and Admin-2 boundaries and displaced EA locations in Guinea for which data were collected. Based on this data, we generate estimates using design-based methods and unit level models. For the unit level models, we use two satellite-derived covariate surfaces based on estimated travel times to cities in 2015 (Weiss et al. 2018) and the intensity of night time lights as observed via satellite imagery in 2016 (WorldPop 2018a). Note that these covariates are themselves estimated using statistical modeling. We also use estimated population counts produced by WorldPop (2020) to create a third covariate classifying pixels as either urban or rural by the highest population pixels in an area to be urban so that the proportion of individuals classified urban equals what is reported in the 2018 Guinea DHS report.

Covariate information for the entire population is required to generate estimates for each individual. However, since complete population data is not available, instead of making a separate prediction for each child, we make predictions for each pixel and aggregate the pixel level predictions to get an area level estimate of the mean outcome of interest. When aggregating, we weight each pixel by its estimated age 1-5 population (WorldPop 2018b). We project all covariate values to the 1km by 1km grid used by WorldPop. Figure 5.2 compares point estimates of measles vaccination rates and Figure 5.3 provides the length of interval estimates among children aged 12–23 months for Guinea's prefectures in 2018. We provide point and interval estimates for all methods in the Appendix. In general, the point estimates produced by all methods are quite similar, but the estimates of uncertainty vary considerably. In general, the unweighted Bayes and weighted but not rescaled Bayes estimates have the shortest prediction intervals, indicating the least uncertainty. The design-based Hájek and GREG approaches produce longer prediction

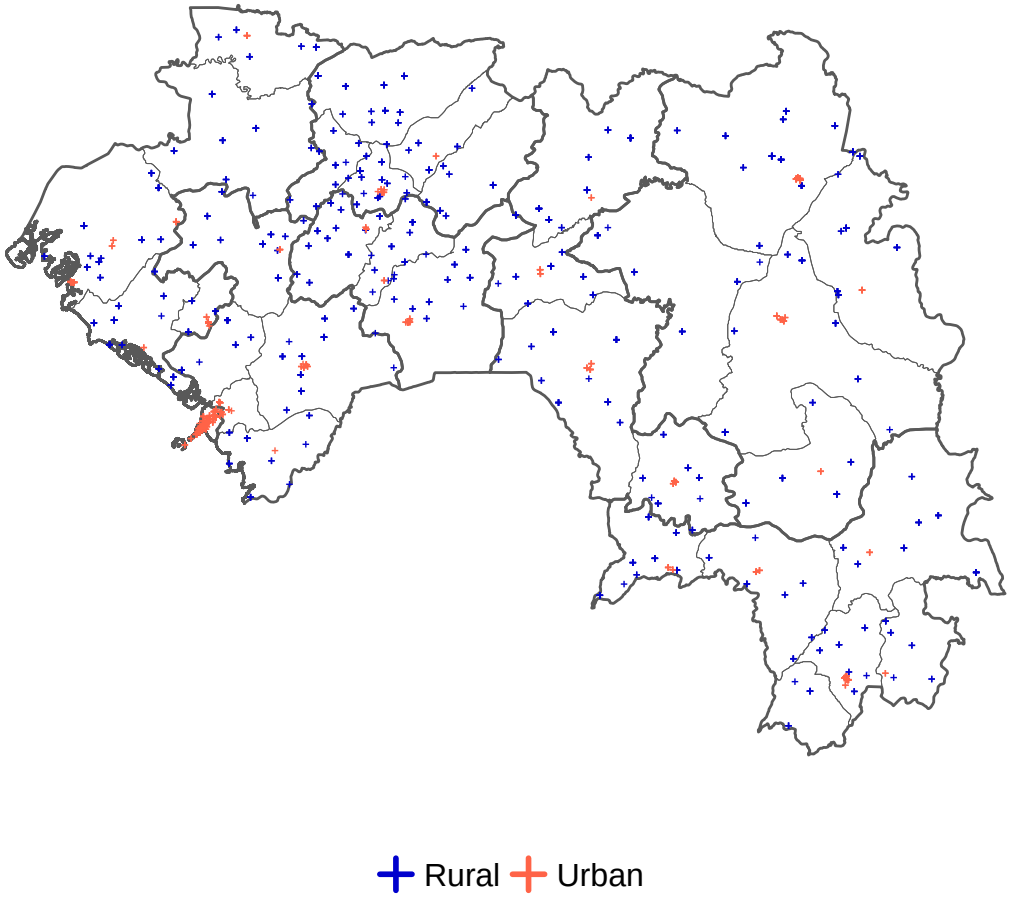

Figure 5.1: Map of Guinea with Admin-1 level boundaries (thick borders) and Admin-2 level boundaries (thin borders). Points indicate enumeration area locations for which data on measles vaccination is available.

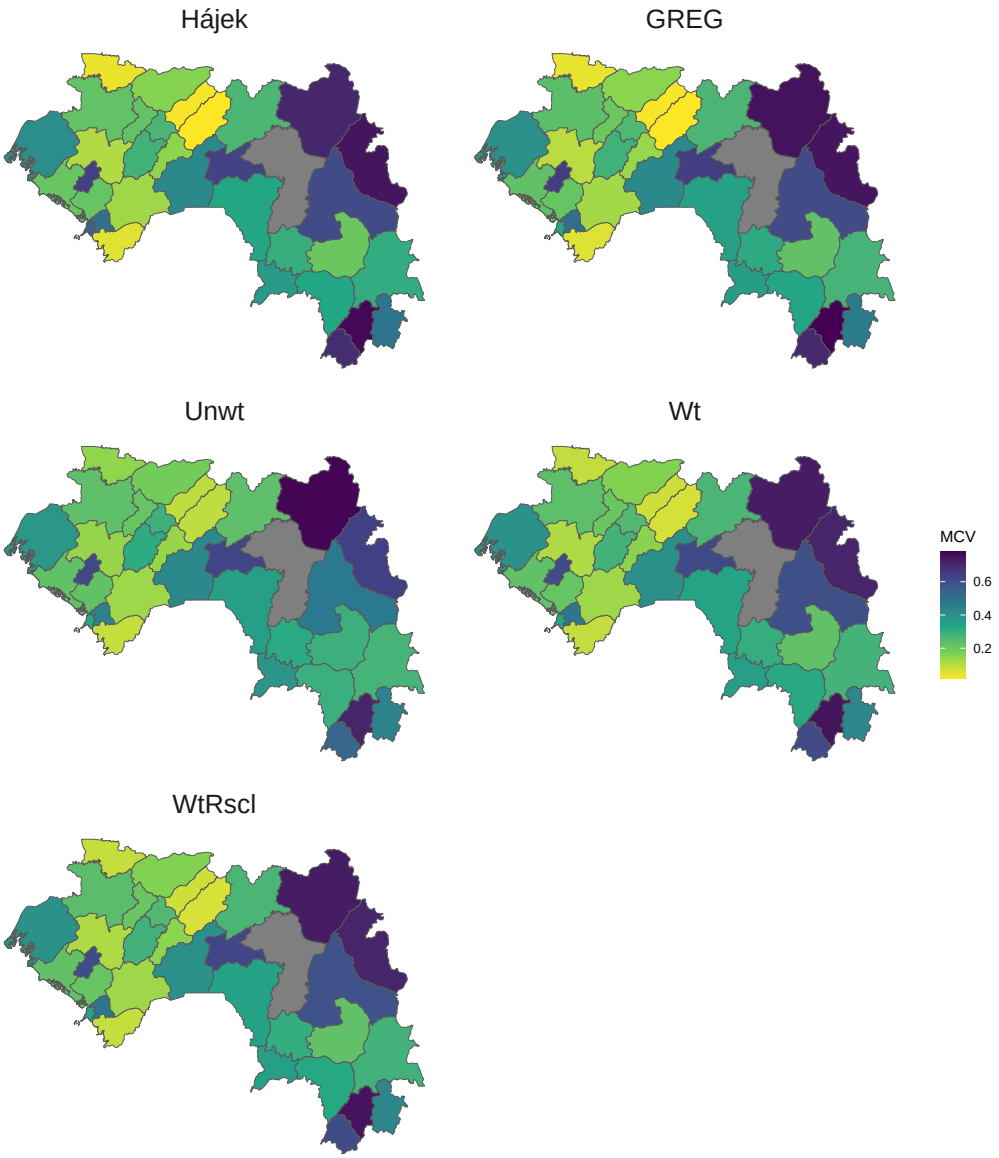

Figure 5.2: Estimated measles vaccination rates among children aged 12–23 months for Admin-2 areas in Guinea in 2018.

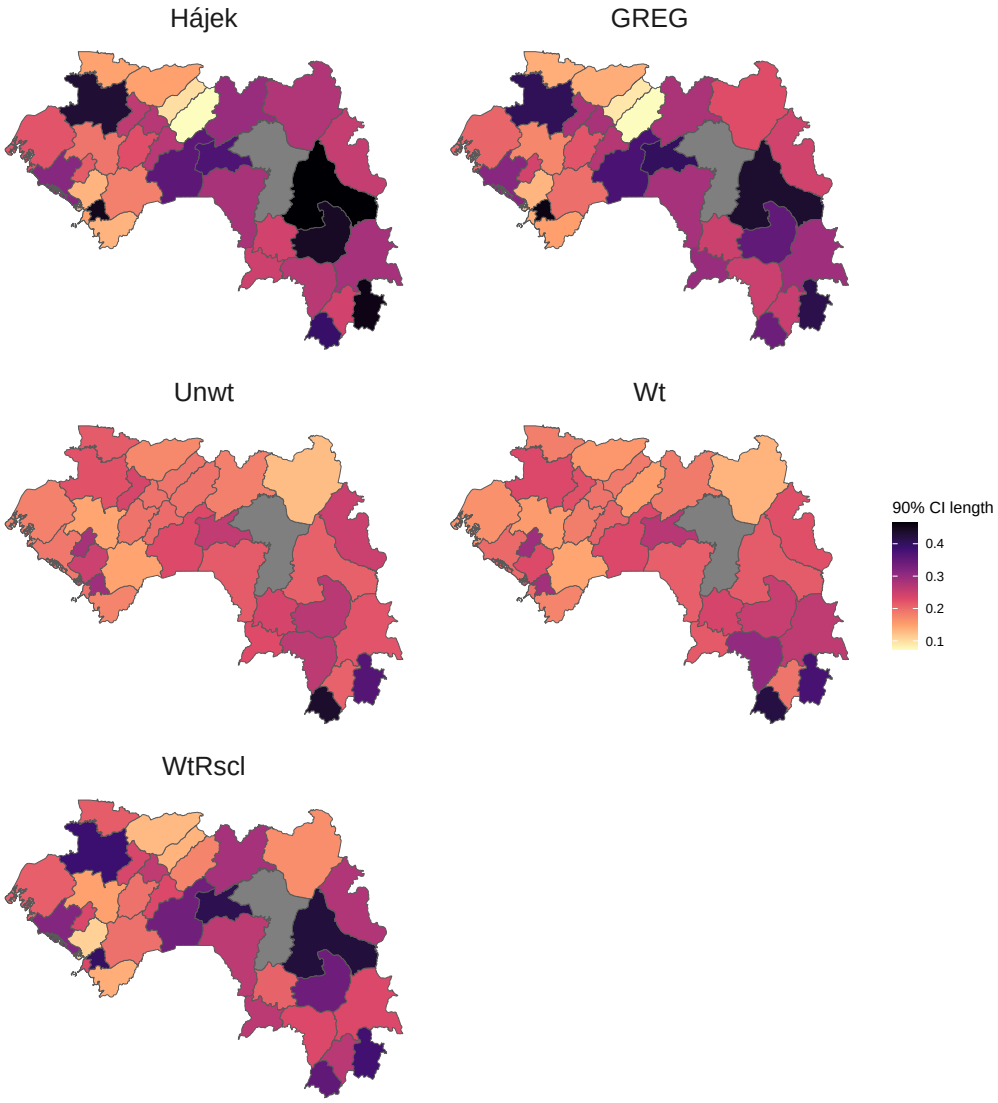

Figure 5.3: Prediction interval lengths for estimated measles vaccination rates for Admin-2 areas in Guinea in 2018.

intervals. The weighted and rescaled method generally produces larger estimates of uncertainty, generating intervals whose lengths more closely resemble those of the design-based approaches. Note that the direct Hájek estimate of the vaccination rate for the Admin-2 prefecture of Kouroussa is zero and thus a direct estimator of the associated variance is unavailable. As a result, we omit estimates for Kouroussa, which is depicted in gray in Figures 5.2 and 5.3. The unweighted and weighted (without rescaling) methods can provide estimates of the rates for Kourossa, but may yield unreliable estiamtes of uncertainty.

5.7 Discussion

Pseudo-Bayesian approaches inference enables analysts to leverage available sampling weights to adjust for features of the survey design that cannot be incorporated into a model. However, since credible sets based on a naive pseudo-posterior exhibit poor empirical coverage rates (León-Novelo and Savitsky 2019), the pseudo-posterior must be rescaled to produce credible sets that quantify uncertainty meaningfully. Although pseudo-Bayesian methods have been previously used for SAE (Parker et al. 2022), this miscalibration has not been explicitly addressed. Using a rescaling post-processing adjustment proposed by Williams and Savitsky (2021), we show that pseudo-Bayesian approaches can be used to generate improved point and interval estimators for small area means of continuous and binary outcomes.

Previous applications of pseudo-Bayesian approaches rely on scaling the sampling weights to sum to the sample size as an ad hoc solution for scaling the pseudo-posterior. The approach proposed by Williams and Savitsky first scales the sampling weights but then also estimates a multivariate design effect for the parameters of interest using the available data, which is subsequently used to rescale the pseudo-posterior. Both the initial scaling of sampling weights and the rescaling of parameter samples are potentially valuable. The initial scaling of sampling weights controls the degree of shrinkage induced

by the Gaussian prior on the random effects. If the unscaled weights are used, then the degree of shrinkage may be too low because inference on the random effects proceeds as though a population of size N is observed. The subsequent rescaling of the samples from the pseudo-posterior is aimed at improving the coverage of credible sets for parameters of interest. Similarly to Williams and Savitsky, Han and Wellner (2021) propose an rescaling approach motivated by explicit assumptions about asymptotic properties of the proposed sequence of sampling designs. Their approach, however, does not explicitly encourage rescaling the sampling weights when defining the pseudo-posterior.

The rescaling adjustment produces Bayesian credible sets that converge asymptotically on valid frequentist confidence sets based on a likelihood analysis that treats the small area means μ_i as fixed parameters. From this perspective, the use of random effects modeling is primarily to facilitate shrinkage of the resulting estimators and prevent overfitting. For small sample sizes, we find that the pseudo-Bayesian approach with corrected credible sets improves upon a naive Bayesian approach that does not incorporate survey weights as well as a pseudo-Bayesian approach with uncorrected interval estimates.

A key limitation of the approach presented here is that we focus on estimation targets that can be expressed in terms of fixed parameters for which we have asymptotically increasing amounts of data. Our approach relies upon having sufficient data to rescale model-based estimates of uncertainty, so as observed in Section 5.6, when we have severely limited or no data in an area, we may be unable to construct valid prediction intervals. When using unit level models for SAE, it has become increasingly common to model outcomes of interest as continuous spatial processes and generate area level estimates by aggregating predictions made on a high-resolution spatial grid. Under such an approach, predictions may be required for each individual cluster or location. However, our approach does not account for the case in which the parameters of interest correspond to intercepts for areas that are themselves observed at random. For example, we do not seek to estimate individual cluster level effects.

The models explored here are simple and do not incorporate spatial modeling or cluster level random effects. Moreover, dependence between units induced by clustering or unobserved covariates is not directly modeled. One direction for future work is to extend the pseudo-Bayesian approach to spatial models and to investigate the benefits of incorporating sampling weights into spatial modeling of survey data.

Chapter 6

CONCLUSION

In the past decade, advances in machine learning and spatial modeling approaches have been leveraged to carry out large-scale global health mapping projects in a variety of contexts, including for mapping excess death associated with COVID-19, under-five mortality rates, and vaccination rates. The demand for subnational estimates of key health outcomes and demographic indicators continues to increase, especially at levels of resolution at which traditional survey statistics estimators may be unreliable. New statistical and machine learning approaches that are considerably more complicated than the standard model-based SAE strategies are becoming more common, but many of these geospatial and machine learning approaches may neglect the possible impact of survey design. As such, it is critical to relate these new approaches to existing methods in survey statistics and SAE. and to offer simple, interpretable alternatives.

The methods proposed in this dissertation aim to bridge the gap between fully design-based weighted estimators such as the Horvitz-Thompson estimator and model-based estimation approaches. Traditional design-based estimators from the survey statistics literature have are typically design consistent and asymptotically design unbiased, but have historically been used in contexts in which high quality census or auxiliary data are available. From a pragmatic perspective, given sufficient data, such estimators may be preferred since they generally rely on fewer assumptions than model-assisted or model-based approaches. On the other hand, when data are particularly limited, or estimates are desired for areas in which no data are available, the only way to produce estimates may be via model-based synthetic estimators. Geostatistical approaches treating quantities of

interest as continuous spatial processes allow prediction at all locations across a spatial domain, even those where data are not available. Such approaches are also not reliant on assumptions on neighborhood structure of discrete domains, though they do depend on the choice of correlation function. However, these approaches are necessarily more model-dependent and do not always acknowledge the design of the surveys used to collect data. A flawed estimate may not necessarily be better than no estimate, of course, so care is needed when determining if such model-based estimates are sufficiently reliable for publication.

More work is needed to improve subnational estimation of critical indicators, especially in low- and middle- income countries. Three directions for future work are developing diagnostics and metrics for evaluating small area estimators, improving software and sharing educational materials for SAE, and further exploring theory for hierarchical and spatial models under complex survey sampling. In most cases there are no guarantees on finite sample performance for small area estimators and evaluating them in low data settings is not well understood.

Evaluation of SAE methods is difficult as typically all available data is incorporated into the estimation process and there are no natural validation datasets. Some standard approaches for generating data for validation include simulation of a population based on a data generating model (as done throughout this dissertation) or simulations based on resampling an artificial population from available sample data. In rare circumstances, estimates based on survey data can be compared with "ground truth" values from a census covering the same areas in a similar or identical timeframe. Standard cross-validation methods based on leaving out part of the sample and computing estimators using the remainder are difficult to apply for SAE problems. If we believe that there may be spatial dependence in the responses, then cross-validation methods that rely on partitioning the dataset into independent components may fail. As such, developing new metrics and diagnostics for model-based small area estimators may be useful in helping decision

makers understand results and uncertainty related to modeling decisions and sampling variability.

As discussed above, existing geostatistical approaches to SAE on a global scale can be computationally expensive and difficult to replicate. Developing open source software and simple, interpretable estimators to help researchers produce and adapt their own maps of health and demographic indicators should thus be a priority. Having the ability to consult multiple competing estimators derived from varying approaches, instead of relying on one set of estimates to make decisions, will help provide insight as to the sensitivity of estimates to modeling choices.

Finally, in terms of potential future research, more work is needed to understand the use of generalized linear mixed models for SAE, especially with regards to models with multiple layers of random effects and estimation of parameters under complex survey sampling. For example, pairwise likelihood methods, which have previously been used for fitting mixed models under informative sampling, may be adapted for prediction of random effects in a SAE context. Similarly, using multilevel models for prediction with pseudo-likelihood or pseudo-Bayesian approaches under complex sampling presents difficulties related to scaling the sampling weights in order to balance the bias and variance of the resulting estimators. Adapting such methods specifically for prediction of small area quantities, as opposed to for estimating model parameters could potentially lead to small area estimates with improved precision.

Despite these opportunities for methodological improvement, the fundamental challenges of mapping subnational health and demographic indicators using limited survey data will not be solved by improved statistical models. In truth, the largest improvements in the coverage and precision of estimates may be found by supporting those engaged in the crucial work of survey design and data collection and building active collaborations with researchers and data analysts situated in the countries studied.

Appendix A

A VARIANCE SMOOTHING AREAL MODEL FOR ESTIMATING PROPORTIONS

A.1 Parameter estimation

Below we provide further details on the estimation process and priors used for each method described in the simulations and applications.

A.1.1 Direct estimation

We use the R package `survey` for computing direct weighted Hájek estimators (and corresponding variance estimates) for all areas.

A.1.2 Mean-smoothing model-based estimation

We adopt a fully Bayesian approach to estimating the mean-smoothing unmatched model described above, assuming priors for model parameters and then using MCMC as implemented in the R package `STAN` to sample from the posterior distributions of the area level proportions p_i for all $i = 1, \ldots, m$. We place a $N(0, 1000)$ prior on the area level model intercept and fixed effects. We use penalized complexity priors for the variance parameter σ_u, as described by Simpson et al. (2017). We specify these priors such that $P(\sigma_u > 1) = 0.01$. For the spatial models, we place a $\mathrm{Beta}(1/2, 1/2)$ prior on the spatial correlation prior ϕ.

A.1.3 Joint-smoothing model-based estimation

We keep the same priors as for the mean-smoothing model on area level model intercept and fixed effects as well as the variance parameters σ_u and for the spatial models, ϕ. We then place a penalized complexity prior for the variance parameter σ_τ such that $P(\sigma_\tau > 1) = 0.01$. We place a $N(0, 1)$ prior on γ_0, a $N(1, .5)$ prior on γ_1, and a $N(-1, .5)$ prior on γ_2 to shrink the resulting variances estimates towards that of a binomial random variable.

A.2 Additional results

A.2.1 Covariate maps

Figure A.1 provides maps of the simulated cluster locations and covariate values used in the simulations described in Chapter 4.

A.2.2 Large sample simulations

Table A.1 provides results for an additional set of simulations that were identical to the simulations with $\mu = 0.5$ described in Section 3.4, except with a larger sample size, with twenty-five clusters sampled per stratum rather than the eight used in the main text. The results illustrate that for large sample sizes, the joint smoothing and mean smoothing model-based estimators perform similarly to the direct weighted estimators with 90% prediction interval coverage rates that are close to nominal.

A.2.3 Applications

Tables A.2 and A.3 provide point estimates and corresponding 90% prediction intervals for model hyperparameters.

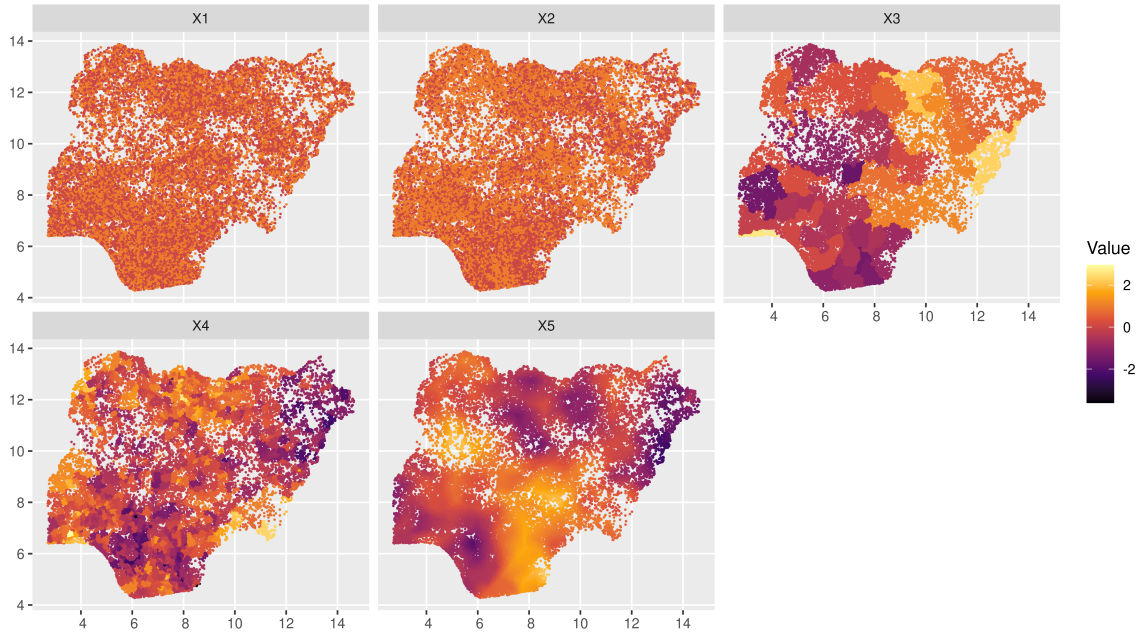

Figure A.1: Simulated cluster locations and covariate values used to generate population data.

Method	RMSE	MAE	90% Cov.	MIL
Direct (Hájek)	2.40	1.91	90	7.98
Unmatched MS	2.35	1.87	90	7.87
Spatial Unmatched MS	2.33	1.86	90	7.80
Unmatched JS	2.37	1.89	91	7.99
Spatial Unmatched JS	2.35	1.87	91	7.92

Table A.1: RMSE ($\times 100$), MAE ($\times 100$), coverage rates, and mean interval length ($\times 100$) of estimators of area level means across 1,000 simulated populations with spatially correlated binary responses based on large sample (25 clusters) obtained via informative sampling. The reduced model omits one of the spatial covariates in the full model.

Parameter	*Spatial Unmatched MS*	*Spatial Unmatched JS*
$\text{logit}(\mu)$	0.36 (0.22, 0.50)	0.36 (0.23, 0.49)
σ_u	0.71 (0.54, 0.96)	0.71 (0.54, 0.94)
ϕ	0.84 (0.38, 1.00)	0.86 (0.39, 1.00)
γ_0		-0.73 (-2.22, 0.74)
γ_1		0.47 (-0.14, 1.01)
γ_2		-0.90 (-1.18, -0.63)
σ_τ		0.42 (0.29, 0.59)

Table A.2: Point estimates and 90% interval estimates for model parameters for Nigeria measles vaccination example.

Tables A.4 and A.5 provide full estimates with prediction intervals for all areas of interest for the methods described in the manuscript applied to both the measles vaccination rate and HIV prevalence rate applications.

Parameter	*Spatial Unmatched MS*		*Spatial Unmatched JS*	
$\text{logit}(\mu)$	-2.03	(-2.14, -1.93)	-2.02	(-2.14, -1.92)
σ_u	0.41	(0.28, 0.62)	0.43	(0.30, 0.64)
ϕ	0.89	(0.36, 1.00)	0.88	(0.36, 1.00)
γ_0			-0.22	(-1.90, 1.50)
γ_1			0.91	(0.51, 1.27)
γ_2			-0.96	(-1.29, -0.64)
σ_τ			0.19	(0.07, 0.37)

Table A.3: Point estimates and 90% interval estimates for model parameters for Malawi HIV prevalence example.

State	*Hájek*		*Spatial Unmatched MS*		*Spatial Unmatched JS*	
Lagos	0.89	(0.84, 0.94)	0.88	(0.83, 0.93)	0.89	(0.83, 0.93)
Ekiti	0.87	(0.79, 0.94)	0.83	(0.76, 0.89)	0.83	(0.75, 0.89)
Anambra	0.80	(0.73, 0.88)	0.79	(0.73, 0.85)	0.79	(0.72, 0.85)
Enugu	0.80	(0.71, 0.88)	0.77	(0.7, 0.85)	0.77	(0.69, 0.85)
Edo	0.79	(0.7, 0.88)	0.78	(0.7, 0.85)	0.78	(0.69, 0.85)
Osun	0.77	(0.69, 0.85)	0.76	(0.69, 0.83)	0.76	(0.68, 0.83)
Abia	0.75	(0.69, 0.82)	0.75	(0.69, 0.81)	0.75	(0.68, 0.82)
Delta	0.75	(0.69, 0.8)	0.75	(0.7, 0.8)	0.76	(0.69, 0.82)
Abuja	0.73	(0.68, 0.79)	0.72	(0.67, 0.77)	0.72	(0.65, 0.78)
Imo	0.73	(0.63, 0.84)	0.74	(0.65, 0.83)	0.73	(0.64, 0.82)
Bayelsa	0.73	(0.65, 0.8)	0.73	(0.66, 0.81)	0.73	(0.65, 0.81)
Ondo	0.69	(0.58, 0.8)	0.70	(0.6, 0.79)	0.69	(0.6, 0.78)
Rivers	0.68	(0.59, 0.77)	0.69	(0.61, 0.77)	0.69	(0.61, 0.77)
Cross River	0.65	(0.52, 0.78)	0.66	(0.55, 0.77)	0.66	(0.55, 0.76)
Adamawa	0.65	(0.57, 0.73)	0.63	(0.56, 0.71)	0.63	(0.55, 0.71)
Nassarawa	0.65	(0.54, 0.75)	0.63	(0.54, 0.72)	0.63	(0.54, 0.72)
Ebonyi	0.63	(0.57, 0.7)	0.64	(0.58, 0.7)	0.64	(0.58, 0.71)
Akwa Ibom	0.63	(0.55, 0.71)	0.64	(0.56, 0.72)	0.64	(0.56, 0.73)
Benue	0.63	(0.54, 0.71)	0.63	(0.55, 0.71)	0.63	(0.55, 0.7)
Oyo	0.60	(0.51, 0.7)	0.61	(0.52, 0.7)	0.61	(0.52, 0.7)
Plateau	0.59	(0.52, 0.65)	0.58	(0.52, 0.65)	0.58	(0.51, 0.65)
Kano	0.56	(0.5, 0.62)	0.56	(0.5, 0.62)	0.56	(0.5, 0.62)
Jigawa	0.54	(0.48, 0.6)	0.53	(0.48, 0.59)	0.53	(0.47, 0.59)
Kwara	0.51	(0.35, 0.67)	0.55	(0.42, 0.68)	0.54	(0.43, 0.66)
Ogun	0.51	(0.4, 0.62)	0.55	(0.45, 0.65)	0.55	(0.45, 0.66)
Borno	0.49	(0.42, 0.57)	0.49	(0.42, 0.56)	0.49	(0.41, 0.57)
Yobe	0.45	(0.4, 0.5)	0.45	(0.4, 0.5)	0.45	(0.39, 0.51)
Kaduna	0.43	(0.35, 0.5)	0.43	(0.36, 0.5)	0.43	(0.36, 0.5)
Kogi	0.42	(0.32, 0.53)	0.49	(0.39, 0.59)	0.50	(0.39, 0.6)
Taraba	0.42	(0.36, 0.48)	0.43	(0.37, 0.49)	0.43	(0.36, 0.5)
Niger	0.39	(0.27, 0.51)	0.40	(0.3, 0.51)	0.41	(0.32, 0.5)
Bauchi	0.36	(0.3, 0.43)	0.37	(0.31, 0.43)	0.37	(0.31, 0.44)
Katsina	0.34	(0.26, 0.41)	0.34	(0.27, 0.41)	0.34	(0.27, 0.41)
Kebbi	0.31	(0.25, 0.38)	0.31	(0.24, 0.37)	0.31	(0.24, 0.37)
Gombe	0.28	(0.2, 0.36)	0.31	(0.23, 0.38)	0.31	(0.24, 0.39)
Sokoto	0.18	(0.13, 0.23)	0.18	(0.13, 0.23)	0.17	(0.12, 0.23)
Zamfara	0.14	(0.07, 0.21)	0.17	(0.12, 0.23)	0.18	(0.12, 0.24)

Table A.4: Point estimates of measles vaccination rates and 90% interval estimates for Admin-1 areas among children aged 12–23 months in Nigeria in 2018.

State	*Hájek*		*Spatial Unmatched MS*		*Spatial Unmatched JS*	
Mulanje	0.27	(0.23, 0.32)	0.25	(0.21, 0.29)	0.25	(0.21, 0.29)
Blantyre	0.24	(0.2, 0.28)	0.21	(0.17, 0.26)	0.22	(0.18, 0.26)
Phalombe	0.24	(0.19, 0.29)	0.23	(0.19, 0.27)	0.23	(0.19, 0.28)
Zomba	0.20	(0.16, 0.24)	0.19	(0.16, 0.22)	0.19	(0.16, 0.22)
Ntcheu	0.16	(0.12, 0.21)	0.15	(0.11, 0.19)	0.16	(0.13, 0.2)
Mangochi	0.16	(0.12, 0.19)	0.15	(0.12, 0.18)	0.15	(0.12, 0.18)
Nsanje	0.16	(0.11, 0.2)	0.15	(0.11, 0.19)	0.16	(0.12, 0.2)
Thyolo	0.15	(0.11, 0.2)	0.16	(0.13, 0.19)	0.16	(0.13, 0.2)
Balaka	0.15	(0.11, 0.19)	0.15	(0.12, 0.18)	0.15	(0.12, 0.19)
Chiradzulu	0.15	(0.1, 0.19)	0.16	(0.12, 0.19)	0.16	(0.12, 0.2)
Neno	0.14	(0.1, 0.18)	0.15	(0.11, 0.18)	0.15	(0.12, 0.18)
Mwanza	0.13	(0.09, 0.17)	0.14	(0.1, 0.17)	0.14	(0.1, 0.17)
Karonga	0.12	(0.09, 0.15)	0.11	(0.08, 0.13)	0.10	(0.07, 0.13)
Chikwawa	0.11	(0.07, 0.15)	0.13	(0.1, 0.16)	0.13	(0.1, 0.16)
Nkhata Bay	0.10	(0.07, 0.14)	0.09	(0.06, 0.12)	0.09	(0.07, 0.12)
Nkhotakota	0.10	(0.07, 0.13)	0.09	(0.07, 0.11)	0.09	(0.07, 0.11)
Machinga	0.09	(0.07, 0.12)	0.11	(0.08, 0.13)	0.10	(0.08, 0.13)
Kasungu	0.09	(0.06, 0.12)	0.08	(0.06, 0.11)	0.09	(0.07, 0.11)
Rumphi	0.09	(0.05, 0.12)	0.08	(0.06, 0.11)	0.09	(0.06, 0.11)
Lilongwe	0.09	(0.06, 0.11)	0.08	(0.06, 0.11)	0.08	(0.06, 0.11)
Dedza	0.08	(0.04, 0.12)	0.09	(0.07, 0.12)	0.10	(0.08, 0.13)
Mchinji	0.08	(0.06, 0.1)	0.08	(0.06, 0.1)	0.07	(0.05, 0.09)
Salima	0.07	(0.05, 0.1)	0.08	(0.06, 0.1)	0.07	(0.06, 0.1)
Ntchisi	0.07	(0.05, 0.1)	0.07	(0.06, 0.09)	0.07	(0.05, 0.09)
Dowa	0.07	(0.04, 0.1)	0.07	(0.05, 0.09)	0.07	(0.05, 0.09)
Chitipa	0.06	(0.03, 0.09)	0.07	(0.05, 0.09)	0.06	(0.04, 0.09)
Mzimba	0.06	(0.04, 0.08)	0.06	(0.05, 0.08)	0.06	(0.05, 0.08)

Table A.5: Point estimates of HIV prevalence and 90% interval estimates for Admin-1 areas among women aged 15–49 in Nigeria in 2018.

Appendix B

COMBINING AREA LEVEL AND UNIT LEVEL MODELING FOR SMALL AREA ESTIMATION FOR PROPORTIONS

B.1 Design consistency of survey regression LGREG estimator

We now consider the design consistency of the model-assisted estimator specified by Equation (4.5) and discuss the relevant regularity assumptions. Our proof adapts the one presented by Kennel and Valliant (2020) for a multivariate logistic model-assisted estimator for clustered samples. Rather than showing design consistency for $\widehat{p}_i^{MA}$, we instead consider the area-specific total estimator $\widehat{t}_i^{MA}$:

$$\widehat{t}_i^{MA} = \sum_{j \in U(i)} \widehat{y}_{ij} + \sum_{j \in S(i)} w_{ij}(y_{ij} - \widehat{y}_{ij}) \tag{B.1}$$

Let $\widehat{y}_{ij}$ denote predictions from our working logistic regression model:

$$P(y_{ij} = 1 \mid \mathbf{x}_{ij}, \boldsymbol{\beta}) = q_{ij} \tag{B.2}$$

$$\text{logit}(q_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} \tag{B.3}$$

If we had full population data, we could estimate $\boldsymbol{\beta}$ by maximizing the population log-likelihood to obtain finite population parameters $\mathbf{B}$:

$$\mathbf{B} = \arg\max_{\boldsymbol{\beta}} \sum_i \sum_{j \in U(i)} \ell(y_{ij}\boldsymbol{\beta}) \tag{B.4}$$

Since we only have data for sampled units, in practice, we maximize the survey-weighted log-likelihood to obtain $\widehat{\mathbf{B}}$, an estimator of $\mathbf{B}$:

$$\widehat{\mathbf{B}} = \arg\max_{\boldsymbol{\beta}} \sum_i \sum_{j \in S(i)} \frac{1}{\pi_{ij}} \ell(y_{ij}; \boldsymbol{\beta}) \tag{B.5}$$

To reflect the dependence of our predictions on the estimated regression parameters we introduce the following notation, letting $\widetilde{y}$ denote predictions if we observed the finite population parameters $\mathbf{B}$:

$$\widehat{y}_{ij} = \mu(\mathbf{x}_{ij}, \widehat{\mathbf{B}}) \tag{B.6}$$

$$\widetilde{y}_{ij} = \mu(\mathbf{x}_{ij}, \mathbf{B}) \tag{B.7}$$

We assume an asymptotic regime with a fixed number of m areas, where area i has sample size $n(i)$ and population size $N(i)$. We let N denote the overall population size and n denote the overall sample size. We assume a sequence of designs and populations such that $N, N(i) \to \infty$ and assume the following conditions:

1. The regression parameter estimates satisfy $\widehat{\mathbf{B}} = \mathbf{B} + O(n^{-1/2})$. Moreover, $\mathbf{B} \to \boldsymbol{\beta}$ as $N \to \infty$.

2. For each area i, for each j, $|\frac{\partial \mu}{\partial \mathbf{t}}| \leq h(\mathbf{x}_{ij}, \boldsymbol{\beta})$ for all $\mathbf{t}$ in a neighborhood centered on $\boldsymbol{\beta}$ such that $\frac{1}{N(i)} \sum_{j \in U(i)} h(\mathbf{z}_{ij}, \boldsymbol{\beta}) = O(1)$.

3. For each area i, $\sum_{j \in S(i)} w_{ij} \widetilde{y}_{ij}$ is design-consistent for $\sum_{j \in U(i)} \widetilde{y}_{ij}$ and $\sum_{j \in S(i)} w_{ij} y_{ij}$ is design-consistent for $\sum_{j \in U(i)} y_{ij}$.

Note that Assumption 1 requires that the working model parameter estimator converges to some limit. Since we may sample from many different areas, we assume the same working model is used for all areas or alternatively, that the survey design calls for proportional sampling of all areas a. Assumption 2 requires that the derivative term $|\frac{\partial \mu}{\partial \mathbf{t}}| \leq h(\mathbf{x}_{ij}, \boldsymbol{\beta})$ is bounded in each small area. Assumption 3 requires that Horvitz-Thompson type estimators are design-consistent under the sequence of designs specified.

By Taylor's theorem, for all $j \in U(i)$, there is some vector $\mathbf{B}^*_{ij}$ such that

$$\widehat{y}_{ij} = \widetilde{y}_{ij} + \left[\left. \frac{\partial \mu}{\partial \mathbf{t}} \right|_{\mathbf{t} = \mathbf{B}^*_{ij}} \right]^T \text{vec}(\widehat{\mathbf{B}} - \mathbf{B}) \tag{B.8}$$

Here, $\frac{\partial \mu}{\partial \mathbf{t}}$ is a $(p+1)\times 1$ vector of the partial derivatives of μ with respect to the components of $\mathbf{t}$. By summing over all units $j \in U(i)$ and dividing by the population size $N(i)$, we obtain the following:

$$\frac{1}{N(i)}\sum_{j\in U(i)}\widehat{y}_{ij} = \frac{1}{N(i)}\sum_{j\in U(i)}\widetilde{y}_{ij} + \frac{1}{N(i)}\sum_{j\in U(i)}\left[\frac{\partial \mu}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\mathbf{B}^*_{ij}}\right]^T \text{vec}(\widehat{\mathbf{B}}-\mathbf{B}) \tag{B.9}$$

Under Assumptions 1 and 2, we have that

$$\frac{1}{N(i)}\sum_{j\in U(i)}\widehat{y}_{ij} - \frac{1}{N(i)}\sum_{j\in U(i)}\widetilde{y}_{ij} = O_{\mathbb{P}_d}(n^{-1/2}) \tag{B.10}$$

and

$$\frac{1}{N(i)}\sum_{j\in S(i)} w_{ij}\widehat{y}_{ij} - \frac{1}{N(i)}\sum_{j\in S(i)} w_{ij}\widetilde{y}_{ij} = O_{\mathbb{P}_d}(n^{-1/2}) \tag{B.11}$$

implying that

$$\frac{1}{N(i)}\left[\sum_{j\in U(i)}\widehat{y}_{ij} - \sum_{j\in S(i)} w_{ij}\widehat{y}_{ij}\right] = \frac{1}{N(i)}\left[\sum_{j\in U(i)}\widetilde{y}_{ij} - \sum_{j\in S(i)} w_{ij}\widetilde{y}_{ij}\right] + O_{\mathbb{P}_d}(n^{-1/2}) \tag{B.12}$$

We can thus rewrite $\widehat{t}_i^{MA}$ as follows:

$$\frac{1}{N(i)}\widehat{t}_i^{MA} = \frac{1}{N(i)}\left[\sum_{j\in U(i)}\widehat{y}_{ij} + \sum_{j\in S(i)} w_{ij}(y_{ij}-\widehat{y}_{ij})\right] \tag{B.13}$$

$$= \frac{1}{N(i)}\left[\sum_{j\in S(i)} w_{ij}y_{ij} + \sum_{j\in U(i)}\widehat{y}_{ij} - \sum_{j\in S(i)} w_{ij}\widehat{y}_{ij}\right] \tag{B.14}$$

$$= \frac{1}{N(i)}\left[\sum_{j\in S(i)} w_{ij}y_{ij} + \sum_{j\in U(i)}\widetilde{y}_{ij} - \sum_{j\in S(i)} w_{ij}\widetilde{y}_{ij}\right] + O_{\mathbb{P}_d}(n^{-1/2}) \tag{B.15}$$

Therefore, as long as Assumption 3 holds, $\widehat{t}_i^{MA}$ will converge to the desired population total $\sum_{j\in U(i)} y_{ij}$.

B.2 Parameter estimation

The analyses were carried out using the R programming language (R Core Team 2023). The R `survey` package provides tools for analyzing survey data and calculating com-

monly used small area estimators (Lumley 2004). We also use the R package `INLA` to conduct approximate Bayesian inference (Rue et al. 2017). The `tidyverse` (Wickham 2014), `sf` (Pebesma 2018), and `raster` (2021) packages were used to process data. The `R` package `SUMMER` (Li et al. 2022) can be used to fit similar models and functions for smoothed model-assisted estimation are currently in development.

We compute the Hájek estimators for all areas using the R package `survey`, which also provides associated variance estimates. For the simulations and application, the working logistic regression models are fit via survey-weighted maximum likelihood using the R package `survey`. Based on the working model predictions, model-assisted estimators are computed for each area and associated variance estimates are calculated using Kennel and Valliant's (2010) with-replacement cluster sampling variance estimator.

The area level models described in the main text take as input a set of direct or model-assisted estimates for all areas with associated variance estimates. We adopt a fully Bayesian approach to estimation by assuming priors on model parameters and using `INLA` to approximate the posterior distributions for area level proportions p_i for all $i = 1, \ldots, m$. We generate predictions by repeatedly sampling from these posterior distributions, enabling us to produce point estimates (from the posterior medians) and interval estimates (by taking relevant quantiles of the posteriors). The uncertainty of the resulting estimates may be quantified either using posterior variance or by taking the length of interval estimates.

For all area level models, we place a flat prior on the area level model intercept and fixed effects, so $\pi(\boldsymbol{\beta}) \propto 1$. As described above, we use penalized complexity priors for the variance parameters, as implemented in `INLA`. For the non-spatial area-level models, we specify the prior for the area effect variance σ_u^2 such that $P(\sigma_u > 5) = 0.01$. For the spatial area-level models, we specify the prior for the area effect variance σ_u^2 such that $P(\sigma_u > 5) = 0.01$ and for the spatial correlation parameter ϕ such that $P(\phi > .5) = 2/3$. Here, we select these priors to be relatively flat.

The unit level models described in the main text take as input survey microdata with covariate information for each sampled individual. As with the area level models, we use a fully Bayesian approach implemented using `INLA`. In order to generate predictions, we require covariate information for all sampled and non-sampled individuals to enable us to generate predictions for all individuals in our population of interest.

For all unit level models, we place a flat prior on the intercept and fixed effects, so $\pi(\boldsymbol{\beta}) \propto 1$. As described above, we use penalized complexity priors for the variance parameters. For the non-spatial area-level models, we specify the prior for the area effect variance σ_u^2 such that $P(\sigma_u > 5) = 0.01$. For the spatial area-level models, we specify the prior for the area effect variance σ_u^2 such that $P(\sigma_u > 5) = 0.01$ and for the spatial correlation parameter ϕ such that $P(\phi > .5) = 2/3$.

B.3 Additional results

State	Hájek	Sp. SH	MA	Spatial SMA	Sp. Betabinomial
Lagos	0.89 (0.84, 0.94)	0.83 (0.78, 0.88)	0.82 (0.77, 0.86)	0.87 (0.81, 0.92)	0.86 (0.81, 0.9)
Ekiti	0.87 (0.79, 0.94)	0.86 (0.77, 0.94)	0.79 (0.68, 0.87)	0.80 (0.7, 0.88)	0.78 (0.71, 0.85)
Anambra	0.80 (0.73, 0.88)	0.79 (0.72, 0.86)	0.77 (0.69, 0.83)	0.78 (0.71, 0.84)	0.78 (0.72, 0.83)
Enugu	0.80 (0.71, 0.88)	0.77 (0.69, 0.85)	0.75 (0.67, 0.81)	0.76 (0.67, 0.84)	0.75 (0.67, 0.81)
Edo	0.79 (0.7, 0.88)	0.79 (0.71, 0.87)	0.77 (0.69, 0.84)	0.76 (0.67, 0.84)	0.77 (0.7, 0.84)
Osun	0.77 (0.69, 0.85)	0.73 (0.65, 0.81)	0.71 (0.63, 0.78)	0.75 (0.68, 0.81)	0.73 (0.66, 0.8)
Abia	0.75 (0.69, 0.82)	0.74 (0.68, 0.81)	0.73 (0.67, 0.79)	0.74 (0.68, 0.8)	0.78 (0.72, 0.84)
Delta	0.75 (0.69, 0.8)	0.74 (0.67, 0.8)	0.73 (0.67, 0.79)	0.74 (0.69, 0.79)	0.72 (0.65, 0.78)
Abuja	0.73 (0.68, 0.79)	0.74 (0.67, 0.8)	0.71 (0.65, 0.78)	0.72 (0.66, 0.77)	0.73 (0.67, 0.79)
Imo	0.73 (0.63, 0.84)	0.70 (0.61, 0.79)	0.70 (0.61, 0.78)	0.73 (0.64, 0.81)	0.65 (0.58, 0.72)
Bayelsa	0.73 (0.65, 0.8)	0.70 (0.62, 0.78)	0.70 (0.62, 0.77)	0.73 (0.65, 0.8)	0.71 (0.64, 0.77)
Ondo	0.69 (0.58, 0.8)	0.67 (0.56, 0.78)	0.67 (0.57, 0.76)	0.69 (0.58, 0.78)	0.68 (0.6, 0.75)
Rivers	0.68 (0.59, 0.77)	0.66 (0.59, 0.72)	0.66 (0.6, 0.72)	0.69 (0.61, 0.76)	0.69 (0.63, 0.75)
Cross River	0.65 (0.52, 0.78)	0.65 (0.51, 0.79)	0.65 (0.53, 0.76)	0.65 (0.54, 0.76)	0.65 (0.56, 0.73)
Adamawa	0.65 (0.57, 0.73)	0.66 (0.58, 0.74)	0.63 (0.55, 0.71)	0.62 (0.54, 0.7)	0.65 (0.58, 0.71)
Nassarawa	0.65 (0.54, 0.75)	0.62 (0.5, 0.73)	0.60 (0.5, 0.7)	0.63 (0.53, 0.72)	0.67 (0.6, 0.73)
Ebonyi	0.63 (0.57, 0.7)	0.63 (0.58, 0.69)	0.63 (0.58, 0.69)	0.64 (0.58, 0.7)	0.61 (0.55, 0.68)
Akwa Ibom	0.63 (0.55, 0.71)	0.64 (0.56, 0.72)	0.64 (0.56, 0.72)	0.64 (0.56, 0.71)	0.65 (0.58, 0.72)
Benue	0.63 (0.54, 0.71)	0.62 (0.54, 0.7)	0.62 (0.54, 0.7)	0.63 (0.55, 0.7)	0.68 (0.62, 0.74)
Oyo	0.60 (0.51, 0.7)	0.57 (0.46, 0.67)	0.57 (0.47, 0.67)	0.61 (0.52, 0.69)	0.54 (0.47, 0.62)
Plateau	0.59 (0.52, 0.65)	0.60 (0.53, 0.66)	0.59 (0.52, 0.65)	0.58 (0.52, 0.64)	0.61 (0.54, 0.67)
Kano	0.56 (0.5, 0.62)	0.58 (0.52, 0.63)	0.57 (0.52, 0.62)	0.56 (0.5, 0.61)	0.58 (0.54, 0.63)
Jigawa	0.54 (0.48, 0.6)	0.54 (0.49, 0.6)	0.54 (0.48, 0.6)	0.53 (0.48, 0.59)	0.60 (0.55, 0.66)
Kwara	0.51 (0.35, 0.67)	0.48 (0.34, 0.62)	0.52 (0.41, 0.63)	0.55 (0.43, 0.67)	0.52 (0.45, 0.58)
Ogun	0.51 (0.4, 0.62)	0.50 (0.4, 0.6)	0.53 (0.44, 0.62)	0.55 (0.45, 0.65)	0.53 (0.46, 0.6)
Borno	0.49 (0.42, 0.57)	0.43 (0.35, 0.51)	0.43 (0.36, 0.51)	0.49 (0.42, 0.56)	0.45 (0.39, 0.52)
Yobe	0.45 (0.4, 0.5)	0.47 (0.42, 0.52)	0.47 (0.42, 0.52)	0.45 (0.4, 0.5)	0.47 (0.41, 0.53)
Kaduna	0.43 (0.35, 0.5)	0.45 (0.38, 0.52)	0.45 (0.39, 0.52)	0.43 (0.36, 0.5)	0.48 (0.43, 0.53)
Kogi	0.42 (0.32, 0.53)	0.41 (0.31, 0.5)	0.46 (0.37, 0.56)	0.49 (0.39, 0.59)	0.46 (0.38, 0.54)
Taraba	0.42 (0.36, 0.48)	0.42 (0.36, 0.48)	0.43 (0.36, 0.49)	0.43 (0.37, 0.49)	0.45 (0.39, 0.51)
Niger	0.39 (0.27, 0.51)	0.38 (0.27, 0.5)	0.40 (0.3, 0.51)	0.41 (0.31, 0.51)	0.44 (0.38, 0.49)
Bauchi	0.36 (0.3, 0.43)	0.36 (0.3, 0.42)	0.37 (0.31, 0.44)	0.37 (0.32, 0.43)	0.43 (0.38, 0.49)
Katsina	0.34 (0.26, 0.41)	0.32 (0.25, 0.4)	0.34 (0.26, 0.41)	0.35 (0.28, 0.43)	0.37 (0.32, 0.43)
Kebbi	0.31 (0.25, 0.38)	0.33 (0.26, 0.39)	0.33 (0.27, 0.39)	0.31 (0.25, 0.37)	0.37 (0.31, 0.43)
Gombe	0.28 (0.2, 0.36)	0.26 (0.17, 0.36)	0.31 (0.22, 0.41)	0.31 (0.24, 0.4)	0.35 (0.29, 0.4)
Sokoto	0.18 (0.13, 0.23)	0.18 (0.13, 0.22)	0.18 (0.14, 0.23)	0.18 (0.14, 0.23)	0.19 (0.15, 0.24)
Zamfara	0.14 (0.07, 0.21)	0.12 (0.07, 0.17)	0.16 (0.11, 0.23)	0.19 (0.13, 0.28)	0.16 (0.12, 0.21)

Table B.1: Estimated measles vaccination rates (left) with 90% prediction intervals for Admin-1 areas in Nigeria in 2018.

Appendix C

UNIT LEVEL MODELING FOR SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING

C.1 Misspecification of the superpopulation model

C.1.1 Linear regression

Under the fixed intercepts linear nested error regression model, we can derive closed-form expressions for the census maximum likelihood estimators:

$$\widehat{\boldsymbol{\beta}_1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\widehat{\beta}_{0i} = \overline{Y}_i - \overline{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}_1}$$

where $\mathbf{X}$ denotes the population covariate matrix and $\mathbf{Y}$ denotes the population response vector. As a result, we can define a model-based predictor of μ_i as follows:

$$\widehat{\mu}_i = \widehat{\beta}_{0i} + \overline{\mathbf{X}}_i^T\widehat{\boldsymbol{\beta}_1} = \overline{Y}_i,$$

so the population estimate of μ_i coincides with the true finite population mean $\overline{Y}_i$. In this sense, assuming the pseudo-posterior concentrates on the census maximum likelihood estimators, even if the model is misspecified, the small area mean estimators will converge to the finite population mean because the model includes area-specific fixed intercept terms. For example, if there are missing relevant covariates, the estimates of $\boldsymbol{\beta}_1$ may be biased, but the resulting small area mean estimates can be robust to misspecification.

C.1.2 Logistic regression

Closed-form expressions may not be available for the maximum likelihood estimators for nested error regression models with non-Gaussian likelihoods. Below, we consider how misspecification of a logistic regression model affects the resulting small area estimators. Given parameter estimates for β_{0i} and $\boldsymbol{\beta}_1$, a model-based predictor of μ_i can be computed if covariate information for

$$\widehat{\mu}_i = \frac{1}{N(i)} \sum_{j \in U(i)} \text{expit}(\widehat{\beta}_{0i} + \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}_1}) = \frac{1}{N(i)} \sum_{j \in U(i)} \widehat{q}_{ij}$$

Given full population data, an estimating equations approach for estimating β_{0i} and $\boldsymbol{\beta}_1$ involves solving the following sets of equations:

$$\sum_{j \in U(i)} (y_{ij} - q_{ij}) = 0; \; i = 1, \ldots, m \tag{C.1}$$

$$\sum_i \sum_{j \in U(i)} (y_{ij} - q_{ij}) \mathbf{x}_{ij} = 0 \tag{C.2}$$

As such, an approximate solution to these equations will generally yield an estimator $\widehat{\mu}_i$ that is close to $\overline{Y}_i$ due to the first equation.

C.2 Pseudo-posterior convergence

It remains to show that the pseudo-posterior converges asymptotically to a Gaussian distribution centered on the census maximum likelihood estimators. We rely upon the following result, originally from Kleijn and van der Vaart (2012) and reformulated for a complex sampling context in Proposition B.1 in the Appendix of Han and Wellner (2021). Note that we consider a sequence of populations indexed by ν, rather than by the population size N.

Proposition 1. Suppose the following conditions hold:

1. (Local asymptotic normality) There exist random vector $\Delta_{N_\nu,\theta^*} = \mathcal{O}_{\mathbb{P}}(1)$ and a non-singular matrix H_{θ^*} such that for every compact $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \left| N_\nu \mathbb{P}_\nu^\pi \log \frac{p_{\theta^* + h/\sqrt{N_\nu}}}{p_{\theta^*}} - h^T H_{\theta^*} \Delta_{N_\nu,\theta^*} - h^T H_{\theta^*} h \right| = o_{\mathbb{P}}(1).$$

2. (Sufficient mass condition) The prior Π on Θ has a Lebesgue density that is continuous and positive on a neighborhood of θ^*.

3. (Posterior contraction) For every $L_{N_\nu} \to \infty$,

$$P_{\theta^*} \Pi_{N_\nu}^\pi (\theta \in \Theta : ||\theta - \theta^*|| > L_n/\sqrt{N} \mid D^{(N)}) \to 0$$

Then the sampling weighted pseudo-posterior distribution $\Pi_{N_\nu}^\pi$ converges to a sequence of normal distributions in total variation:

$$\sup_B \left| \Pi_{N_\nu}^\pi (\sqrt{N_\nu}(\theta - \theta^*) \in B \mid D^{(N)}) - \mathcal{N}_{\delta_{N_\nu,\theta^*}, H_{\theta^*}^{-1}}(B) \right| = o_{\mathbb{P}}(1).$$

Williams and Savitsky provide an alternative reformulation of the above theorem for a complex survey setting (Williams and Savitsky 2021). Han and Wellner provide conditions under which the first and third conditions hold. In particular, they consider the following conditions on the design:

1. For some nonrandom $\pi_0 > 0$,

$$\min_{1 \le j \le N_\nu} \pi_j \ge \pi_0$$

2. The weights satisfy a central limit theorem

$$\frac{1}{\sqrt{N_\nu}} \sum_{j=1}^{N} \left(\frac{\delta_j}{\pi_j} - 1 \right) = \mathcal{O}_{\mathbb{P}}(1)$$

In addition, they require the conditions on the true superpopulation measure and the model, which we modify to reflect that the true superpopulation measure may not be in the model space:

1. The map $\theta \mapsto \log p_\theta(x) = \ell_\theta(x)$ is differentiable at θ^* for all x with derivative $\dot{\ell}_{\theta^*}(x)$ for θ_1, θ_2 close enough to θ^*

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \leq m(x)||\theta_1 - \theta_2||$$

   for some P_0 square integrable function m.

2. The Kullback-Leibler divergence relative to the true superpopulation measure P_0 has a second-order Taylor-expansion:

$$P_0 \log \frac{p_{\theta^*}}{p_\theta} = \frac{1}{2}(\theta - \theta^*)^\top H_{\theta^*}(\theta - \theta^*) + o(||\theta - \theta^*||^2)$$

   where H_{θ^*} is a positive-definite Hessian matrix.

The first design assumption is standard in the survey statistics literature and ensures that sampling probabilities are bounded away from zero. The second design assumption must be examined for each design. In our setting, we largely consider two-stage cluster designs in which the first-stage clusters or PSUs are sampled with probability proportional to size. However, sampling at the subsequent stage is independent between clusters and at random, so all final inclusion probabilities are equal and

$$\frac{1}{\sqrt{N_\nu}} \sum_{j=1}^{N} \left(\frac{\delta_j}{\pi_j} - 1 \right) = 0$$

C.3 Estimating the multivariate design effect

As described above, Williams and Savitsky propose a post-processing adjustment that scales the pseudo-posterior by rescaling samples:

$$\widehat{\theta}^{WS(k)} = \left(\widehat{\theta}^{(k)} - \overline{\theta} \right) R_2^{-1} R_1 + \overline{\theta} \tag{C.3}$$

where $\widehat{\theta}^{(k)}$ is the kth sample from the pseudo-posterior, $\hat{\theta}^{WS(k)}$ the kth adjusted sample, $\overline{\theta}$ is the vector mean of the samples $\widehat{\theta}^{(k)}$, $R_1^T R_1 = H_{\theta^*}^{-1} J_{\theta^*}^{\pi} H_{\theta^*}^{-1}$, and $R_2^T R_2 = H_{\theta^*}^{-1}$. We define H_{θ^*} as

$$H_{\theta^*} = -\frac{1}{N_\nu} \sum_{j \in U_\nu} \mathbb{E}_{P_{\theta^*}} \ddot{\ell}_{\theta^*}(y_j, \mathbf{x}_j) \tag{C.4}$$

and $J_{\theta^*}^{\pi}$ is the variance matrix of the weighted score functions under $\mathbb{P}$:

$$J_{\theta^*}^{\pi} = \mathbb{E}_{\theta^*, \nu} \left[\mathbb{P}_\nu^{\pi} \dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T \right] \tag{C.5}$$

In practice, the matrices H_{θ^*} and $J_{\theta^*}^{\pi}$ can be estimated via Algorithm 1 from Williams and Savitsky (2021). Based on the model (5.2) with fixed area-specific intercepts, we obtain sample vectors $\widehat{\theta}^{(k)} = (\beta_{0i}^{(k)}, \boldsymbol{\beta}_1^{(k)}, \sigma_\varepsilon^{2(k)})$. We can then compute the mean of the pseudo-posterior draws $\overline{\theta}$. We compute $\widehat{H}_{\theta^*}$ as the negative Hessian of the weighted log-likelihood for model (5.2) with fixed β_{0i} at the pseudo-MLE.

We estimate $J_{\theta^*}^{\pi}$ using a resampling approach that repeatedly subsamples PSUs without replacement, within strata, from the sample (Preston 2009). For each subsample, we can compute a weighted score function. We then compute the sample covariance of the weighted score functions, across 100 subsamples, to obtain an estimator $\widehat{J}_{\theta^*}^{\pi}$. Finally, we can compute $\widehat{R}_1$ and $\widehat{R}_2$ via Cholesky decomposition, plugging in $\widehat{H}_{\theta^*}$ and $\widehat{J}_{\theta^*}^{\pi}$ in the definitions of $R_2^T R_2$ and $R_2^T R_2$.

C.4 Estimation procedure

The simulations and application described above were implemented in the R programming language (R Core Team 2023). For general data cleaning and processing covariate information, we used the `tidyverse` (Wickham 2014), `sf` (Pebesma 2018), and `terra` (Hijmans 2023) packages. For computing the design-based Hájek and GREG estimators, in addition to carrying out the resampling for estimating the multivariate design, we use the `survey` package (Lumley 2011).

We approximate the unscaled pseudo-posterior distributions using the `INLA` package (Rue et al. 2017), which enables the user to input weights when carrying out Bayesian inference. We obtain samples from the approximated pseudo-posterior distributions, and then transform and rescale the samples as described above. Credible sets are constructed by taking relevant quantiles of the rescaled samples for μ_i for all i.

For all unit level models, we place a flat prior on the intercept and fixed effects, so $\pi(\boldsymbol{\beta}_1) \propto 1$. We use penalized complexity priors for the variance parameters which place a prior on the Kullback-Leibler distance between a full model to a simplified base model, shrinking variance components σ_ε and σ_u to zero (Simpson et al. 2017). In particular, we specify the prior for σ_u^2 and σ_ε^2 such that $P(\sigma_u > 3) = 0.05$ and $P(\sigma_\varepsilon > 3) = 0.05$.

REFERENCES

Arora, V., and Lahiri, P. (1997), "On the superiority of the Bayesian method over the BLUP in small area estimation problems," *Statistica Sinica*, 7, 1053–1063.

Asparouhov, T. (2006), "General multi-level modeling with sampling weights," *Communications in Statistics - Theory and Methods*, 35, 439–460.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83, 28–36. https://doi.org/10.2307/2288915.

Bell, W. R. (2008), *Examining sensitivity of small area inferences to uncertainty about sampling error variances*, United States Census Bureau.

Bell, W. R., Basel, W. W., and Maples, J. J. (2016), "An overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program," in *Analysis of Poverty Data by Small Area Estimation*, John Wiley & Sons, Ltd, pp. 349–378.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–20. https://doi.org/10.1007/BF00116466.

Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review / Revue Internationale de Statistique*, 51, 279–292. https://doi.org/10.2307/1402588.

Breidt, F. J., and Opsomer, J. D. (2017), "Model-assisted survey estimation with modern prediction techniques," *Statistical Science*, 32, 190–205. https://doi.org/10.1214/16-STS589.

Brewer, K. R. W. (1999), "Design-based or prediction-based inference? Stratified random vs stratified balanced sampling," *International Statistical Review / Revue Internationale de Statistique*, 67, 35–47. https://doi.org/10.2307/1403564.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. https://doi.org/10.18637/jss.v076.i01.

Chen, C., Wakefield, J., and Lumley, T. (2014), "The use of sampling weights in Bayesian hierarchical models for small area estimation," *Spatial and Spatio-temporal Epidemiology*, 11, 33–43. https://doi.org/10.1016/j.sste.2014.07.002.

Chung, H. C., and Datta, G. S. (2020), *Bayesian hierarchical spatial models for small area estimation*, Center for Statistical Research & Methodology, Research; Methodology Directorate, U.S. Census Bureau.

Congdon, P., and Lloyd, P. (2010), "Estimating small area diabetes prevalence in the US using the Behavioral Risk Factor Surveillance System," *Journal of Data Science*, 8, 235–252. https://doi.org/10.6339/JDS.2010.08(2).583.

Corral, P., Himelein, K., McGee, K., and Molina, I. (2021), "A map of the poor or a poor map?" *Mathematics*, 9, 2780. https://doi.org/10.3390/math9212780.

Corral Rodas, P., Molina, I., and Nguyen, M. (2021), "Pull your small area estimates up by the bootstraps," *Journal of Statistical Computation and Simulation*, 91, 3304–3357. https://doi.org/10.1080/00949655.2021.1926460.

Dass, S. C., Maiti, T., Ren, H., and Sinha, S. (2012), "Confidence interval estimation of small area parameters shrinking both means and variances," *Survey Methodology*, 38, 173–187.

Database of Global Administrative Areas (GADM) (2022), "Database of Global Administrative Areas 4.1."

Diggle, P. J., and Giorgi, E. (2016), "Model-based geostatistics for prevalence mapping in low-resource settings," *Journal of the American Statistical Association*, 111, 1096–1120. https://doi.org/10.1080/01621459.2015.1123158.

Dong, T. Q., and Wakefield, J. (2021), "Modeling and presentation of vaccination coverage estimates using data from household surveys," *Vaccine*, 39, 2584–2594. https://doi.org/10.1016/j.vaccine.2021.03.007.

Efron, B., and Morris, C. (1975), "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311–319. https://doi.org/10.2307/2285814.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2019), "Model-based county level crop estimates incorporating auxiliary sources of information," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 283–303. https://doi.org/10.1111/rssa.12390.

Faraway, J. J. (2014), *Linear Models with R, Second Edition*, CRC Press.

Fay, R. E., and Herriot, R. A. (1979), "Estimates of income for small places: An application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277. https://doi.org/10.2307/2286322.

Franco, C., and Bell, W. R. (2013), "Applying bivariate binomial/logit normal models to small area estimation," in *Proceedings of the American Statistical Association, Survey Research Section*, pp. 690–702.

Fuglstad, G.-A., Li, Z. R., and Wakefield, J. (2022), "The two cultures for prevalence mapping: Small area estimation and spatial statistics," *arXiv:2110.09576 [stat]*.

Gao, P. A., and Wakefield, J. (2023+), "Smoothed model-assisted small area estimation," *to appear in Canadian Journal of Statistics*.

Gao, P. A., and Wakefield, J. (2022), "A spatial variance-smoothing area level model for small area estimation of demographic rates," arXiv. https://doi.org/10.48550/

arXiv.2209.02602.

Gelman, A. (2006), "Multilevel (hierarchical) modeling: What it can and cannot do," *Technometrics*, 48, 432–435. https://doi.org/10.1198/004017005000000661.

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Ghosh, M. (2020), "Small area estimation: Its evolution in five decades," *Statistics in Transition*, 21. https://doi.org/10.21307/stattrans-2020-022.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998), "Generalized linear models for small-area estimation," *Journal of the American Statistical Association*, 93, 273–282. https://doi.org/10.2307/2669623.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., Dwyer-Lindgren, L., Farag, T. H., Flaxman, A. D., Fraser, M. S., Gething, P. W., Gibson, H. S., Graetz, N., Krause, L. K., Kulikoff, X. R., Lim, S. S., Mappin, B., Morozoff, C., Reiner, R. C., Sligar, A., Smith, D. L., Wang, H., Weiss, D. J., Murray, C. J. L., Moyes, C. L., and Hay, S. I. (2017), "Mapping under-5 and neonatal mortality in Africa, 2000–15: A baseline analysis for the Sustainable Development Goals," *The Lancet*, 390, 2171–2182. https://doi.org/10.1016/S0140-6736(17)31758-0.

Goldstein, H. (2010), "The 2-level model," in *Multilevel Statistical Models*, John Wiley & Sons, Ltd, pp. 15–72. https://doi.org/10.1002/9780470973394.ch2.

Goldstein, H., and Spiegelhalter, D. J. (1996), "League tables and their limitations: Statistical issues in comparisons of institutional performance," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 385–443. https://doi.org/10.2307/2983325.

Hájek, J. (1971), "Discussion of "An essay on the logical foundations of survey sampling, part I" , by D. Basu." in *Foundations of Statistical Inference*, eds. V. P. Godambe and D.

A. Sprott, Holt, Rinehart; Winston.

Han, Q., and Wellner, J. A. (2021), "Complex sampling designs: Uniform limit theorems and applications," *The Annals of Statistics*, 49, 459–485. https://doi.org/10.1214/20-AOS1964.

Hartley, H. O., and Sielken, R. L. (1975), "A "super-population viewpoint" for finite population sampling," *Biometrics*, 31, 411–422. https://doi.org/10.2307/2529429.

Harville, D. (1976), "Extension of the Gauss-Markov theorem to include the estimation of random effects," *The Annals of Statistics*, 4, 384–395. https://doi.org/10.1214/aos/1176343414.

Hawala, S., and Lahiri, P. (2018), "Variance modeling for domains," *Statistics and Applications*, 16, 399–409.

Hijmans, R. J. (2023), *Terra: Spatial data analysis*.

Hijmans, R. J., Etten, J. van, Sumner, M., Cheng, J., Baston, D., Bevan, A., Bivand, R., Busetto, L., Canty, M., Fasoli, B., Forrest, D., Ghosh, A., Golicher, D., Gray, J., Greenberg, J. A., Hiemstra, P., Hingee, K., Geosciences, I. for M. A., Karney, C., Mattiuzzi, M., Mosher, S., Naimi, B., Nowosad, J., Pebesma, E., Lamigueiro, O. P., Racine, E. B., Rowlingson, B., Shortridge, A., Venables, B., and Wueest, R. (2021), *Raster: Geographic data analysis and modeling*.

Hirose, M., Ghosh, M., and Ghosh, T. (2023), "Arc-Sin transformation for binomial sample proportions in small area estimation," *Statistica Sinica*. https://doi.org/10.5705/ss.202020.0446.

Hodges, J. S. (2016), *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, CRC Press.

Horvitz, D. G., and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685. https://doi.org/10.2307/2280784.

Huang, X. (2019), "Mixed models for complex survey data," Thesis, University of Auckland.

Hwang, J. T. G., Qiu, J., and Zhao, Z. (2009), "Empirical Bayes confidence intervals shrinking both means and variances," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71, 265–285.

Institut National de la Statistique and ICF (2019), *Guinea Demographic and Health Survey (EDS V) 2016-18*, Conakry, Guinea: INS/Guinea; ICF.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Science & Business Media.

Jiang, J. (2017), *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*, New York: Chapman; Hall/CRC. https://doi.org/10.1201/9781315119281.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for standard errors of estimators of fixed and random effect in mixed linear models," *Journal of the American Statistical Association*, 79, 853–862. https://doi.org/10.2307/2288715.

Kennel, T. L., and Valliant, R. (2010), "Logistic Generalized Regression (LGREG) estimator in cluster samples," in *Section on Survey Research Methods*, Vancouver.

Kennel, T. L., and Valliant, R. (2020), "Multivariate logistic-assisted estimators of totals from clustered survey samples," *Journal of Survey Statistics and Methodology*, 0, 1–35. https://doi.org/10.1093/jssam/smaa017.

Kish, L. (1965), *Survey Sampling*, Wiley.

Kleffe, J., and Rao, J. N. K. (1992), "Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model," *Journal of Multivariate Analysis*, 43, 1–15. https://doi.org/10.1016/0047-259X(92)90107-Q.

Kleijn, B. J. K., and Vaart, A. W. van der (2012), "The Bernstein-Von-Mises theorem under

misspecification," *Electronic Journal of Statistics*, 6, 354–381. https://doi.org/10.1214/12-EJS675.

Korn, E. L., and Graubard, B. I. (2003), "Estimating variance components by using survey data," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65, 175–190.

Laird, N. M., and Ware, J. H. (1982), "Random-effects models for longitudinal data," *Biometrics*, 38, 963–974. https://doi.org/10.2307/2529876.

Lehtonen, R., and Veijanen, A. (1998), "Logistic generalized regression estimators," *Survey Methodology*, 24, 51–55.

León-Novelo, L. G., and Savitsky, T. D. (2019), "Fully Bayesian estimation under informative sampling," *Electronic Journal of Statistics*, 13, 1608–1645. https://doi.org/10.1214/19-EJS1538.

Li, Z. R., Martin, B. D., Hsiao, Y., Godwin, J., Paige, J., Gao, P., Wakefield, J., Clark, S. J., Fuglstad, G.-A., and Riebler, A. (2022), *SUMMER: Small-area-estimation unit/area models and methods for estimation in r*.

Lindgren, F., and Rue, H. (2015), "Bayesian spatial modelling with R-INLA," *Journal of Statistical Software*, 63, 1–25. https://doi.org/10.18637/jss.v063.i19.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 1–41.

Little, R. J. (2004), "To model or not to model? Competing modes of inference for finite population sampling," *Journal of the American Statistical Association*, 99, 546–556.

Liu, B., Lahiri, P., and Kalton, G. (2014), "Hierarchical Bayes modeling of survey-weighted small area proportions," *Survey Methodology*, 40, 1–13.

Lohr, S. L. (2019), *Sampling : Design and Analysis*, Chapman; Hall/CRC. https://doi.org/10.1201/9780429296284.

Lumley, T. (2004), "Analysis of complex survey samples," *Journal of Statistical Software*, 9, 1–19. https://doi.org/10.18637/jss.v009.i08.

Lumley, T. (2011), *Complex Surveys: A Guide to Analysis Using R*, John Wiley & Sons.

Lumley, T., and Scott, A. (2017), "Fitting regression models to survey data," *Statistical Science*, 32, 265–278. https://doi.org/10.1214/16-STS605.

Lumley, T., Shaw, P. A., Dai, J. Y., Tsiatis, A. A., Davidian, M., Handcock, M. S., Lawless, J. F., Kalbfleisch, J. D., Scott, A. J., and Wild, C. J. (2011), "Connections between survey calibration estimators and semiparametric models for incomplete data [with discussions]," *International Statistical Review / Revue Internationale de Statistique*, 79, 200–232.

Maiti, T., Ren, H., and Sinha, S. (2014), "Prediction error of small area predictors shrinking both means and variances," *Scandinavian Journal of Statistics*, 41, 775–790.

Maples, J. J. (2016), "Estimating design effects in small areas/domains through aggregation," in *Proceedings of the American Statistical Association, Survey Research Section*, pp. 671–681.

Maples, J. J., Bell, W. R., and Huang, E. T. (2009), "Small-area variance modeling with application to county poverty estimates from the American Community Survey," 12.

Marhuenda, Y., Molina, I., and Morales, D. (2013), "Small area estimation with spatio-temporal Fay–Herriot models," *Computational Statistics & Data Analysis*, The Third Special Issue on Statistical Signal Extraction and Filtering, 58, 308–325. https://doi.org/10.1016/j.csda.2012.09.002.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J. N. K. (2017), "Poverty mapping in

small areas under a twofold nested error regression model," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 1111–1136. https://doi.org/10.1111/rssa.12306.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015), "Space-time smoothing of complex survey data: Small area estimation for child mortality," *The Annals of Applied Statistics*, 9, 1889–1905. https://doi.org/10.1214/15-AOAS872.

Miller, J. W. (2021), "Asymptotic normality, concentration, and coverage of generalized posteriors," *The Journal of Machine Learning Research*, 22, 168:7598–168:7650.

Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., and Kerckhove, W. V. de (2012), "Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models," *Journal of the Indian Society of Agricultural Statistics*, 55–63.

Myrskylä, M. (2007), "Generalised regression estimation for domain class frequencies," PhD thesis,.

National Population Commission - NPC/Nigeria and ICF. (2019), *Nigeria Demographic and Health Survey 2018*, Abuja, Nigeria,; Rockville, Maryland: NPC; ICF.

NSO/Malawi and ICF, N. S. O. - (2017), *Malawi Demographic and Health Survey 2015-16*, Zomba, Malawi: NSO; ICF.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2022a), "Design- and model-based approaches to small-area estimation in a low- and middle-income country context: Comparisons and recommendations," *Journal of Survey Statistics and Methodology*, 10, 50–80. https://doi.org/10.1093/jssam/smaa011.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2022b), "Spatial aggregation with respect to a population distribution: Impact on inference," *Spatial Statistics*, 52, 100714. https://doi.org/10.1016/j.spasta.2022.100714.

Parker, P. A., Holan, S. H., and Janicki, R. (2022), "Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling with application to estimation of health insurance coverage," *The Annals of Applied Statistics*, 16, 887–904. https://doi.org/10.1214/21-AOAS1524.

Parker, P. A., Janicki, R., and Holan, S. H. (2020), "Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions," *arXiv:1908.10488 [stat]*.

Pebesma, E. (2018), "Simple Features for R: Standardized Support for Spatial Vector Data," *The R Journal*, 10, 439–446. https://doi.org/10.32614/RJ-2018-009.

Petrucci, A., and Salvati, N. (2006), "Small area estimation for spatial correlation in watershed erosion assessment," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 169. https://doi.org/10.1198/108571106X110531.

Pfeffermann, D. (2013), "New important developments in small area estimation," *Statistical Science*, 28, 40–68. https://doi.org/10.1214/12-STS395.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for unequal selection probabilities in multilevel models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 23–40.

Pfeffermann, D., and Sverchkov, M. (2007), "Small-area estimation under informative probability sampling of areas and within the selected areas," *Journal of the American Statistical Association*, 102, 1427–1439. https://doi.org/10.1198/016214507000001094.

Polettini, S. (2017), "A generalised semiparametric Bayesian Fay–Herriot model for small area estimation shrinking both means and variances," *Bayesian Analysis*, 12, 729–752.

Porter, A. T., Holan, S. H., Wikle, C. K., and Cressie, N. (2014), "Spatial Fay–Herriot models for small area estimation with functional covariates," *Spatial Statistics*, 10, 27–42. https://doi.org/10.1016/j.spasta.2014.07.001.

Pratesi, M., and Salvati, N. (2008), "Small area estimation: the EBLUP estimator based on spatially correlated random area effects," *Statistical Methods and Applications*, 17, 113–141. https://doi.org/10.1007/s10260-007-0061-9.

Preston, J. (2009), "Rescaled bootstrap for stratified multistage sampling," *Survey Methodology*, 35, 227–234.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rabe-Hesketh, S., and Skrondal, A. (2006), "Multilevel modelling of complex survey data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 805–827. https://doi.org/10.1111/j.1467-985X.2006.00426.x.

Rao, J. N. K., and Molina, I. (2015), *Small Area Estimation*, Wiley Series in Survey Methodology, John Wiley & Sons.

Rao, J. N. K., Verret, F., and Hidiroglou, M. (2013), "A weighted composite likelihood approach to inference for two-level models from survey data," *Survey Methodology*, 39, 263–282.

Raudenbush, S. W., and Willms, J. D. (1995), "The estimation of school effects," *Journal of Educational and Behavioral Statistics*, 20, 307–335. https://doi.org/10.2307/1165304.

Ribatet, M., Cooley, D., and Davison, A. C. (2012), "Bayesian inference from composite likelihoods, with an application to spatial extremes," *Statistica Sinica*, 22, 813–845.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016), "An intuitive Bayesian spatial model for disease mapping that accounts for scaling," *Statistical Methods in Medical Research*, 25, 1145–1165. https://doi.org/10.1177/0962280216660421.

Rivest, L.-P., and Vandal, N. (2002), "Mean squared error estimation for small areas when the small area variances are estimated," in *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Ottawa, Canada.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866. https://doi.org/10.2307/2290910.

Robinson, G. K. (1991), "That BLUP is a good thing: The estimation of random effects," *Statistical Science*, 6, 15–32. https://doi.org/10.1214/ss/1177011926.

Royall, R. M. (1970), "On finite population sampling theory under certain linear regression models," *Biometrika*, 57, 377–387. https://doi.org/10.2307/2334846.

Royall, R. M. (1992), "The model based (prediction) approach to finite population sampling theory," in *Current issues in statistical inference: Essays in honor of D. Basu*, Institute of Mathematical Statistics.

Royall, R. M., and Herson, J. (1973), "Robust estimation in finite populations I," *Journal of the American Statistical Association*, 68, 880–889. https://doi.org/10.2307/2284516.

Rubin-Bleuer, S., and Schiopu-Kratina, I. (2005), "On the two-phase framework for joint model and design-based inference," *The Annals of Statistics*, 33, 2789–2810. https://doi.org/10.1214/009053605000000651.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017), "Bayesian computing with INLA: A review," *Annual Review of Statistics and Its Application*, 4, 395–421.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003), *Model Assisted Survey Sampling*, Springer Science & Business Media.

Savitsky, T. D., and Toth, D. (2016), "Bayesian estimation under informative sampling," *Electronic Journal of Statistics*, 10, 1677–1708. https://doi.org/10.1214/16-EJS1153.

Savitsky, T. D., and Williams, M. R. (2022), "Pseudo Bayesian mixed models under informative sampling," *Journal of Official Statistics*, 38, 901–928. https://doi.org/10.

2478/jos-2022-0039.

Si, Y., Trangucci, R., Gabry, J. S., and Gelman, A. (2020), "Bayesian hierarchical weighting adjustment and survey inference," *Survey Methodology*, 46, 181–214.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017), "Penalising model component complexity: A principled, practical approach to constructing priors," *Statistical Science*, 32, 1–28. https://doi.org/10.1214/16-STS576.

Skinner, C. J. (1989), "Domain means, regression and multivariate analysis." eds. C. J. Skinner, D. Holt, and T. M. F. Smith, Chichester, UK: Wiley, pp. 59–87.

Skrondal, A., and Rabe-Hesketh, S. (2009), "Prediction in multilevel generalized linear models," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 659–687. https://doi.org/10.1111/j.1467-985X.2009.00587.x.

Slud, E. V. (2020), *Model-assisted estimation of mixed-effect model parameters in complex surveys*, Center for Statistical Research & Methodology.

Smith, A. F. M. (1973), "A general Bayesian linear model," *Journal of the Royal Statistical Society. Series B (Methodological)*, 35, 67–75.

Smith, T. M. F. (1976), "The foundations of survey sampling: A review," *Journal of the Royal Statistical Society. Series A (General)*, 139, 183–204. https://doi.org/10.2307/2345174.

Smith, T. M. F. (1994), "Sample surveys 1975-1990; an age of reconciliation?" *International Statistical Review / Revue Internationale de Statistique*, 62, 5–19. https://doi.org/10.2307/1403539.

Sugasawa, S., and Kubokawa, T. (2020), "Small area estimation with mixed models: A review," *Japanese Journal of Statistics and Data Science*. https://doi.org/10.1007/s42081-020-00076-x.

Sugasawa, S., Kubokawa, T., and Rao, J. N. K. (2018), "Small area estimation via un-

matched sampling and linking models," *TEST*, 27, 407–427.

Sugasawa, S., Tamae, H., and Kubokawa, T. (2017), "Bayesian estimators for small area models shrinking both means and variances," *Scandinavian Journal of Statistics*, 44, 150–167.

Tatem, A. J., Gething, P. W., Bhatt, S., Weiss, D., and Pezzulo, C. (2017), "Pilot high resolution poverty maps," University of Southampton/Oxford.

Thompson, M. E., Sedransk, J., Fang, J., and Yi, G. Y. (2022), "Bayesian inference for a variance component model using pairwise composite likelihood with survey data," *Survey Methodology*, 48, 73–93.

Tillé, Y., and Matei, A. (2021), *Sampling: Survey sampling*.

Torabi, M., and Rao, J. N. K. (2010), "Mean squared error estimators of small area means using survey weights," *Canadian Journal of Statistics*, 38, 598–608. https://doi.org/10.1002/cjs.10078.

United Nations (2015), *Transforming our World: The 2030 Agenda for Sustainable Development*.

United Nations Inter-agency Group for Child, and Mortality Estimation (2021), *Subnational under-five mortality estimates, 1990–2019*, New York.

Utazi, C. E., Wagai, J., Pannell, O., Cutts, F. T., Rhoda, D. A., Ferrari, M. J., Dieng, B., Oteri, J., Danovaro-Holliday, M. C., Adeniran, A., and Tatem, A. J. (2020), "Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys," *Vaccine*, 38, 3062–3071. https://doi.org/10.1016/j.vaccine.2020.02.070.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite population sampling and inference: A prediction approach*, Wiley.

Wakefield, J., Okonek, T., and Pedersen, J. (2020), "Small area estimation for disease preva-

lence mapping," *International Statistical Review*, 88, 398–418. https://doi.org/10.1111/insr.12400.

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

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S., Kang, S. Y., Cameron, E., Bisanzio, D., Battle, K. E., Bhatt, S., and Gething, P. W. (2018), "A global map of travel time to cities to assess inequalities in accessibility in 2015," *Nature*, 553, 333–336. https://doi.org/10.1038/nature25181.

Wickham, H. (2014), "Tidy Data," *Journal of Statistical Software*, 59, 1–23.

Williams, M. R., and Savitsky, T. D. (2021), "Uncertainty estimation for pseudo-bayesian inference under complex sampling," *International Statistical Review*, 89, 72–107. https://doi.org/10.1111/insr.12376.

Wolter, K. (2007), *Introduction to Variance Estimation*, Springer Science & Business Media.

WorldPop (2018a), "Global 100m Covariates," University of Southampton. https://doi.org/10.5258/SOTON/WP00644.

WorldPop (2018b), "Global 100m Age/Sex Structures," University of Southampton. https://doi.org/10.5258/SOTON/WP00646.

WorldPop (2020), "Global 100m Population total adjusted to match the corresponding UNPD estimate," University of Southampton. https://doi.org/10.5258/SOTON/WP00660.

WorldPop, and Center for International Earth Science Information Network (CIESIN), Columbia University (2006), "Global high resolution population denominators project - funded by the bill and melinda gates foundation (OPP1134076)."

Yi, G. Y., Rao, J. N. K., and Li, H. (2016), "A weighted composite likelihood approach for analysis of survey data under two-level models," *Statistica Sinica*, 26, 569–587.

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

You, Y., and Rao, J. N. K. (2002), "A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30, 431–439. https://doi.org/10.2307/3316146.