Bayesian Small Area Estimation Methods for Sparse Complex Survey Data

Alana McGovern

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

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Jon Wakefield, Chair
Geir-Arne Fuglstad
Katherine Wilson

Program Authorized to Offer Degree
Statistics

University of Washington

Abstract

Bayesian Small Area Estimation Methods for Sparse Complex Survey Data

Alana McGovern

Chair of the Supervisory Committee:

Jon Wakefield

Department of Statistics and Biostatistics

Small area estimation (SAE) of health and demographic outcomes in low- and middle-income countries (LMICs) is central to policy evaluation and resource allocation, yet it is complicated by the sparsity and complex design of the survey data on which such estimates typically rely. Household surveys like the Demographic and Health Surveys (DHS) are generally powered at the first administrative (admin1) level, so design-based estimates at lower levels are often imprecise or unavailable, particularly for rare outcomes. Model-based approaches can address this sparsity by borrowing strength across areas through spatial random effects, but this borrowing can induce extreme shrinkage when data are limited, affecting both point and variance estimates. This dissertation develops a set of Bayesian hierarchical spatial models which address instability and excessive shrinkage in SAE, motivated throughout by subnational estimation of child health outcomes in LMICs.

The first methodological contribution is the direct-assisted Bayesian unit-level (DABUL) model, designed for estimating rare event prevalence at an administrative level too sparse to support standard area-level models. The DABUL model incorporates higher-level design-based estimates from the same underlying data as benchmarks via either a hard constraint or a soft constraint which accounts for uncertainty of the benchmarks. By applying this method to simulated data and neonatal mortality rate (NMR) estimation in Zambia using the 2013 DHS, we show that the soft DABUL model reduces aggregation discrepancy and

both variants of the DABUL model yield more conservative intervals than those under a standard unit-level model, which may be particularly preferable for extreme areas.

The second contribution addresses a different source of instability. The design-based variance estimates used as input in Fay-Herriot (FH) models are themselves estimated with error which is frequently ignored despite their affect on interval estimation, particularly when sample size is low. Two candidate sampling distributions for the design variance estimator are proposed and incorporated into a variance-smoothing FH framework. One sampling distribution incorporates complex survey design assumptions, while the other, simpler sampling distribution relies on stronger design assumptions. A simulation study and application to height-for-age z-scores in Kenya using the 2022 DHS show that both variance-smoothing approaches outperform a standard FH model without variance smoothing under proper scoring rules, with the simpler sampling distribution recommended for its more favorable shrinkage behavior and ease of implementation.

The third contribution extends multivariate areal spatial modeling frameworks from the disease mapping literature to the FH setting and proposes a class of multivariate latent models using a BYM2 spatial structure that includes independent, correlated, and shared-component variants. Theoretical results characterize how the conditional posterior mean and variance of a primary outcome depend on the relationship between sampling and latent covariance structures. A simulation study and application to NMR estimation in Kenya using the 2022 DHS show that joint modeling with correlated outcomes can substantially improve estimation of a noisy primary outcome relative to a univariate FH model, particularly when latent correlation exceeds sampling correlation or auxiliary outcomes carry lower uncertainty than the primary outcome.

Together, these three contributions demonstrate that explicitly modeling design-based sampling error and structured dependence across administrative levels, areas, or outcomes, offers a coherent strategy for improving SAE for LMICs in the data-sparse settings characteristic of complex surveys.

Contents

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Admin1: First administrative

Admin2: Second administrative

DABUL: Direct-assisted Bayesian unit-level

DCM: Dirichlet compound multinomial

BYM: Besag-York-Mollíe

BYM2: Besag-York-Mollíe 2

DHS: Demographic Health Surveys

ICAR: Intrinsically conditional autoregressive

IID: Independent and identically distributed

FH: Fay-Herriot

GMBYM2: Generalized multivariate Besag-York-Mollíe 2

GVF: Generalized variance function

HAZ: Height-for-age z-score

LMIC: Low-and middle-income country

MICS: Multiple Indicator Cluster Surveys

NMR: Neonatal mortality rate

PC: Penalize complexity

PPS: Probability proportional to size

RMSE: Root mean squared error

SAE: Small area estimation

SASW: Satterthwhaite-approximated survey-weighted

SCM: Shared component model

SDGs: Sustainable Development Goals

WHO: World Health Organization

DEDICATION

To my mom and dad, for their dedicated, generous spirits which inspire me every day.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Jon Wakefield, without whom this work would not be possible. Thank you for answering the email I sent you my very first week in the department and supporting me every day since. I cannot imagine having reached this major accomplishment without having you as my advisor. I would also like to thank the rest of my reading committee, Geir-Arne Fuglstad and Katherine Wilson, as well as my Graduate Student Representative, Elizabeth Sanders. I offer a special thank you to Geir-Arne for his generosity in our ongoing collaborations. I always appreciate your insights and attention to detail. I would also like to thank all the members of the Space Time Analysis Bayesian (STAB) research group, past and present, for their generous mentorship and thoughtful discussions which have helped me develop as a researcher.

Thank you to the entire community at UW for their guidance and support. To the department staff, Ellen Reynolds, Kristine Chan, Tracy Pham, Veronica Bae, Vickie Graybeal, and Asa Sourdiffe: thank you for all the work you do to keep everything running smoothly. Thank you to all my friends and colleagues in the Statistics and Biostatistics departments, including Medha Agarwal, Andrea Boskovic, James Buenfil, Ameer Dharamshi, Jillian Fisher, Nina Galanter, Ellen Graham, Saksham Jain, Shirley Mathur, Ronak Mehta, Alan Min, Jess Phillips, Anupreet Porwal, Shreya Prakash, and Apara Venkateswaran. This accomplishment would not have been possible without all of your generous collaboration and support. I am humbled to be a member of such an incredible cohort of statisticians and I cannot wait to see what you all accomplish.

Thank you to all of my friends in Seattle for all of the support and great times over these last five years. In particular, I want to thank Danielle for being my home away from home; I'm so glad our paths crossed again on the other side of the country. A huge thank you to the dance community at eXit SPACE for providing me with a much-needed creative outlet and taking my mind out of my research, if only for 90 minutes at a time. I would also like to thank the lovely, hardworking baristas at Stone's Throw and Diva Espresso in Wallingford.

I want to thank all of my long-distance friends, including Noah, Jess, Aliye, Meghan, Karina, and Marisa. Thank you for all the silly texts, late-night phone calls, and weekend trips; for checking in on me to make sure my research didn't swallow me whole; and for making the distance not feel so far.

I would also like to thank some of the educators and mentors whose paths I crossed years ago, but were so influential in my being here today. Thank you to the teachers at Barnert Temple, for nurturing my inquisitive nature; to my 5th grade teacher, Gail Cordello, for supporting me; to my high school calculus teacher, Mark Jacobus, for seeing my aptitude for math before I did; to Dr. Kalyani Madhu, for encouraging me to pursue academia; and to Dr. Clement Ma, for guiding me through the beginning of my research career.

Most of all, I want to thank my family for all their love and support. Thank you for taking my move across the country in stride and facetiming me into every gathering. Thank you to my sister, Nicole, for the random texts, all the trips out to Seattle, and for always knowing exactly what I mean. I'm not sure how you're meant to properly thank the woman who raised you with more love, support, and strength than even seems possible. Thank you for encouraging me every step of the way and always being a soft place to land. Thank you to my dad, for setting such a good example of how to live a life of kindness, hard work, and fun, even with the little time he had. As with everything else I accomplish in my life, this work is dedicated to your memory. Thank you to my Nana and Pop-pop for always cheering me on in that unrelenting way only grandparents can. I know if you were here today you would be determined to read this dissertation front to back, no matter how many tries it took. Finally, thank you to my beloved rabbit, Theo, for being the best listener and always reminding me when it's time for dinner. We did it, Bubby!

Chapter 1

INTRODUCTION

1.1 Motivation

Estimation of health and demographic outcomes in low- and middle-income countries (LMICs) is essential for goal-setting and policy evaluation for both national statistical agencies and international organizations, like the World Health Organization (WHO). Subnational estimation is of particular importance because understanding the variability across a country can allow for more targeted assessments and interventions. The context of subnational estimation in LMICs introduces distinct challenges because we often do not have access to high-quality, high-resolution data that may be available in higher income countries (AbouZahr et al., 2015). We often must rely on household surveys like the Demographic Health Surveys (DHS; Corsi et al., 2012) and the Multiple Indicator Clusters Surveys (MICS; UNICEF, 2026) for subnational estimation. These surveys are generally powered at the first administrative (admin1) area level, so we can usually obtain reliable design-based estimates at this level. However, these design-based estimates can be very imprecise, especially if the outcome is rare or has a high rate of missingness (Wakefield et al., 2019). Design-based estimates at a lower level, such as the second administrative (admin2) level, are generally not reliable due to missingness and high uncertainty. In these cases, we turn to model-based estimation methods, which use area-level random effects, and covariates when available, to borrow information across areas through latent spatial models (Rao and Molina, 2015; Besag et al., 1991).

The inclusion of area-level random effects induces shrinkage, which can be useful in attenuating noisy data. However, when data are particularly sparse, the shrinkage effect can be too extreme, yielding estimates which severely underestimate spatial variation (Wakefield et al., 2020). The inclusion of spatial structure in the random effects model, along with covariates, can help to reduce this overshrinkage effect. While spatial modeling is often

neglected in the SAE literature, leveraging spatial structure across areas is particularly useful in the context of LMICs because high-quality explanatory covariates are not always available (Wakefield et al., 2025). This dissertation works at the intersection of spatial and survey statistics by examining some of the challenges of SAE with sparse data and proposing several Bayesian hierarchical spatial models which address these challenges.

1.2 Surveys in low- and middle-income countries

The methodological developments in this dissertation are motivated by the challenges that arise when using complex survey data for subnational estimation. In particular, we consider the two-stage stratified cluster sampling design that is used in the DHS and MICS. Under this design, all households in a country are divided into enumeration areas, which we will simply refer to as 'clusters'. Each cluster is classified as either urban or rural and grouped into strata, which are usually defined by admin1 area crossed with urban/rural status. For example, Zambia has 10 admin1 areas and 20 sampling stratum: two for each admin1 area. In the first stage of sampling, a fixed number of clusters is sampled from each strata, with probability proportional to listed size, and in the second stage, a fixed number of households are sampled from each sampled cluster with equal probability. A household roster is created for each sampled household, which lists all usual residents and any visitors who stayed the night before. Of those listed on the household roster, each woman aged 15 to 49 is interviewed and each child under the age of 5 has biometric data recorded. In a subsample of households, adult men are also interviewed. As a result of this design, the number of clusters sampled from lower area levels is random, often leading to high data sparsity which can affect quality of estimates.

In Figure 1.1 we map the number of sampled clusters in the two datasets used in this dissertation: the 2013 Zambia DHS (ICF International, 2014) and 2022 Kenya DHS (ICF International, 2022). We observe that there is a large proportion of admin2 areas for which fewer than 5 clusters are sampled, and 6 admin2 areas in Kenya for which no clusters are

sampled. This problem of data sparsity is compounded when the outcome measure is a rare event, which we will explore in Chapters 2 and 4. For rare events, reliable estimation can be difficult at even the admin1 level, because sampled clusters do not guarantee sampled events. For example, in the 2022 Kenya DHS, each admin1 area sampled an average of 9.2 neonatal deaths across the 2018-2022 period, with 10 out of the 47 areas sampling 5 or fewer deaths.

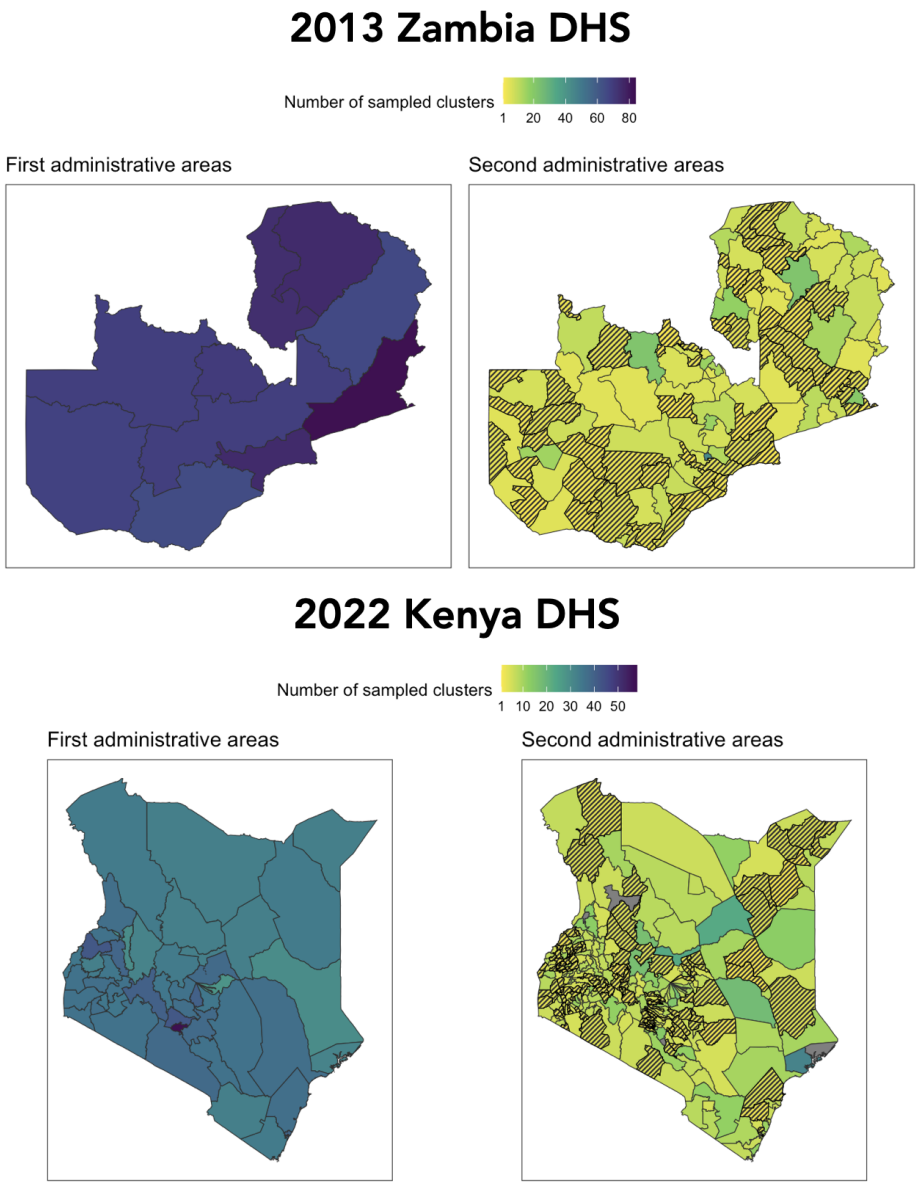


Figure 1.1. Number of sampled clusters in each admin1 and admin2 area in the 2013 Zambia and 2022 Kenya DHS. Admin2 areas with fewer than 5 sampled clusters are marked by hatching and areas with no sampled clusters are filled in gray.

Aside from the level of data sparsity that is often present in surveys in LMICs, another important consideration is accounting for the survey design itself. In particular, surveys generally oversample urban strata, so if the outcome of interest is associated with urban/rural status (as is often the case) then this must be accounted for in the model. Ideally, sampling weights should also be included in the modeling framework, as is the case in area-level models and some novel unit-level models as discussed in Section 1.3.3 and Chapter 2.

1.3 Small area estimation methods

1.3.1 Area-level models

The most commonly used area-level model in SAE is the Fay-Herriot (FH) model (Fay and Herriot, 1979), which uses design-based estimates as model input, rather than the raw data. A primary advantage of this model is that it directly accounts for the survey design (Rao and Molina, 2015). However, when data are sparse, design-based point or variance estimates can be unstable, limiting the areas for which data can be incorporated into the model. Moreover, excessive shrinkage of the point and interval estimates can occur when there is a large proportion of areas for which stable design-based estimates are not available (Wakefield et al., 2020; Gao and Wakefield, 2023).

In the sampling level of the FH model, the design-based variance is generally assumed to be known, but in practice, it is estimated from the data, often via Taylor linearization or jackknife procedures (Wolter, 2007, chap. 6; Korn and Graubard, 2011, chap. 2.4; Särndal et al., 2003, chap. 5.5; Lohr, 2021, chap. 9.1). While such estimators are design-consistent, they can be imprecise or have downward bias for finite samples, particularly when sample size is small (Särndal et al., 2003, p. 176; Lohr, 2021, p. 369). When the outcome of interest is a proportion or total, one may smooth these variance estimates using a generalized variance function (GVF), which exploits the mean-variance relationship assumed under a binomial distribution (Valliant, 1987; Wolter, 2007, chap. 7; Korn and Graubard, 2011, chap. 5.6; Lohr, 2021, chap. 9.4). Another common strategy, for any type of outcome,

is to jointly model the mean and variance. Previous work in this area uses a chi-square sampling distribution for the design variance estimator which makes strong implicit design and population assumptions (You and Chapman, 2006; Maiti et al., 2014; Sugasawa et al., 2017; Erciulescu et al., 2019). In Chapter 3 we examine the assumptions implied by this sampling distribution and compare its performance in a FH with that of a newly derived sampling distribution which more explicitly considers the complex survey design.

1.3.2 Unit-level models

When data are too sparse for an area-level model, we may consider using a unit-level model, such as a generalized linear mixed model which uses the cluster-level data as model input (Battese et al., 1988). This type of model is useful when data are too sparse to obtain design-based estimates, but generally has the limitation of not explicitly accounting for survey design (Pfeffermann, 2013; Wakefield et al., 2020). There has, however, been some development towards unit-level SAE models which incorporate survey weights directly into the model likelihood, most commonly through pseudo-EBLUP approaches that combine unit-level mixed models with design-weighted estimating equations (You and Rao, 2002; You and Rao, 2003), or through Bayesian pseudo-likelihood methods that weight unit-level likelihood contributions within a hierarchical model (Savitsky and Toth, 2016; Si et al., 2020). In Chapter 2 we propose a Bayesian hierarchical unit-level model which partially accounts for survey design by incorporating higher level design-based estimates into the likelihood.

In the latent layer of a unit-level model, one can use either area-level random effects or some continuous spatial model, such as a Matérn covariance model (Diggle and Giorgi, 2019; Paige et al., 2022), which is commonly known as a geostatistical model. We will use the former because it better serves our primary goal of obtaining the most precise administrative-level estimates possible under the constraints of data sparsity. A primary disadvantage of geostatistical models for SAE is that unit-level estimates must then be aggregated to the area level. A secondary reason we disfavor continuous spatial models in this context is

because listed cluster locations for these surveys are jittered for privacy reasons. Another consideration when using a unit-level model is that we must assume the distribution of the cluster-level data. When the outcome is continuous, a Gaussian likelihood is a logical choice, whereas for binary outcomes we may choose an overdispersed Poisson or Binomial likelihood to account for within-cluster dependence (Dong and Wakefield, 2021).

1.3.3 Current challenges and developments

In the preceding sections, we have discussed the issues of data sparsity, overshrinkage, covariate availability, and accounting for survey design. Here we discuss some other current challenges in SAE modeling, all of which are considered throughout this dissertation, where some are explored more thoroughly than others. Some of these considerations apply more specifically to the Bayesian spatial modeling context which this dissertation concerns.

Benchmarking The shrinkage bias introduced by SAE models generally yields estimates which are inconsistent across different area levels. For example, the aggregation of state-level model-based estimates may not be consistent with the national design-based estimate. This is undesirable for many of the statistical offices which use these estimates. Consistency in aggregation is not only desirable for logistical reasons, but also because estimates at higher levels are generally more precise and accurate because there are more data in each area. Many benchmarking procedures have been introduced in the SAE literature to ensure consistency of aggregated small area estimates with higher-level estimates. The majority of these procedures are for small area means (Pfeffermann and Barnard, 1991; You and Rao, 2002; Wang et al., 2008; Bell et al., 2013), while some have also been developed for generalized linear models (Datta et al., 2011; Berg and Fuller, 2018). There is also a body of work exploring the use of posterior projections to conduct Bayesian inference in constrained parameter spaces (Dunson and Neelon, 2003; Astfalck et al., 2018; De Nicolò et al., 2024). However, the majority of these benchmarking procedures do not account for the uncertainty

of the benchmarks and are sensitive to benchmark misspecification, which can be problematic when using design-based estimates as benchmarks because they are often quite noisy. Zhang and Bryant (2020), Nandram and Sayit (2011), and Okonek and Wakefield (2024) all propose inexact Bayesian benchmarking methods which can incorporate the uncertainty of the benchmarks into the posterior distribution via a soft constraint. However, these frameworks require that the benchmarks and small area estimates are produced from separate data sources. In Chapter 2 we propose a unit-level Bayesian model for rare events which uses higher-level design-based estimates from the same data as benchmarks by imposing either a hard benchmarking constraint or a soft constraint which accounts for uncertainty of the benchmarks.

Interval coverage The uncertainty intervals resulting from models with area-level random effects typically attain the nominal coverage rate only on average across areas rather than for each area individually (Yu and Hoff, 2018). As Yu and Hoff (2018) and Burris and Hoff (2020) demonstrate, areas whose means are far from the global mean can have coverage rates well below the nominal level, approaching zero in extreme cases. To address this, Yu and Hoff (2018) propose an adaptive interval procedure that extends the approach of Pratt (1963) by treating the implied distribution for the area mean as a prior and using it to determine the shape of the interval rather than the center, subject to an exact frequentist coverage constraint. This construction guarantees coverage at the nominal level for every area, albeit with wider intervals than those under the conventional random effects model. Burris and Hoff (2020) extend this framework to linking models commonly used in estimating small area means, including the spatial Fay–Herriot model.

Model validation Model validation in SAE presents a fundamental methodological challenge distinct from conventional predictive settings, as independent validation data are rarely available at the area level. External validation against census values, though widely regarded as the gold standard (Merfeld et al., 2022), is possible only for a narrow set of variables,

subject to temporal lag, and frequently absent altogether in LMICs (Dong et al., 2026). Information criteria such as AIC, DIC, and WAIC are in-sample measures that rely on asymptotic approximations to a complexity penalty which can be unreliable when the number of areas is moderate (Kawano et al., 2026). Moreover, leave-one-area-out cross-validation evaluates extrapolation to unsampled areas, rather than estimation quality for sampled areas (Marshall and Spiegelhalter, 2003; Kawano et al., 2026). Kawano et al. (2026) explore the properties of Gaussian data thinning (Neufeld et al., 2024) in the SAE setting, splitting each direct estimate into two marginally independent components that sum to the original observation, thereby creating a train-test split that does not require unit-level data. However, this method, like other area-level cross-validation methods, implicitly assumes the distribution of the design-based estimator is known, which is only true asymptotically and thus difficult to justify in the sparse data settings that motivate model-based estimation.

Computation Within Bayesian spatial modeling, there is an active area of computational research aimed at developing scalable methods for inference with large numbers of areas (Rue et al., 2009; Banerjee et al., 2015; Botella-Rocamora et al., 2015). However, in the sparse data contexts we consider here, these issues are of less practical concern. Instead, we are more burdened by the tradeoffs between data sparsity and model complexity. Perhaps the most commonly used software for Bayesian spatial modeling is the `INLA` package in `R`, which provides fast approximate inference using integrated nested Laplace approximation (Rue et al., 2009). Although `INLA` is quite useful for routine modeling, we use `Stan` software (Stan Development Team, 2024) throughout this dissertation because of its high flexibility in specifying novel modeling frameworks.

1.4 Structure of dissertation

The rest of the dissertation is structured as follows. In Chapter 2 we introduce the direct-assisted Bayesian unit-level (DABUL) model, which uses admin1-level design-based estimates

as benchmarks in a unit-level model for admin2-level rare event prevalence. We apply this model to the estimation of admin2-level neonatal mortality rate (NMR) in Zambia using the 2013 DHS. In Chapter 3 we consider the case where the design variance estimator is downwardly biased due to data sparsity, and explore sampling distributions to jointly model the design variance along with the mean in a FH model. We apply these variance smoothing FH models to the estimation of admin1-level average height-for-age z-score (HAZ) in Kenya using the 2022 DHS. In Chapter 4 we address over-shrinkage in FH models by studying the theoretical properties of multivariate FH models and extending multivariate spatial models from the disease mapping literature to the SAE context. We explore how these multivariate FH models can improve estimation of a noisy outcome when jointly modeled with other, less noisy, outcomes. We apply these multivariate FH models to the estimation of admin1-level NMR in Kenya using the 2022 DHS, and two auxiliary outcomes. We conclude with a summary of our findings and discussion of future work in Chapter 5.

Chapter 2

DIRECT-ASSISTED BAYESIAN UNIT-LEVEL MODELING FOR SMALL AREA ESTIMATION OF RARE EVENT PREVALENCE

2.1 Introduction

In terms of simplicity and fewest assumptions, the ideal method for obtaining estimates from complex surveys is a design-based weighted estimator, such as the Horvitz-Thompson (Horvitz and Thompson, 1952) or Hájek estimator (Hájek, 1971). These estimates use sampling weights which incorporate the sampling probability, and possibly non-response and post-stratification adjustments. While these design-based weighted estimators are consistent and asymptotically normal (Breidt and Opsomer, 2017), they also produce very large design-based variance estimates when data are sparse. In SAE problems there is often insufficient data to produce design-based estimates with reasonable precision, making it necessary to use a model-based approach with random effects. By borrowing information across areas, precision is increased at the cost of introducing some shrinkage bias (Datta and Ghosh, 2012; Wakefield et al., 2020). The bias introduced by these model-based approaches often makes higher level combinations of these small area estimates inconsistent in aggregate. For example, the aggregation of state-level model-based estimates may not be consistent with the national design-based estimate.

There have been many procedures introduced in the SAE literature which ensure aggregation accuracy of small area estimates to more reliable, higher-level estimates, referred to as benchmarks. There is a vast literature on this topic and we highlight a relevant subset here. One early development is the use of nested error linear regression models for estimating small area means (Pfeffermann and Barnard, 1991; You and Rao, 2002). Wang et al. (2008) derived a unique best linear unbiased estimator for small area means from augmented area-level linear models under a single benchmarking constraint, and Bell et al. (2013) extended this result to accommodate multiple benchmarking constraints. Datta et al. (2011) developed

a class of benchmarked Bayes estimators under area-level generalized linear models. All of these methods modify the best linear unbiased predictor of the small area means to achieve the benchmarking constraint, which leads to increased variance under the model (Bell et al., 2013). Berg and Fuller (2018) proposed two benchmarking procedures for nonlinear models: one which uses a linear additive adjustment, and the other which uses an augmented model for the expectation function. All of the aforementioned methods consider benchmark constraints which are estimated from the same data. There is also a body of work exploring the use of posterior projections to conduct Bayesian inference in constrained parameter spaces (Dunson and Neelon, 2003; Astfalck et al., 2018; De Nicolò et al., 2024). As Astfalck et al. (2018) notes, this framework is quite sensitive to misspecification of constraints, which implies that it would not provide reliable estimates when using design-based weighted estimates as constraints because of their large uncertainty. Zhang and Bryant (2020) and Nandram and Sayit (2011) proposed inexact Bayesian benchmarking methods which can incorporate the uncertainty of the benchmarks into the posterior distribution via a soft constraint. However, both of these frameworks require that the benchmarks and small area estimates are produced from separate data sources. Violating this assumption would yield benchmarked small area estimates that underestimate uncertainty (Okonek and Wakefield, 2024).

The contribution of this chapter is a pair of unit-level Bayesian mixed effects models with a likelihood modified to incorporate higher-level design-based direct estimates obtained from the same data source. The set of models we will introduce use design-based estimates as constraints while accounting for the fact that these constraints are estimated from the same data. As Stefan and Hidiroglou (2021) note, "statistical agencies favor an overall agreement between the sum of the model-based small area estimates and the direct estimate at a higher level that corresponds to the union of the small areas". The two models differ in that one enforces a hard constraint, while the other uses a soft constraint, accounting for the uncertainty of the benchmark. The goal of the latter is fundamentally different from exact benchmarking procedures, because its primary goal is to use higher-level estimates to

encourage consistency in aggregation while taking the uncertainty of the benchmark into account, as opposed to imposing a hard benchmark constraint. Consistency in aggregation with design-based estimates is not only desirable for logistical reasons, but also because these estimates take the survey design into account, which model-based estimates typically do not. We propose two variations of our model because while taking the uncertainty of the benchmark into account is the more principled approach, there may be practical cases in which exact benchmarking is preferred by the practitioner. We will focus on the estimation of rare event prevalence, as this is a setting in which issues of data sparsity tend to be most severe, making this direct-assisted unit-level model necessary. Examples of commonly measured outcomes that can be considered rare events, depending on the study population, include neonatal mortality, vaccination (or no vaccination) status, and HIV status. While the monitoring of these rare events at subnational levels is of substantive importance, the exceedingly small number of sampled events poses unique challenges in prevalence estimation.

In section 2.2 we will introduce the context of SAE in LMICs, which motivates the new model. In section 2.3 we will present a standard Bayesian unit-level model for rare events and in section 2.4 we will extend this model and introduce the soft and hard direct-assisted Bayesian unit-level (DABUL) models and discuss its implementation. In section 2.5 we conduct a simulation study comparing the DABUL models to a standard unit-level Bayesian model. Section 2.6 presents an application of the DABUL models to estimation of the neonatal mortality rate (NMR) in Zambia and a corresponding cross-validation study. We conclude with a discussion in Section 2.7.

2.2 Small area estimation in low- and middle-income countries

As described in Chapter 1, it is common to utilize nationally representative samples collected by the Demographic and Health Surveys (DHS) (Corsi et al., 2012) and Multiple Indicator Cluster Surveys (MICS) (UNICEF, 2026) to obtain estimates in LMICs at the admin1 or admin2 level. DHS and/or MICS carry out surveys in the majority of LMICs, often as

frequently as every five years. The DHS and MICS are conducted using a two-stage stratified cluster sampling design, where the sampling strata are defined by urban/rural crossed with the first (or for some countries, second) administrative area. In the first stage, a selection of enumeration areas (EAs) are sampled from each strata, where the probability a given EA will be selected is proportional to the number of households in that EA, relative to the others in its strata. In the second stage, a fixed number of households are sampled with equal probability from each EA. Under this design, a 'cluster' refers to either an EA or a segment of an EA.

The NMR is one of a multitude of demographic and health outcomes that is often estimated from these surveys. Accurate and precise estimation of the NMR at subnational levels is particularly important because it allows government officials at the country or regional level to evaluate which regions need more targeted interventions. The Sustainable Development Goals (SDGs) provide a target of no more than 12 deaths per 1000 live births by 2030 (`sdgs.un.org/2030agenda`). As previously stated, using a design-based weighted estimator to estimate the subnational NMR would be ideal, but is often not feasible due to small area-level sample sizes and the relative rarity of neonatal death. The sampling frame for most DHS and MICS are powered at the admin1 level, so estimates at smaller area levels using design-based methods, or even area-level models, such as the Fay-Herriot model, are often not reliable. However, estimates at the admin2 level are desired because this is often the level at which health interventions are administered. As a result of this sampling design, it is often helpful to use unit-level models to obtain small area estimates (Rao and Molina, 2015), where the units are the sampled clusters. Because the households in each cluster are recorded with the same location and sampling weight, it is natural to combine data within the same cluster and model these summed counts.

The methods of this chapter are motivated by estimation of the NMR in cases where design-based estimates are reliable at the first, but not second, administrative level. The model can easily be simplified to accommodate a case where the design-based estimates

are reliable at the national level, but not the admin1 level. We will use NMR-specific terminology throughout (i.e., births and neonatal deaths), but our proposed model can be applied to prevalence estimation of any rare event using complex survey data. For example, if HIV status is the indicator of interest, 'births' can be replaced by 'individuals' and 'neonatal deaths' replaced by 'individuals with positive HIV status'.

2.3 Bayesian unit-level models for rare event prevalence

2.3.1 Sampling Model

Suppose we seek to estimate the prevalence of a rare event, neonatal mortality, at the admin2 level using sparse survey data. It is common to fit such a model under the assumption that the number of events follows an overdispersed Poisson distribution (Diggle and Giorgi, 2019), where the overdispersion accounts for the within-cluster dependence in outcomes. Specifically, let n_c and Z_c represent the number of births and neonatal deaths in a fixed time period in the sampled households in sampled cluster c, respectively, where each cluster is contained in one of m_2 admin2 areas. Then, for neonatal mortality rate in cluster c (i.e., probability of death in the neonatal population in cluster c), r_c, and overdispersion parameter λ, we assume that for each cluster, c,

$$Z_c \mid r_c, \lambda \sim \text{Negative-binomial}(n_c r_c, \lambda) \tag{1}$$

with the parameterization

$$P(Z_c \mid n_c r_c, \lambda) = \frac{\Gamma(Z_c + n_c r_c/\lambda)}{\Gamma(Z_c+1)\Gamma(n_c r_c/\lambda)} \lambda^{Z_c} \left(1+\lambda\right)^{-(Z_c + n_c r_c/\lambda)}. \tag{2}$$

Under this parameterization, $\mathbb{E}[Z_c \mid n_c, r_c, \lambda] = n_c r_c$ and $\text{Var}(Z_c \mid n_c, r_c, \lambda) = (1+\lambda) n_c r_c$. We link this distribution with the regression model,

$$\log(r_c) = \alpha + b_j \mathbb{1}_{c \in \delta_2(j)} \tag{3}$$

where b_j is an admin2-level random effect (independent and identically distributed or with spatial structure) and $\delta_k(j)$ is the set of clusters in k^{th} administrative area j. Throughout, we use $b_j \mathbb{1}_{c\in\delta_2(j)}$ to denote the only nonzero term of the summation, $\sum_j b_j \mathbb{1}_{c\in\delta_2(j)}$. Appendix A1 provides a summary of definitions and notation introduced throughout this chapter. When this model is fit with sparse data there is often a large amount of shrinkage because information in each admin2 area is very limited. One way to mitigate this shrinkage effect is to use a regression model with nested spatial effects which includes a set of fixed effects at the admin1 level and a set of random effects at the admin2 level, i.e., replacing (3) with

$$\log(r_c) = \alpha + \beta_i \mathbb{1}_{c\in\delta_1(i)} + b_j \mathbb{1}_{c\in\delta_2(j)} \tag{4}$$

where β_i is an admin1-level fixed effect and β_1 is fixed at 0 to preserve identifiability. As above, we use $\beta_i \mathbb{1}_{c\in\delta_1(i)}$ to denote the only nonzero term of the summation, $\sum_i \beta_i \mathbb{1}_{c\in\delta_1(i)}$.

2.3.2 Motivating example: NMR in Zambia

To demonstrate the difference between regression models (3) and (4), consider the example of NMR estimation in Zambia at the admin2-level using all births between 2009 and 2013 sampled in the 2013 Zambia DHS (ICF International, 2014). Zambia has 10 admin1 areas and 115 admin2 areas. A summary of these data is displayed in Figure 2.1. The numbers of observed births and neonatal deaths in each admin1 area are sufficiently large to use an area-level model, such as Fay-Herriot (Fay and Herriot, 1979), with an average of 1319 births and 32 deaths observed in each area. However, the number of observed births and neonatal deaths in each admin2 area are prohibitively small, with an average of 115 births and 2.8 deaths observed in each area, and 21.7% of areas observing no neonatal deaths at all. At this level of data sparsity, a unit level model is useful for estimation at the admin2 level (Wakefield et al., 2025).

We must account for the stratified sampling design of the DHS survey (admin1 area

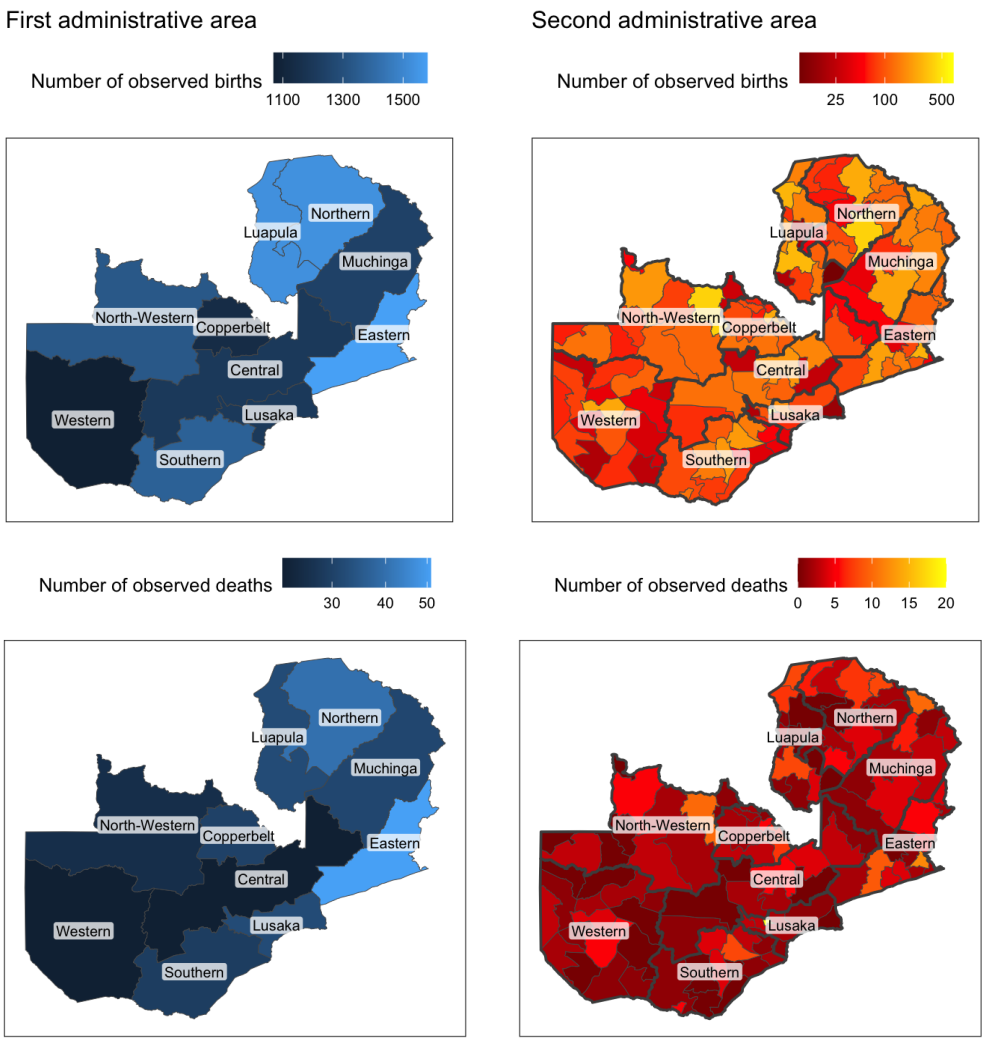


Figure 2.1. Summary of births and neonatal deaths between 2009 and 2013 recorded in the 2013 Zambia DHS. The top row provides maps of the total number of observed births in each admin1 and admin2 area. The bottom row provides maps of the total number of neonatal deaths in each admin1 and admin2 area. Each map has a different color scale to make differences between areas more clear.

crossed with urban/rural), so instead of estimating a global intercept, we include one urban and one rural intercept as follows,

$$\log(r_c) = \alpha_U \mathbb{1}_{c\in\delta_U} + \alpha_R \mathbb{1}_{c\in\delta_R} + b_j \mathbb{1}_{c\in\delta_2(j)} \tag{5}$$

$$\log(r_c) = \alpha_U \mathbb{1}_{c\in\delta_U} + \alpha_R \mathbb{1}_{c\in\delta_R} + \beta_i \mathbb{1}_{c\in\delta_1(i)} + b_j \mathbb{1}_{c\in\delta_2(j)}, \tag{6}$$

where δ_U and δ_R are the sets of urban and rural clusters, respectively, and β_1 is fixed at 0 to preserve identifiability.

We estimate the NMR at the admin2 level under regression models (5) and (6) using the `Stan` software (Stan Development Team, 2024). While a variety of random effect models at the admin2 level could be employed, we use a BYM2 spatial effect. Introduced by Riebler et al. (2016), the BYM2 model is a re-parameterized version of the Besag-York-Mollié (BYM) model (Besag et al., 1991), which includes both unstructured, independent and identically distributed (IID), spatial effects and structured, intrinsically conditional autoregressive (ICAR), spatial effects, (Besag, 1974). The BYM2 model has two parameters: σ^2, which indicates the total variance of the random effects, and ϕ, which indicates the proportion of this variation that is explained by the structured component. The structured component has a sum-to-zero constraint to ensure identifiability. More specifically, in (6), the structured component of the spatial effect has a separate sum-to-zero constraint for the areas within each first adimistrative area, while in (5) there is only one global sum-to-zero constraint. In Appendix A2 we compare NMR estimates and uncertainty under model (6) using IID and BYM2 spatial effects, which shows that, in this example, choice of spatial effect model makes little difference, so we choose BYM2 as it provides the option of more structure (by including an ICAR component in addition to an IID component) and has the added benefit of allowing more spatially informative estimation in areas without observed data.

For hyperpriors, we set $\phi \sim \text{Beta}(0.5, 0.5)$ and use a penalized complexity (PC) prior for σ^2 with hyperparameters $U = 1$ and $\alpha = 0.01$, which corresponds to the prior belief that $P(\sigma > 1) = 0.01$ (Simpson et al., 2017). We place diffuse priors on the overdispersion and regression parameters: $\lambda \sim \text{Exp}(1)$, $(\alpha_U, \alpha_R) \sim^{\text{iid}} \mathcal{N}(-3.5, 3^2)$ and $\beta_i \sim \mathcal{N}(0, 1)$. The priors on the urban and rural intercepts are centered at -3.5 to reflect the prior belief that neonatal death is a rare event ($e^{-3.5} \approx 0.03$). Neonatal mortality rate estimates in many Sub-Saharan African countries, including Zambia, which are available on the UN-IGME Child Mortality

Estimation dashboard (`childmortality.org`), confirm that this is a reasonable prior mean for the overall level. The priors on the regression parameters $(\alpha_U, \alpha_R, \boldsymbol{\beta})$ are chosen with the goal of being relatively diffuse on the exponential scale, while still accounting for the prior belief that the outcome measure is a rare event. More information regarding these prior choices is presented in Appendix A3. We aggregate the urban and rural estimates within each admin2 area using urban/rural population fractions estimated using the method in Wu and Wakefield (2024). This method is a logistic classification model which uses pixel-level population density and defines a threshold for classifying pixels as urban or rural. This threshold is calibrated by urban/rural composition at the admin1 level which is reported by DHS (Corsi et al., 2012). The pixel-level population density is obtained from WorldPop (Tatem, 2017), a large-scale producer of high-resolution age-specific population distributions for a wide array of countries.

In Figure 2.2 we map the NMR (deaths per 1000 live births) estimates under non-nested, nested, and direct methods. We observe that including a set of fixed effects at the admin1 level in the nested model significantly reduces shrinkage, by inducing shrinkage at a more local level. The nested model estimates, mapped on the top right, have a range of $18 - 33$ deaths per 1000 live births, while the estimates from the non-nested model, mapped on the top left, are nearly homogeneous with a range of $22 - 26$ deaths per 1000 live births. In the bottom right panel of Figure 2.2, we also include a map of the admin2-level direct estimates (for areas where at least 1 neonatal death is observed) to illustrate the magnitude of shrinkage, but we caution that they are estimated from very small sample sizes and have large uncertainty (of the 87 out of 115 areas which have valid estimates, 70% of areas have coefficient of variation greater than 50% and 23% have coefficient of variation greater than 100%). In Appendix A4 we provide maps of the coefficients of variation for the estimates in Figure 2.2. To further compare the model-based and direct estimates, we aggregate the admin2-level model-based estimates to the admin1 level with population weights calculated using WorldPop data (Tatem, 2017). In Figure 2.3, observe that the aggregation of the

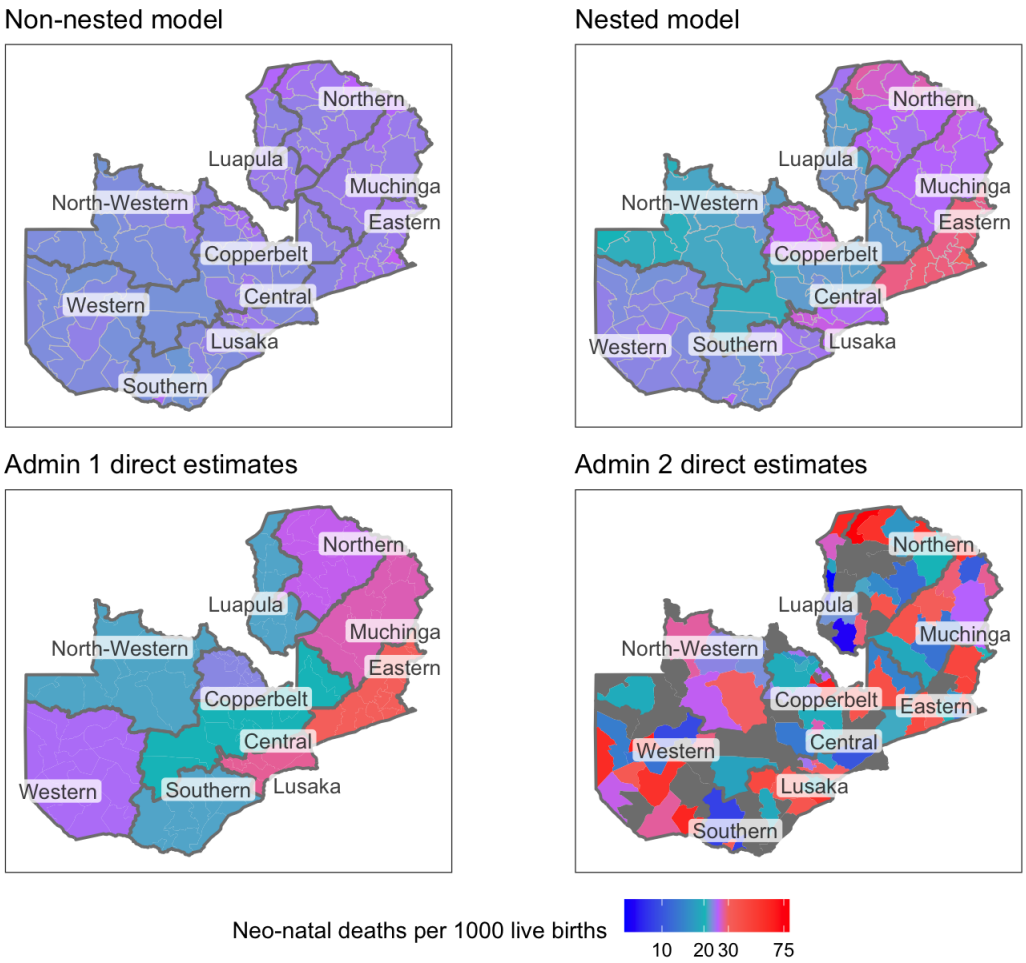


Figure 2.2. Map of NMR estimates in Zambia (2009-2013) under various models. The top row displays results from unit-level Bayesian negative binomial models with BYM2 spatial effects at the admin2 level, while the bottom row displays Hájek direct estimates at the admin1 and admin2 level, respectively. The map on the top left displays the estimates obtained from the model with spatial effects only at the admin2 level (5), while the map on the top right displays the estimates obtained from the model with nested spatial effects (6).

estimates to the admin1 level can be quite far from the consistent design-based estimate. This figure further highlights the overwhelming shrinkage of estimates under the non-nested model. In some admin1 areas, such as the Eastern, Northern, Luapala, and Central regions, aggregated estimates from the nested model are in close agreement with their corresponding design-based estimates. However, for other areas, such as the Lusaka, Muchinga, Western, and Copperbelt regions, aggregated estimates from the nested model are quite different from their corresponding design-based estimates. While employing nested spatial effects can

mitigate shrinkage in unit-level models, it is not sufficient to achieve small area estimates which agree in aggregation with higher-level design-based estimates. In fact, there is no guarantee that aggregated model-based estimates will be at all similar to their design-based counterparts. Because design-based estimators are consistent, it is desirable to consider a model which will encourage consistency with them in aggregation.

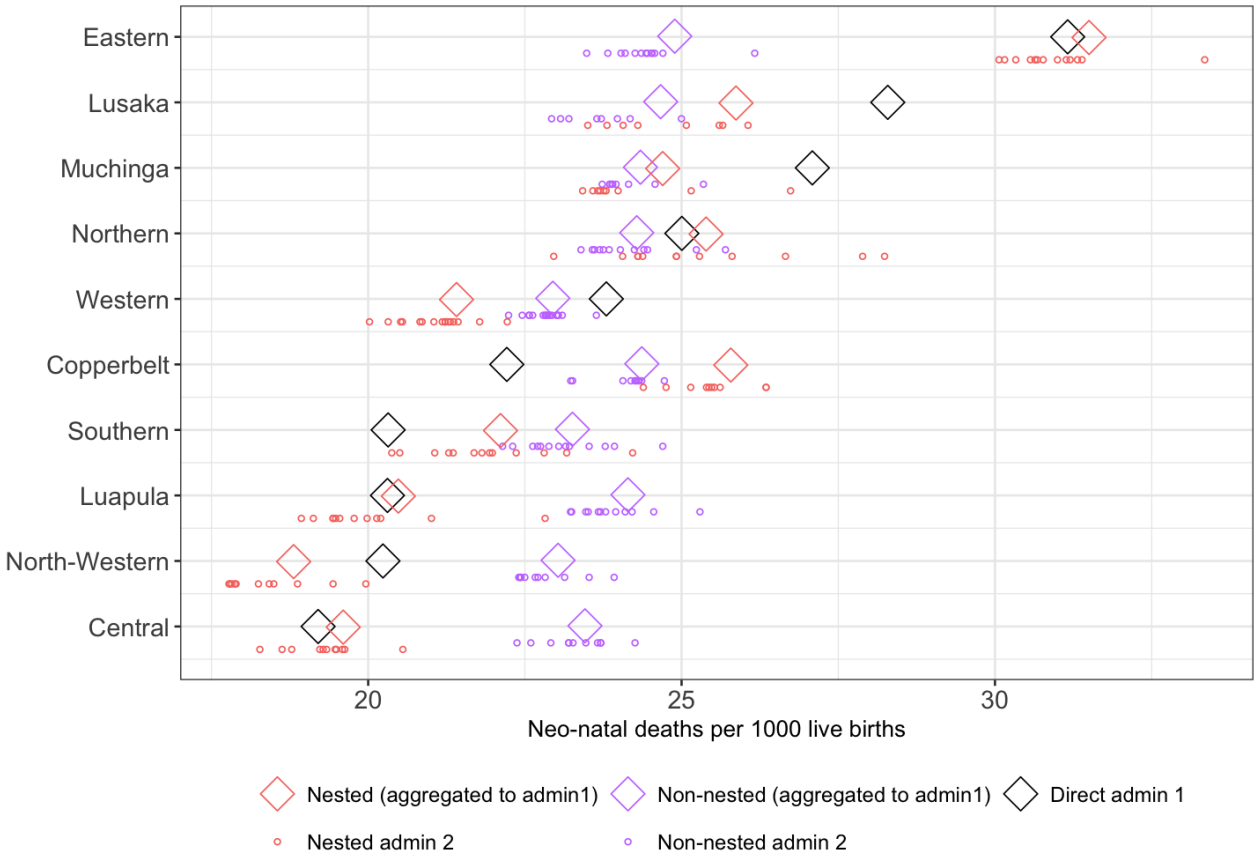


Figure 2.3. NMR estimates at the admin2 level in Zambia (2009-2013) under various models. The colored circles denote the admin2-level estimates using unit-level Bayesian negative binomial models with BYM2 spatial effects at the admin2 level and the colored diamonds denote their aggregations to the admin1 level. The black diamonds denote Hájek direct estimates at the admin1 level.

2.4 Direct-assisted Bayesian unit-level model

In the following section we introduce the DABUL models: an extension of the standard unit-level Bayesian model with nested random effects which incorporates direct design-based

estimates to encourage consistency with design-based estimates in aggregation.

First, consider a model for a full enumeration of the population in which all births in all clusters are observed. Let $\mathbf{Y}$ be the vector containing the total number of neonatal deaths in all households, Y_c, in each cluster c, so that the vector has length equal to the total number of clusters in the population. Let $\mathbf{Y_i}$ be the sub-vector containing the elements of $\mathbf{Y}$ corresponding to all clusters in the admin1 area i, and Y_{i+} be the sum of the elements in this sub-vector. Similarly, let $\mathbf{N}$ be the vector containing the total number of births in all households, N_c, in each cluster c, so that the sum of the elements of $\mathbf{N}$ is the total number of births in the population. Let $\mathbf{N_i}$ be the sub-vector containing the elements of $\mathbf{N}$ corresponding to all clusters in the admin1 area i, and N_{i+} be the sum of the elements in this sub-vector. Also, let $\mathbf{r}$ be the vector containing elements r_c, and $\mathbf{r_i}$ be the sub-vector containing the elements of $\mathbf{r}$ corresponding to clusters in the admin1 area, i. Let $\odot$ denote the Hadamard product (elementwise multiplication; see Appendix A1 for more details). Then an equivalent way to define a negative binomial distribution is

$$P(\mathbf{Y_i} \mid \mathbf{r_i}, \lambda) = P(\mathbf{Y_i}, Y_{i+} \mid \mathbf{r_i}, \lambda) = P(\mathbf{Y_i} \mid Y_{i+}, \mathbf{r_i}, \lambda) \times P(Y_{i+} \mid \mathbf{r_i}, \lambda) \tag{7}$$

for clusters in admin1 area, i, where

$$\mathbf{Y_i} \mid Y_{i+}, \mathbf{r_i}, \lambda \sim \text{DCM}\left(Y_{i+}, \mathbf{N_i} \odot \mathbf{r_i}/\lambda\right) \tag{8}$$

and

$$Y_{i+} \mid \mathbf{r_i}, \lambda \sim \text{Negative-binomial}\left(\sum_{c \in \delta_1(i)} N_c r_c, \lambda\right). \tag{9}$$

Here DCM abbreviates the Dirichlet compound multinomial distribution (Mosimann, 1962), also referred to as the multivariate Pólya distribution , which is defined by the probability

mass function,

$$P(\mathbf{Y_i} \mid Y_{i+}, \mathbf{r_i}, \lambda) = \frac{\Gamma(Y_{i+}+1)\Gamma(\frac{1}{\lambda}\sum_{c\in\delta_1(i)} N_c r_c)}{\Gamma(Y_{i+}+\frac{1}{\lambda}\sum_{c\in\delta_1(i)} N_c r_c)} \prod_{c\in\delta_1(i)} \frac{\Gamma(Y_c + N_c r_c/\lambda)}{\Gamma(N_c r_c/\lambda)\Gamma(Y_c+1)}.$$

This alternative expression follows directly from the additive property of negative binomial random variables and the relationship between the negative binomial and DCM distributions. Specifically, if $\mathbf{x} = (x_1, ..., x_n)$ are independent random variables, each following a negative binomial distribution with mean parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, respectively, and a common overdispersion parameter λ, then $x_+ := \sum_{i=1}^n x_i$ follows a negative binomial distribution with mean $\sum_{i=1}^n \theta_i$, and overdispersion parameter λ. It also holds that $\mathbf{x} \mid x_+, \boldsymbol{\theta}, \lambda$ follows a DCM distribution with parameter vector $\boldsymbol{\theta}/\lambda$.

When all births and neonatal deaths in each cluster are observed (i.e. $\mathbf{Y}$ and Y_{i+} is known), this alternative parameterization of the negative binomial distribution is equivalent to the parameterization presented in (2). We will now consider a survey sampling context in which 1) we use two-stage sampling, so that only some clusters are sampled and only some of the births and corresponding neonatal deaths are observed in each of those sampled clusters, i.e., $\mathbf{Y}$ is not known, and 2) each Y_{i+} can be estimated through a design-based estimate.

2.4.1 Accounting for survey sampling

We first assume Y_{i+} is known for all i, so that only point 1) needs to be considered. Suppose the birth and death observations are collected using the two-stage cluster sampling design described in section 2. In this situation, we must incorporate the fact that not all births, and corresponding deaths, are observed in each sampled cluster. For each cluster c, let γ_c be an indicator variable denoting whether the cluster was sampled and let n_c and Z_c be the number of births and neonatal deaths across all of the sampled households in that cluster. Also, let the vectors $\boldsymbol{\gamma}$, $\boldsymbol{\gamma_i}$, $\mathbf{Z}$ and $\mathbf{Z_i}$ be defined analogously to $\mathbf{Y}$ and $\mathbf{Y_i}$. Note that if $\gamma_c = 0$ for a cluster c, this implies $n_c = Z_c = 0$ because that cluster was not sampled. Observe then,

for each cluster c,

$$Z_c \mid Y_c, \gamma_c = 1 \sim \text{Hypergeometric}\left(N_c, Y_c, n_c\right), \qquad P(Z_c \mid \gamma_c = 0) = I(Z_c = 0). \tag{10}$$

2.4.2 Incorporating higher level design-based estimates

As alluded to in the previous section, in this survey sampling context each Y_{i+} is not known, and must be estimated. We have established that the DABUL models pertain to the situation in which there is sufficient data to use design-based estimates at the admin1 level, though not at the admin2 level. We will denote the design-based prevalence estimate for an admin1 area i as $r_i^{D_1}$. The only difference between the two DABUL models is how the design-based estimates are incorporated into the likelihood. For the hard DABUL model, we can simply set $Y_{i+} = r_i^{D_1} \times N_{i+}$, to apply a hard benchmarking constraint. For the soft DABUL model, we account for the uncertainty of the design-based estimate by using the property that design-based estimators are asymptotically normal (Breidt and Opsomer, 2017, Section 3). On the logit scale, we can specify

$$\text{logit}\left(r_i^{D_1}\right) \mid Y_{i+} \sim \mathcal{N}\left(\text{logit}\left(Y_{i+}/N_{i+}\right), V_i^{D_1}\right) \tag{11}$$

where $V_i^{D_1}$ is the design-based variance of $\text{logit}\left(r_i^{D_1}\right)$, which accounts for the uncertainty induced by the sampling design. This variance can be estimated using the `svydesign` function in the `survey` package (Lumley, 2024), to estimate the prevalence, which is then transformed to the logit scale using the delta method. If this method does not produce stable variance estimates, a variance smoothing method may be employed (e.g., using generalized variance functions described in chapter 7 of Wolter (2007)), but we will not consider this option here.

Then, the joint distribution of $(\mathbf{Z}, \mathbf{r^{D_1}})$ conditional on $(\mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma})$ under the hard DABUL

model can be expressed as,

$$\begin{aligned} P\left(\mathbf{Z}, \mathbf{r^{D_1}} \mid \mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma}\right) &= P\left(\mathbf{Z} \mid \mathbf{r^{D_1}}, \mathbf{Y}, \boldsymbol{\gamma}\right) P\left(\mathbf{r^{D_1}} \mid \mathbf{Y}_+\right) \\ &\propto \prod_{i=1}^{m_1} P\left(\mathbf{Z_i} \mid r_i^{D_1}, \mathbf{Y_i}, \boldsymbol{\gamma}_i\right) I\left(r_i^{D_1} = Y_{i+}/N_{i+}\right). \end{aligned} \tag{12}$$

Similarly, the joint distribution of $(\mathbf{Z}, \mathbf{r^{D_1}})$ conditional on $(\mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma})$ under the soft DABUL model can be expressed as,

$$\begin{aligned} P(\mathbf{Z}, \mathbf{r^{D_1}} \mid \mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma}) &= P(\mathbf{Z} \mid \mathbf{r^{D_1}}, \mathbf{Y}, \boldsymbol{\gamma}) P(\mathbf{r^{D_1}} \mid \mathbf{Y}_+) \\ &= \prod_{i=1}^{m_1} P(\mathbf{Z_i} \mid r_i^{D_1}, \mathbf{Y_i}, \boldsymbol{\gamma}_i) P(r_i^{D_1} \mid Y_{i+}). \end{aligned} \tag{13}$$

By definition, $r_i^{D_1}$ is a weighted sum of $\mathbf{Z_i}$, so it follows that $P(\mathbf{Z_i} \mid r_i^{D_1}, \mathbf{Y_i}, \boldsymbol{\gamma}_i)$ is a product of Hypergeometric distributions conditional on the survey-weighted sum of $\mathbf{Z_i}$, $g_i^{D_1}(\mathbf{Z_i})$, being equal to $r_i^{D_1}$. More explicitly, for admin1 area i,

$$\begin{aligned} P(\mathbf{Z_i} \mid r_i^{D_1}, \mathbf{Y_i}, \boldsymbol{\gamma}_i) &= \frac{P(\mathbf{Z_i}, g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i)}{P(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i)} \\ &\propto \frac{I(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1})}{P(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i)} \underbrace{\left[\prod_{\substack{c \in \delta_1(i): \\ \gamma_c = 1}} P(Z_c \mid Y_c, \gamma_c = 1)\right]}_{\text{Hypergeometric (10)}}. \end{aligned} \tag{14}$$

Note that $g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i$ follows the distribution of a weighted sum of Hypergeometric random variables.

2.4.3 Combining likelihood components

The joint distribution of $(\mathbf{Z}, \mathbf{r}^{\mathbf{D_1}}, \mathbf{Y}_+, \mathbf{Y})$ conditional on $(\boldsymbol{\gamma}, \mathbf{r}, \lambda)$ can be expressed as,

$$\begin{aligned} P(\mathbf{Z}, \mathbf{r}^{\mathbf{D_1}}, \mathbf{Y}_+, \mathbf{Y} \mid \boldsymbol{\gamma}, \mathbf{r}, \lambda) &\propto \prod_{i=1}^{m_1} P(\mathbf{Z}, \mathbf{r}^{\mathbf{D_1}} \mid \mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma}) P(Y_{i+}, \mathbf{Y_i} \mid \mathbf{r_i}, \lambda) \\ &\propto \prod_{i=1}^{m_1} \underbrace{P(\mathbf{Z}, \mathbf{r}^{\mathbf{D_1}} \mid \mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma})}_{\text{(12) or (13)}} \underbrace{P(\mathbf{Y_i} \mid Y_{i+}, \mathbf{r_i}, \lambda)}_{\text{DCM (8)}} \underbrace{P(Y_{i+} \mid \mathbf{r_i}, \lambda)}_{\text{Neg-bin (9)}} \end{aligned} \tag{15}$$

where $\mathbf{Y^{(s)}}$ is the sub-vector of $\mathbf{Y}$ containing the elements corresponding to sampled clusters and m_1 is the number of admin1 areas.

Observe that in this joint distribution, the latent variables corresponding to the unobserved clusters (i.e., $\mathbf{Y} \setminus \mathbf{Y^{(s)}}$) are only included in the DCM distribution. Estimation of these latent variables will cause an identifiability issue because there is no data for these clusters, so the total neonatal death counts from all of the unobserved clusters in an admin1 area can be collapsed into one group. Because we can estimate $\mathbf{Y}^{(s)}$ and $\mathbf{Y}_+$, this collapsing ensures identifiability. Hence, we replace (8) with

$$(\mathbf{Y_i^{(s)}}, Y_{i+} - Y_{i+}^{(s)}) \mid Y_{i+}, \mathbf{r_i}, \lambda \sim \text{DCM}\left(Y_{i+}, \left((\mathbf{N_i^{(s)}} \odot \mathbf{r_i^{(s)}})/\lambda, \left(\sum_{\substack{c \in \delta_1(i) \\ \gamma_c = 0}} N_c r_c\right)/\lambda\right)\right) \tag{16}$$

where $Y_{i+}^{(s)}$ is the sum of the elements in $\mathbf{Y_i^{(s)}}$, and $\mathbf{N}_i^{(s)}$ and $\mathbf{r}_i^{(s)}$ are the sub-vectors of $\mathbf{N_i}$ and $\mathbf{r_i}$ containing elements corresponding to the observed clusters in admin1 area i. Note that this distribution has dimension equal to the total number of observed clusters plus one, as does the corresponding probability vector. By making this adjustment we do not lose any information and reduce the dimension of the latent variable vector from the total number of clusters, to the total number of *observed* clusters, which is of significantly smaller magnitude. This reduction increases numerical stability and computational efficiency.

This completes the derivation of the posterior distribution under the DABUL models,

which can be expressed as follows,

$$\pi(\mathbf{r}, \lambda, \boldsymbol{\kappa}, \mathbf{Y}^{(\mathbf{s})}, \mathbf{Y}_+ \mid \mathbf{Z}, \mathbf{r}^{\mathbf{D_1}}, \boldsymbol{\gamma}) \propto \underbrace{P(\mathbf{Z}, \mathbf{r}^{\mathbf{D_1}}, \mathbf{Y}_+, \mathbf{Y} \mid \boldsymbol{\gamma}, \mathbf{r}, \lambda)}_{(15)} P(\mathbf{r} \mid \boldsymbol{\kappa}) \pi(\lambda, \boldsymbol{\kappa}) \tag{17}$$

where $\boldsymbol{\kappa}$ is the set of regression parameters and hyperparameters. Figure 2.4 depicts the components of this model in a visual form and compares it to the standard unit-level Bayesian model with nested spatial effects. Observe the only difference between the hard and soft DABUL models is how $\mathbf{r}^{\mathbf{D_1}}$ is used to estimate $\mathbf{Y}_+$.

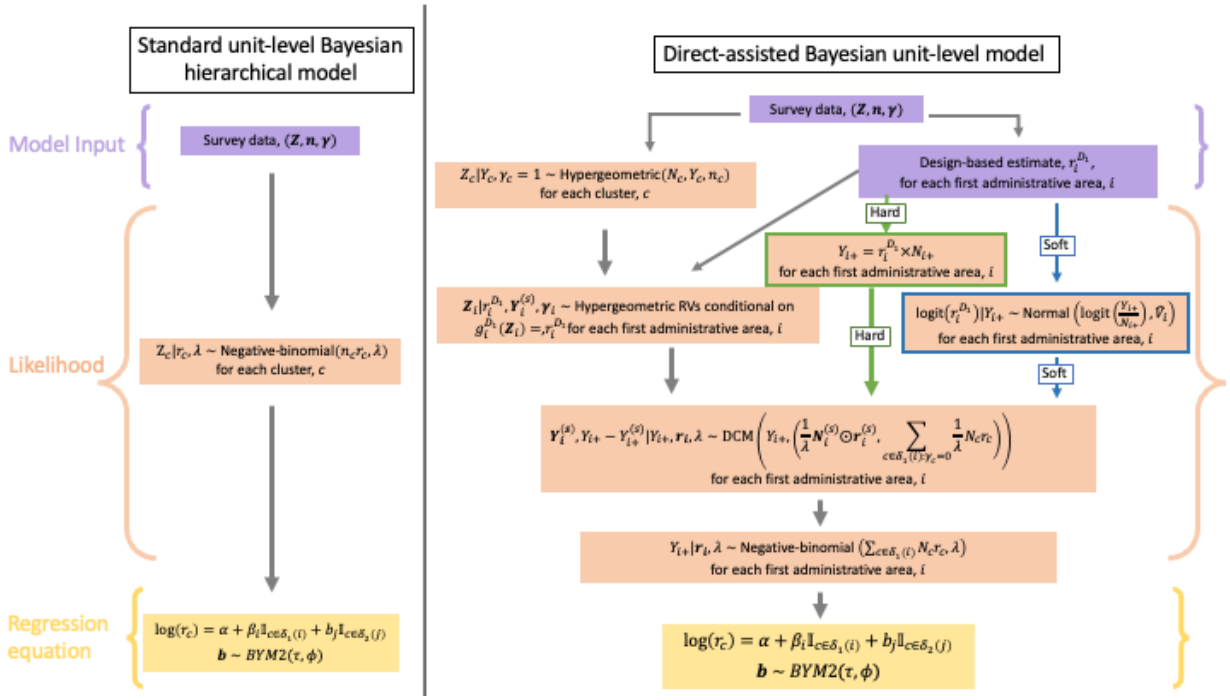


Figure 2.4. Comparison of DABUL models with the standard unit-level Bayesian model. The cluster-length vector $\mathbf{Y}$, indexed by c, denotes the total number of neonatal deaths in each cluster. The vector $\mathbf{Y}^{(\mathbf{s})}$ is the sub-vector of $\mathbf{Y}$ containing the elements corresponding to sampled clusters and $\mathbf{Z}$ is the vector of observed deaths in each sampled cluster, indexed by c.

2.4.4 Implementation

Estimation of the posterior distribution (17) is not straightforward because most Bayesian inference software, including `Stan` (Stan Development Team, 2024) and the `INLA` package

in `R` (Rue et al., 2009), are not built to accommodate discrete latent variables that cannot be marginalized out. As a result, we implement a NUTS-within-Gibbs sampler, described in Algorithms 1 and 2, which uses a No-U-Turn Sampler (NUTS) to update all continuous parameters and inverse transform sampling to update the discrete parameters. Introduced by Hoffman et al. (2014), NUTS is an efficient extension of Hamiltonian Monte Carlo and is the computational framework for `Stan`. We implemented the NUTS step of our algorithm in R according to Algorithm 3 in Hoffman et al. (2014), with assistance from the code provided by Märtens (2017). We modified the algorithm to include a scaling matrix, denoted $\mathbf{\Sigma}$, in the proposal density to better account for correlation between parameters, as is done in `Stan`. The gradient of the log-likelihood with respect to all continuous parameters was derived analytically to avoid costly numerical approximations at each step of the NUTS sampler. The tuning parameter, ϵ, is chosen to optimize the step size of the random walk.

Algorithm 1 NUTS-within-Gibbs sampler for soft DABUL model

Input: $(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{n}, \mathbf{N})$; $(\mathbf{r^{D_1}}, \mathbf{V^{D_1}})$; ϵ, $\mathbf{\Sigma}$, M, $\ell_{REG} := \log \pi(\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda \mid \mathbf{Y^{(s)}}, \mathbf{Y}_+, \mathbf{Z}, \boldsymbol{\gamma})$

for m in 1:M **do**

$(\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda)^{(m)} \longleftarrow NUTSOneStep((\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda)^{(m-1)}, \ell_{REG}, (\mathbf{Y^{(s)}}, \mathbf{Y}_+)^{(m-1)}, \epsilon, \mathbf{\Sigma})$

for each cluster, c **do**

$r_c^{(m)} \longleftarrow exp(\alpha + \beta_i \mathbb{1}_{c \in \delta_1(i)} + b_j \mathbb{1}_{c \in \delta_2(j)})$

end for

for each admin1 area, i **do**

Draw $Y_{i+}^{(m)} \sim P\left(Y_{i+} \mid \mathbf{Y^{(s)}}^{(m-1)}, \boldsymbol{\gamma}, \mathbf{r}^{(m)}, \lambda^{(m)}, r_i^{D_1}, V_i^{D_1}\right)$ using inverse transform sampling

end for

for each observed cluster, c **do**

Draw $Y_c^{(m)} \sim P\left(Y_c \mid \mathbf{Y^{(s)}_{1:(c-1)}}^{(m)}, \mathbf{Y^{(s)}_{(c+1):n_{clusters}}}^{(m-1)}, Y_{i+}^{(m)}, Z_c, \gamma_c, r_c^{(m)}, \lambda^{(m)}\right)$ using inverse transform sampling

end for

end for

Output: $\alpha, \boldsymbol{\beta}, \mathbf{b}, \lambda, \sigma^2, \phi, \mathbf{Y^{(s)}}, \mathbf{Y}_+$

All components of the posterior distribution (17) have a closed form except for $P\left(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i\right)$, the probability mass function of a weighted sum of independent Hypergeometric random variables. Although $g_i^{D_1}(\mathbf{Z_i}) \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i$ follows a discrete distribution,

Algorithm 2 NUTS-within-Gibbs Sampler for hard DABUL model

Input: $(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{n}, \mathbf{N})$; $\mathbf{r^{D_1}}$; ϵ, $\boldsymbol{\Sigma}$, M, $\ell_{REG} := \log \pi(\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda \mid \mathbf{Y^{(s)}}, \mathbf{Y}_+, \mathbf{Z}, \boldsymbol{\gamma})$

$Y_{i+} = r_i^{D_1} N_{i+}$ for each admin1 area i

for m in 1:M **do**

$(\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda)^{(m)} \longleftarrow NUTSOneStep\left((\alpha, \boldsymbol{\beta}, \mathbf{b}, \sigma^2, \phi, \lambda)^{(m-1)}, \ell_{REG}, \mathbf{Y^{(s)}}^{(m-1)}, \mathbf{Y}_+, \epsilon, \boldsymbol{\Sigma}\right)$

for each sampled cluster, c **do**

$r_c^{(m)} \longleftarrow exp(\alpha + \beta_i \mathbb{1}_{c \in \delta_1(i)} + b_j \mathbb{1}_{c \in \delta_2(j)})$

Draw $Y_c^{(m)} \sim P\left(Y_c \mid \mathbf{Y^{(s)}_{1:(c-1)}}^{(m)}, \mathbf{Y^{(s)}_{(c+1):n_{clusters}}}^{(m-1)}, Y_{i+}, Z_c, \gamma_c, r_c^{(m)}, \lambda^{(m)}\right)$ using inverse transform sampling

end for

end for

Output: $\alpha, \boldsymbol{\beta}, \mathbf{b}, \lambda, \sigma^2, \phi, \mathbf{Y^{(s)}}$

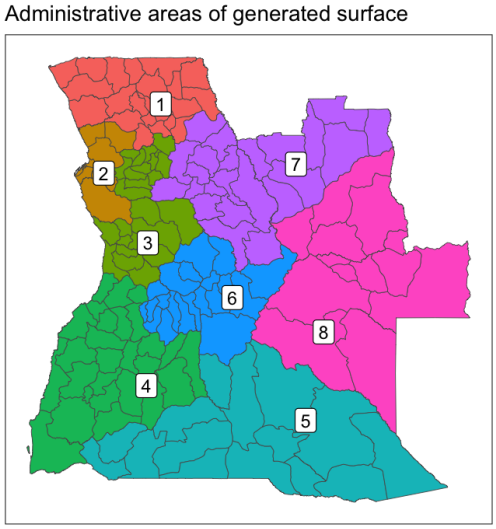
the domain contains all the possible weighted sums of $\mathbf{Z}_i$, which is very dense because each Z_c has a different corresponding weight. Therefore, in practice, we can treat $P(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i)$ as a probability density function (with continuous domain). At each iteration of the algorithm this component must be approximated for the inverse transform sampling of $\mathbf{Y}^{(s)}$ step. We approximate this term by drawing 2000 multivariate samples from the independent Hypergeometric distributions implied by the proposed and current states and taking the weighted sum of each sample vector, resulting in 2000 draws from the distribution of $g_i^{D_1}(\mathbf{Z_i}) \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i$. Gaussian kernel density estimation is then applied to this empirical distribution, using the `density` function in `R`, to approximate $P(g_i^{D_1}(\mathbf{Z_i}) = r_i^{D_1} \mid \mathbf{Y_i}, \boldsymbol{\gamma}_i)$. The bandwidth is chosen using Silverman's rule-of-thumb method (Silverman, 2018).

2.5 Simulation Study

In this study we will compare the performance of the proposed DABUL models to an analogous, standard unit-level Bayesian model with nested spatial effects. We start by defining a neighborhood structure with 8 admin1 areas and 159 admin2 areas. There is a minimum number of 13 admin2 areas in each admin1 area and a maximum of 27. A map of these areas is provided in Figure 2.5. This neighborhood structure and map is modified from that of Angola.

The total number of urban and rural clusters in each admin1 area is drawn from $\mathcal{N}(700, 100^2)$ and $\mathcal{N}(800, 100^2)$ distributions, respectively, and the total number of births in each urban cluster and each rural cluster is drawn from $\mathcal{N}(100, 10^2)$ and $\mathcal{N}(125, 10^2)$ distributions, respectively. All draws are rounded to the nearest natural number. These values are calibrated to be comparable to the sampling frames in LMICs used by DHS and MICS. Within each admin1 area, the clusters are distributed evenly across admin2 areas so that the number of births and clusters in an admin2 area are similar to other admin2 areas within its admin1 area.

Figure 2.5. Admin2 area neighborhood structure for simulation study.



The total number of neonatal deaths in each cluster, Y_c, is drawn from a negative binomial distribution with mean $N_c r_c$ and overdispersion parameter $\lambda = 0.25$, where r_c is defined as

$$\log(r_c) = \alpha_U \mathbb{1}_{c \in \delta_U} + \alpha_R \mathbb{1}_{c \in \delta_R} + \beta_i \mathbb{1}_{c \in \delta_1(i)} + b_j \mathbb{1}_{c \in \delta_2(j)} \tag{18}$$

where $\alpha_U = \log(0.02), \alpha_R = \log(0.025)$, and $\boldsymbol{\beta} = (-0.2, -0.1, -0.05, -0.025, 0.025, 0.05, 0.1, 0.5)$. Separate urban and rural intercepts are specified to replicate real-world conditions in which prevalence are often different between urban and rural areas. This corresponds to a neonatal mortality rate between 16 and 41 per 1000 live births, which Figure 2.2 shows is a reasonable range. Additionally neonatal mortality rate estimates in many Sub-Saharan African countries, including Zambia, which are available on the UN-IGME Child Mortality Estimation dashboard (`childmortality.org`), confirm that this is a reasonable choice. The choice of overdispersion parameter $d = 0.25$ is also reasonable compared to the overdispersion param-

eter estimate in the motivating Zambia example, 0.19 (90% credible interval: $[0.08, 0.32]$). The random spatial effects $\mathbf{b}$ are drawn from a BYM2 model with parameters σ^2 and ϕ, which is equivalent to a mean-zero multivariate normal distribution with covariance matrix, $\sigma^2((1-\phi)\mathbf{I}+\phi\mathbf{Q}_*^-)$, where $\mathbf{Q}_*^-$ is the unique Moore-Penrose generalized inverse of the scaled structure matrix, Q_*, as described in Riebler et al. (2016). The values of σ^2 and ϕ vary over three different hyperparameter settings: in the first, $\sigma^2 = 0.15^2$ and $\phi = 0.25$; in the second, $\sigma^2 = 0.05^2$ and $\phi = 0.25$; and in the third $\sigma^2 = 0.05^2$ and $\phi = 0.7$. The choices of σ are reasonable compared to the σ parameter estimate in the motivating Zambia example, 0.12 (90% credible interval: $[0.03, 0.31]$). For the choice of ϕ, anywhere in the range of $(0, 1)$ is reasonable, so we pick two distinct values in that range. Through these settings, we can observe whether the relative performance of the DABUL model is affected by spatial precision or dependence.

For each simulation setting, 500 datasets are independently sampled from a single generated risk surface using two-stage stratified cluster sampling (the same method which is used by DHS and MICS, except that in the second stage we directly sample births instead of households). A summary of the four simulation settings is provided in Table 2.1. A proportion of urban and rural clusters are sampled from each admin1 area. For the majority of simulation settings, 8% and 5% of urban and rural clusters are sampled, respectively, but for one setting, only 5% and 3% are sampled so that the effect of sample size can be observed. Note that urban clusters are oversampled as this is often the case for DHS and MICS. Similarly, note that clusters are sampled at the admin1 level to mimic DHS and MICS, which are powered at the admin1 level. Due to this sampling design, there may be admin2 areas which have very small samples, or no sampled clusters at all.

For each sampled urban cluster, the number of births sampled is drawn from $\mathcal{N}(15, 4)$ and for each sampled rural cluster, the number of births sampled is drawn from $\mathcal{N}(20, 4)$. Each birth in a cluster has equal probability of being sampled. A summary of the resulting datasets (i.e., average numbers of observed births and deaths per administrative area) is

Table 2.1. Summary of simulation parameters and sampled datasets

	Simulation setting			
	1	2	3	1a
Spatial variance: σ^2	0.15^2	0.05^2	0.05^2	0.15^2
Proportion of variation explained by structured effect: ϕ	0.25	0.25	0.7	0.25
Percentage urban clusters sampled	8%	8%	8%	5%
Percentage rural clusters sampled	5%	5%	5%	3%
Avg. # of births per 1st admin area	1640	1640	1640	1007
Avg. # of neonatal deaths per 1st admin area	40	39	39	23
Avg. # of births per 2nd admin area	83	83	83	51
Avg. # of neonatal deaths per 2nd admin area	2.0	2.0	2.0	1.2
Pct. 2nd admin areas w/ no observed neonatal deaths	23.2	22.9	23.0	39.7

displayed in Table 2.1. For each of these 2000 sample datasets (4 settings $\times$ 500 samples), we obtain the following:

1. Hájek direct estimates and standard errors at the admin1 level using the `SUMMER` (Li et al., 2024) and `survey` (Lumley, 2024) packages in R.

2. Admin2-level prevalence estimates from a standard unit-level Bayesian model using `Stan` (Stan Development Team, 2024), by fitting the nested BYM2 regression model specified in (18) under the assumption that each Y_c follows a negative binomial distribution with mean $N_c r_c$ and overdispersion parameter λ.

3. Admin2-level prevalence estimates from the soft DABUL model with the regression equation specified in (18) using Algorithm 1 for 1000 iterations, after 1000 iterations of burn-in, for 4 chains.

4. Admin2-level prevalence estimates from the hard DABUL model with the regression equation specified in (18) using Algorithm 2 for 1000 iterations, after 1000 iterations of burn-in, for 4 chains.

As in the motivating example, we aggregate the urban and rural estimates within each admin2 area using urban/rural population fractions. The run time of each DABUL model

was approximately 130 minutes, while the run time of each standard unit-level model in `Stan` was approximately 2 minutes.

We compare the performance of the admin2-level prevalence estimates using four metrics. First, we quantify the discrepancy, or absolute difference, between the admin1 level aggregated estimates and the direct estimates, averaged across all simulations. For admin1 area i, the average discrepancy is calculated by

$$\frac{1}{500}\sum_{k=1}^{500}\left|\sum_{j\in\zeta(i)} w_j\hat{r}_j^{(k)} - r_i^{D_1(k)}\right|$$

where $\hat{r}_j^{(k)}$ is the admin2-level prevalence estimate for the k^{th} sample dataset, w_j is the population weight for admin2 area j, and $\zeta(i)$ is the set of admin2 areas in admin1 area i. Second, we evaluate the absolute error of the admin2-level estimates, averaged across all simulations, calculated by $\frac{1}{500}\sum_{k=1}^{500}\left|\hat{r}_j^{(k)} - r_j\right|$, where r_j is the population prevalence for admin2 area j. Then, we compare the coverage and average width of the admin2-level 90% credible intervals, calculated by

$$\frac{1}{500}\sum_{k=1}^{500} I\left(r_j \in \left[Q_j^{(k)}(5), Q_j^{(k)}(95)\right]\right)$$

and

$$\frac{1}{500}\sum_{k=1}^{500}\left(Q_j^{(k)}(95) - Q_j^{(k)}(5)\right),$$

respectively, where $Q_j^{(k)}(q)$ is the q^{th} quantile of the posterior distribution for admin2 area j and sample dataset k. Additionally, coefficients of variation of the admin2-level estimates, calculated by $\frac{1}{500}\sum_{k=1}^{500} 100\times\frac{SD_j^{(k)}}{\hat{r}_j^{(k)}}$, are presented in Appendix A5, where $SD_j^{(k)}$ is the standard deviation of the posterior distribution for admin2 area j and sampled dataset k.

Figure 2.6 displays the discrepancies between aggregated admin2-level model-based estimates and admin1-level direct estimates. The top panel depicts the average discrepancies across simulations for each admin1 area, model, and simulation setting. From this figure

we observe that discrepancies are uniformly lower among the aggregated soft DABUL estimates, compared to the aggregated standard unit-level model estimates. As expected, the discrepancies among the aggregated hard DABUL estimates are consistently near zero, as a result of enforcing a hard benchmarking constraint. Both of these observations are consistent across all four simulation settings. The bottom panel depicts the distribution of the average percent decrease in discrepancy of the soft DABUL estimates for each admin1 area, relative to the standard unit-level estimates, across all simulations, for each simulation setting. We observe that the aggregated soft DABUL estimates have $58.4 - 63.9\%$ lower discrepancy with the admin1-level direct estimates, on average, compared to aggregated estimates from the standard unit-level model. Percentage decrease in discrepancy is larger when the standard unit-level model has larger discrepancy with the direct estimates (scenario 1a). Figure A6 presented in Appendix A5 further shows this pattern.

Figure 2.7 displays the absolute error of the admin2-level estimates, i.e., the absolute value of the difference between the prevalence estimate and the true prevalence in that area. The top panel depicts the average absolute error across simulations for each admin2 area, model, and simulation setting. The bottom panel depicts the distribution of the average absolute error of each admin2 area across all simulations, for each model and simulation setting. This figure suggests that absolute error is comparable across all models and simulation settings, with a slightly lower average absolute error for standard unit-level model estimates.

Figures 2.8 and 2.9 compare the coverage and width of the 90% credible intervals for the admin2-level estimates. Note that coverage of area-specific Bayesian credible intervals in unit-level models (and that of frequentist confidence intervals, for that matter) are expected to attain the desired coverage *on average*, but the interval for each specific area does not necessarily have the target coverage (Yu and Hoff, 2018). Moreover, in an extensive simulation study, Osgood-Zimmerman and Wakefield (2023) found that coverage of area-specific credible intervals from unit-level models did not obtain the nominal level in a range of scenarios. From Figures 2.8 and 2.9, we observe that the DABUL estimates have wider credible

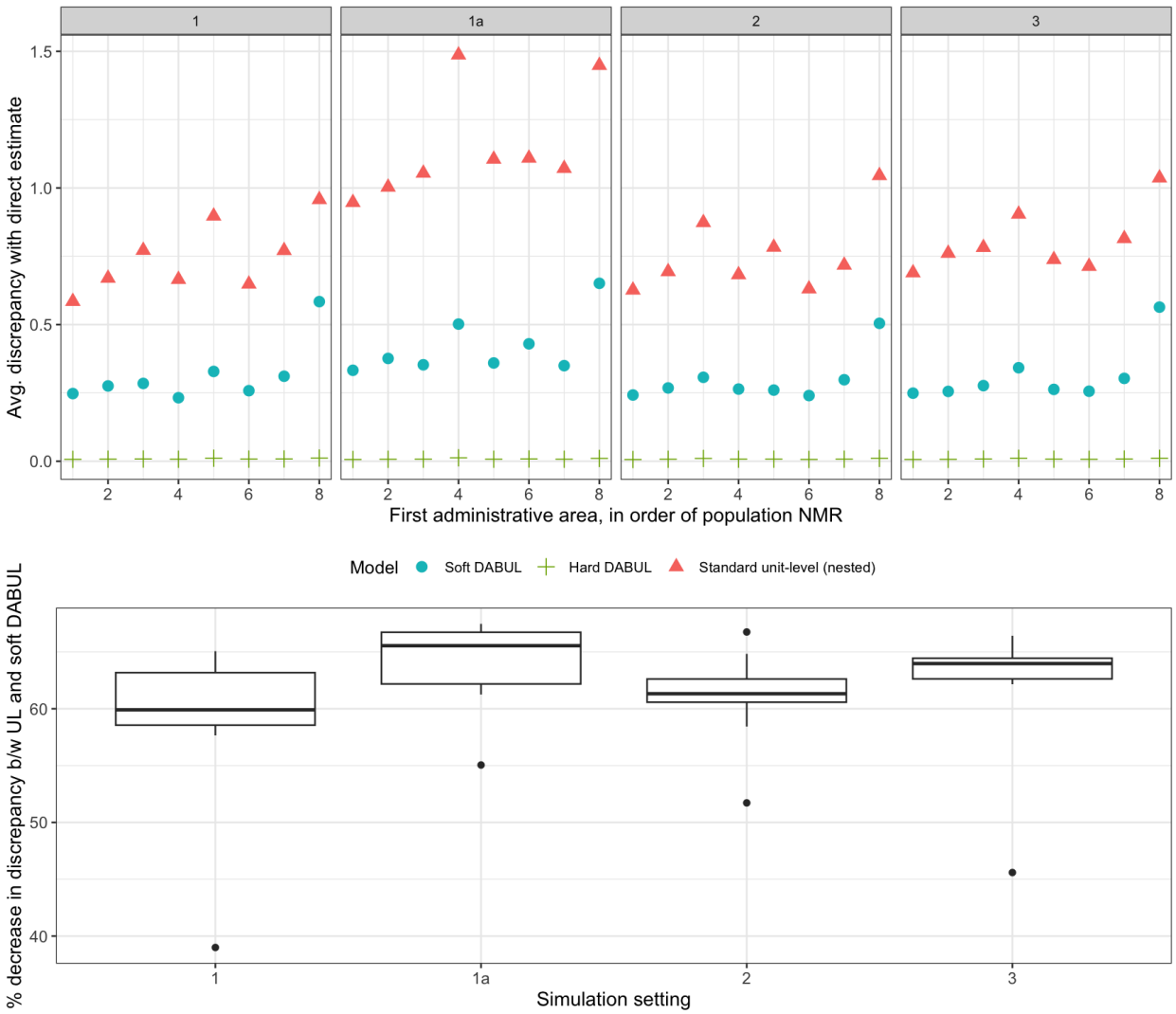


Figure 2.6. Discrepancy between aggregated model-based admin2-level estimates and design-based admin1-level estimates, across 500 simulations for each of 4 settings.

intervals and more consistent coverage at or above the target threshold, compared to the standard unit level model. This higher variability is due to the estimation of the additional parameters $\mathbf{Y}$ (and $\mathbf{Y}_+$ for the soft DABUL model). We observe that the coverage and width of the hard DABUL credible intervals are slightly lower than the soft DABUL credible intervals as a result of not accounting for the uncertainty of the benchmarks. While the credible intervals from the standard unit level model are narrower and have a marginal rate that is closer to the target rate, there are also significantly more areas for which the credi-

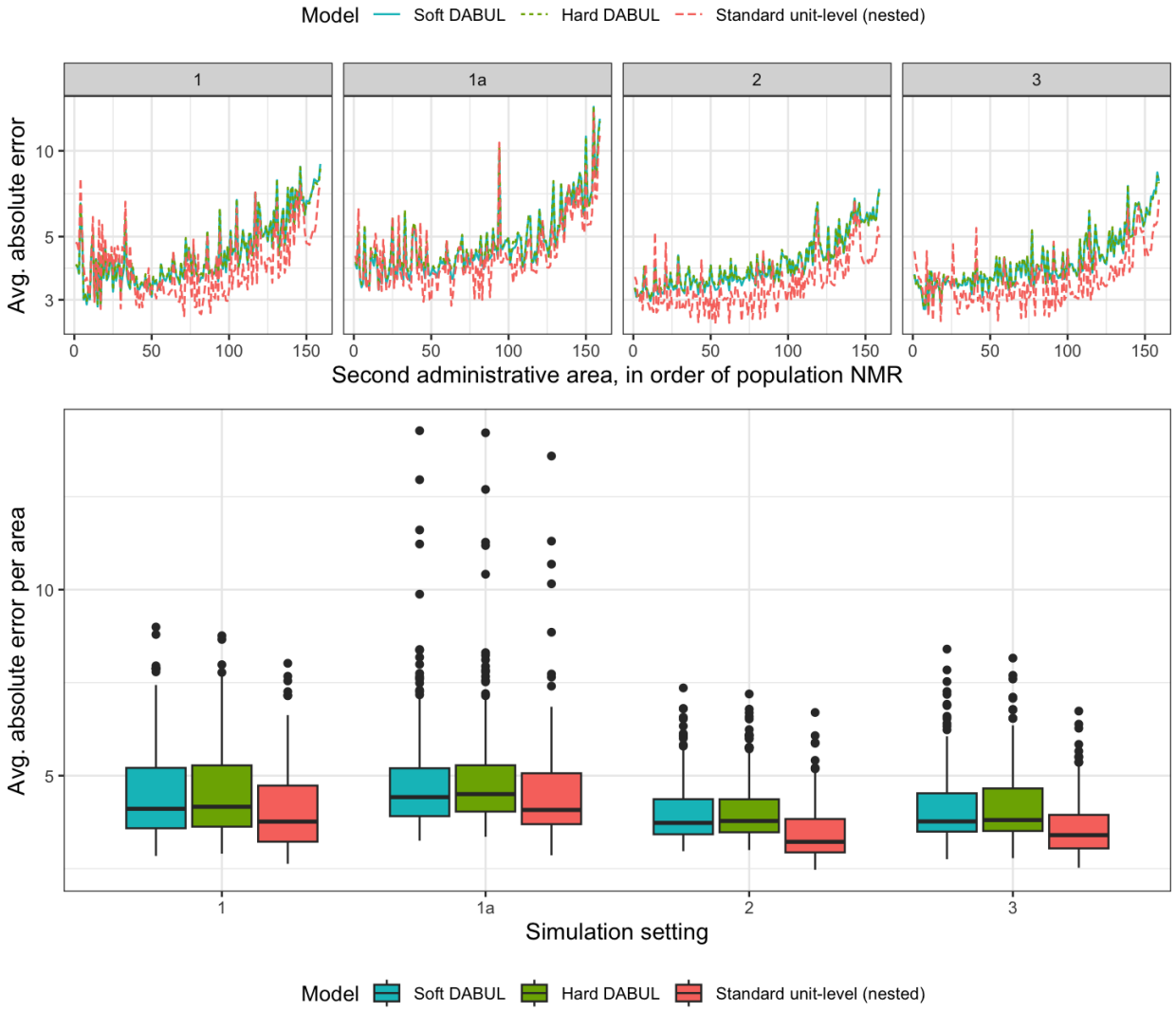


Figure 2.7. Absolute error of admin2-level estimates, across 500 simulations for each of 4 settings.

ble intervals have considerable undercoverage. From these results we observe that DABUL estimates provide more conservative credible intervals, with the benefit that a much higher proportion of areas do not have undercoverage.

2.6 Application to Zambia DHS data

We return to the example of NMR estimation in Zambia and compare the performance of the DABUL models to the standard unit-level model with nested spatial effects. We use the

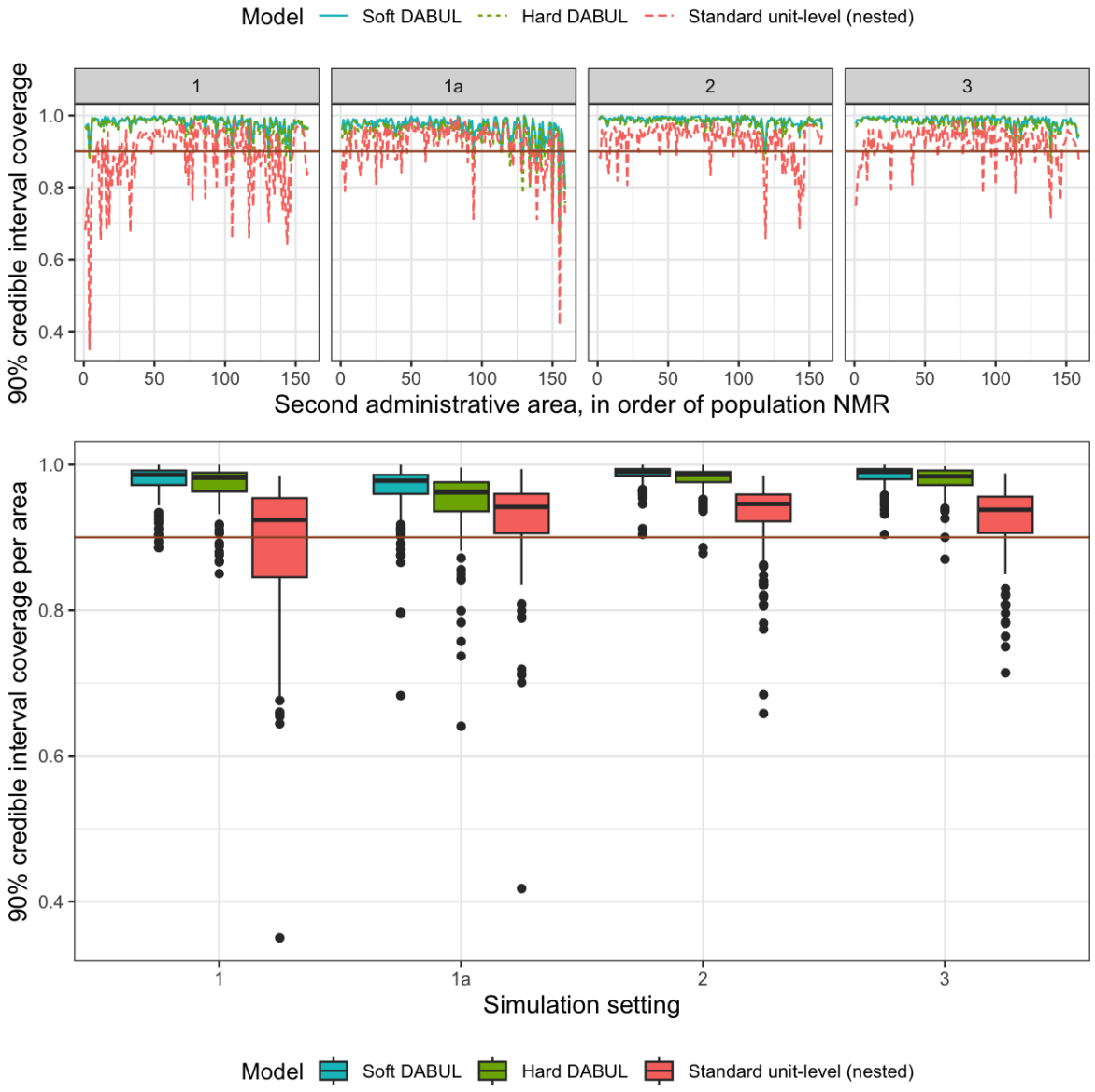


Figure 2.8. Coverage of 90% credible intervals for admin2-level estimates, across 500 simulations for each of 4 settings.

same priors and hyperpriors as in the motivating example. We use WorldPop population estimates (Tatem, 2017) and sampling frame information from the 2013 Zambia DHS (ICF International, 2014) to obtain population counts. The run time of each of the DABUL models was approximately 90 minutes, while the run time of the standard unit-level model in `Stan` was approximately 1 minute.

Figure 2.10 provides a map of the NMR estimates resulting from each of the three models

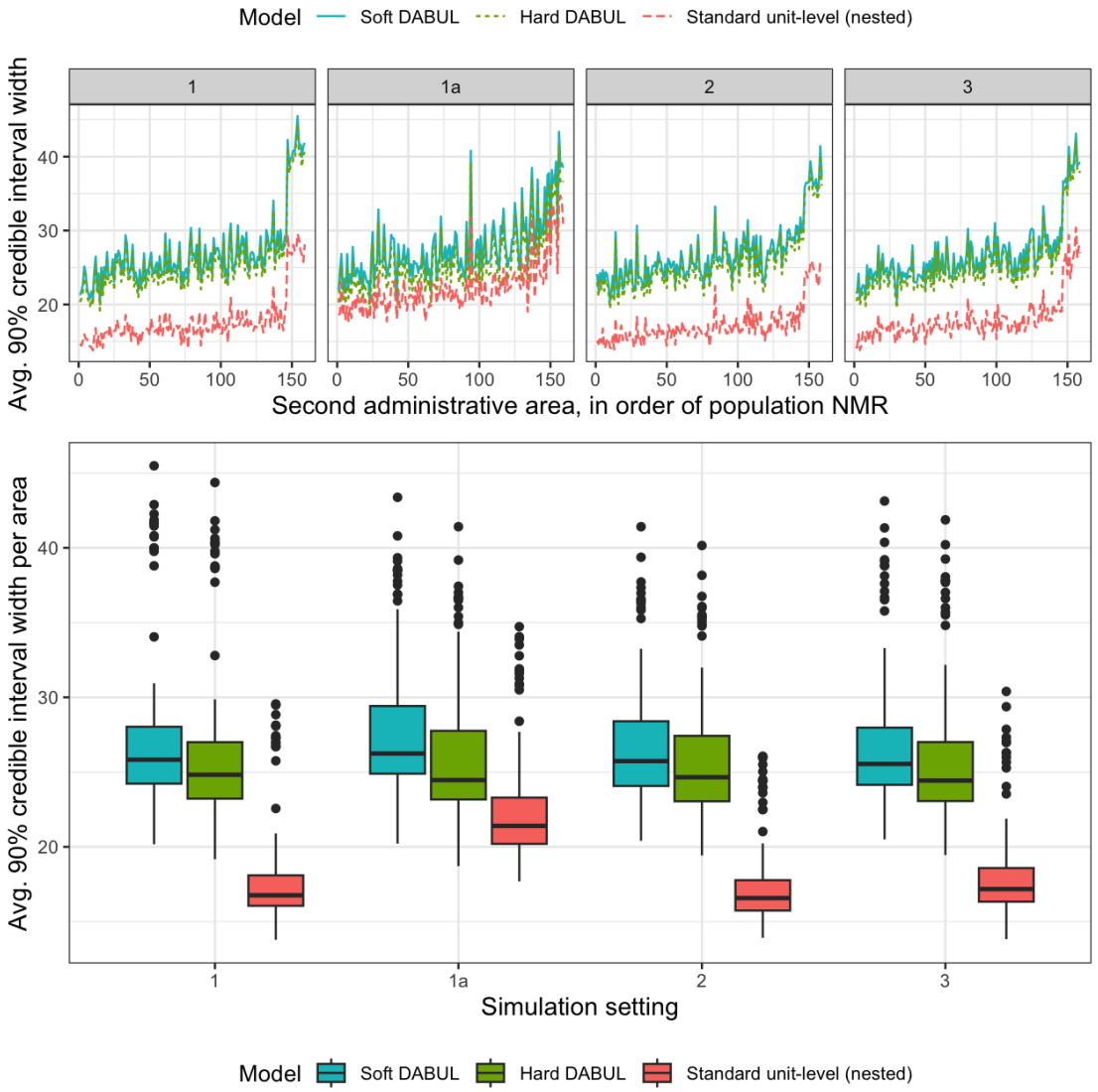


Figure 2.9. Width of 90% credible intervals (in deaths per 1000 births) for admin2-level estimates, across 500 simulations for each of 4 settings.

and Figure 2.11 compares the credible interval width for each admin2 area. A scatter plot comparing the NMR estimates from each of the DABUL models to the standard nested model is presented in Appendix A6. From these figures, we observe that the NMR point estimates are fairly similar between models, with the DABUL estimates having less shrinkage towards the admin1 level. The credible interval widths from the DABUL estimates are uniformly wider than those resulting from the standard unit-level model, which is in agreement with the simulation results.

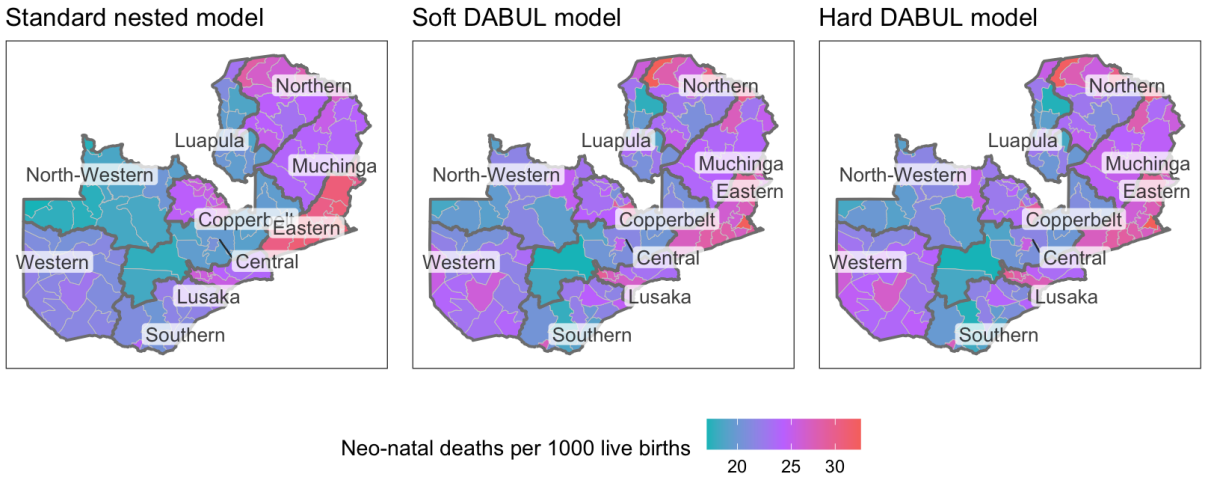


Figure 2.10. Map of NMR estimates at the admin2 level in Zambia (2009-2013) using the DABUL models, as compared to a standard unit-level Bayesian model, both with BYM2 spatial effects at the admin2 level nested based on admin1 level.

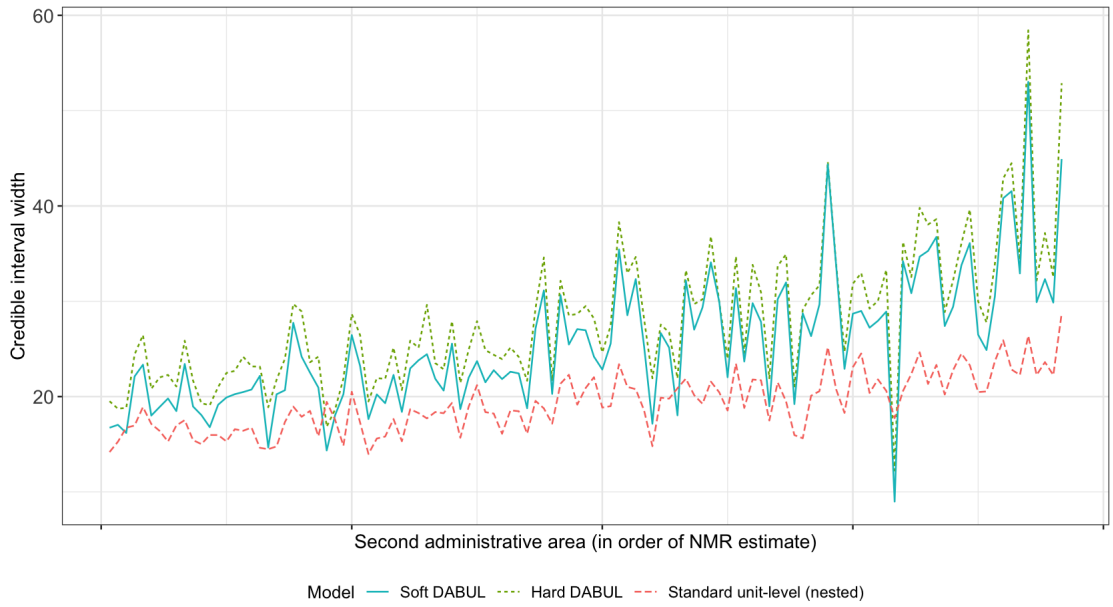


Figure 2.11. Width of 90% credible intervals of NMR estimates (in deaths per 1000 live births) at the admin2 level in Zambia (2009-2013) using the DABUL models, as compared to a standard unit-level Bayesian model.

In Figure 2.12 we evaluate whether the aggregated soft DABUL model estimates have less discrepancy with the direct estimates than the aggregated standard model estimates do. For 6 out of the 10 regions, the DABUL model estimates have similar or less discrepancy

with the direct estimates, while 5 of these regions have a significant reduction in discrepancy. Two of the regions for which discrepancy increases (Luapala and Central) have substantially higher uncertainty, with coefficients of variation of 30% and 31%, respectively, compared to the other admin1-level direct estimates, which have coefficients of variation between 18% and 23%. This is an expected consequence of taking the uncertainty of the benchmarks into account: when uncertainty is higher, the estimates are less likely to adhere to the benchmark. Additionally, the DABUL models produce results with reduced shrinkage, so if the more populous admin2 areas happen to be more extreme areas, the aggregated estimate will be pulled in that direction. If estimates under the soft DABUL model do not have sufficient agreement in aggregation for a particular dataset, the practitioner may instead choose to use the hard DABUL model, with the warning that uncertainty of the benchmarks will not be accounted for.

2.6.1 Model comparison

To compare performance of the DABUL models to the standard unit-level model in this data example, we use a leave-one-area-out cross-validation procedure. While there are cross-validation procedures for count data (Czado et al., 2009), these methods are difficult to use in the context of rare events because the resulting predictive distribution is generally too noisy to meaningfully distinguish predictive ability across the narrow relevant domain. Therefore, instead of leaving out one cluster at a time, we leave out all of the clusters in one admin2 area at a time, and then predict the logit transformed design-based estimate for that admin2 area (Wakefield et al., 2025). By using this approach, we avoid using the predictive distribution of a count variable, in favor of the predictive distribution of a Gaussian variable, which is much more stable. We systematically leave one admin2 area out at a time, and use each of the 3 models to predict the design-based estimate for that admin2 area. This means we can only evaluate predictive ability for admin2 areas which have valid design-based estimates and variances. This condition is met for 75% of admin2 areas (87 out of

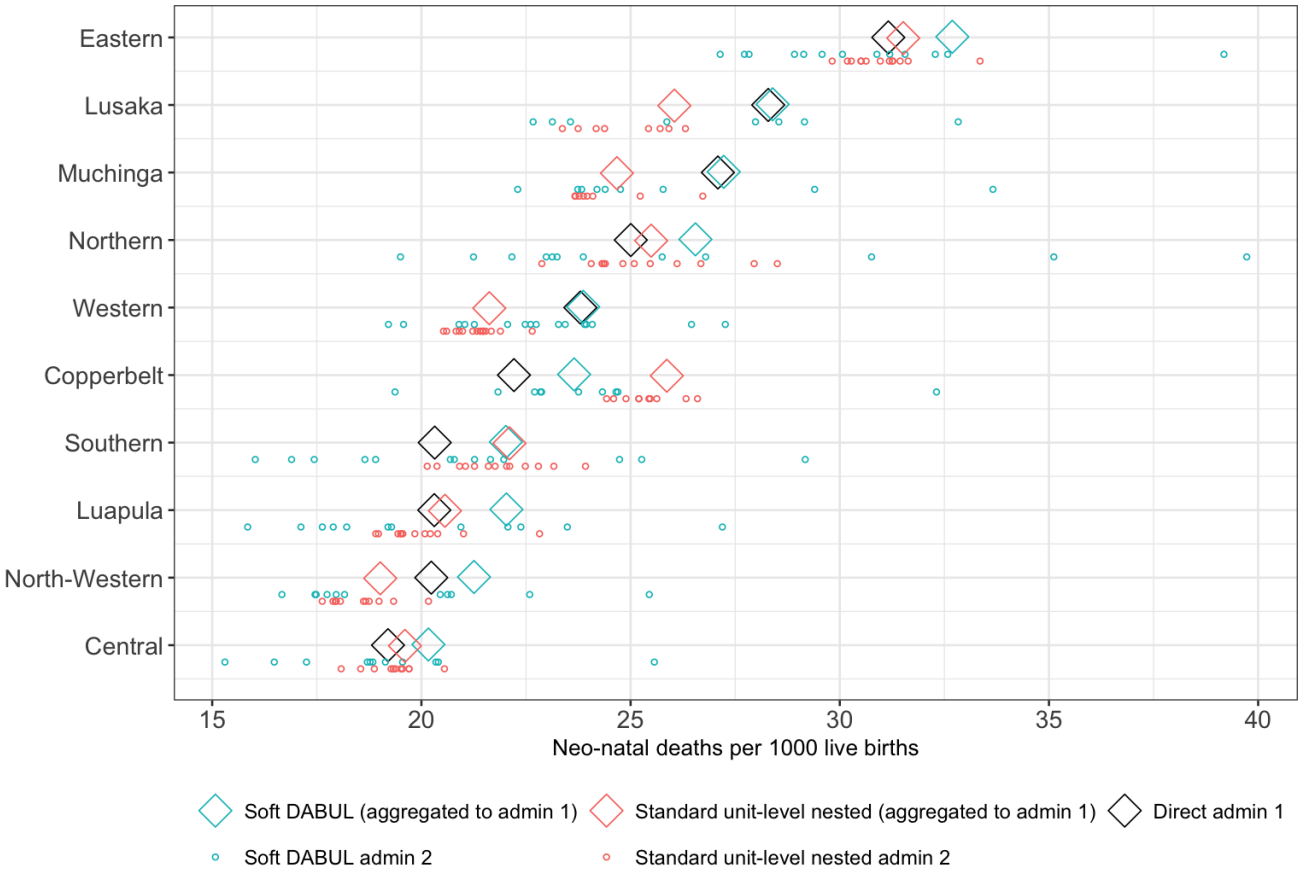


Figure 2.12. NMR estimates at the admin2 level in Zambia (2009-2013) using the soft DABUL model, as compared to a standard unit-level nested Bayesian model. The colored circles denote the admin2-level estimates and the colored diamonds denote their aggregations to the admin1 level. The black diamonds denote Hájek direct estimates at the admin1 level.

115), as the remaining 25% have no observed deaths, so design-based estimates cannot be ascertained. While cross-validation is one of several ways we assess model performance, we emphasize that these results are presented with the caveat that there are valid criticisms of using cross-validation for spatially dependent data and a consensus on best practices has not yet been reached (Bürkner et al., 2020; Adin et al., 2024).

We use three metrics to evaluate predictive ability. First, we calculate absolute error, which is defined as $\left|\hat{r}_{-j} - r_j^{D_2}\right|$, where $\hat{r}_{-j}$ is the admin2-level estimate for area j, when area j data are excluded from the model, and $r_j^{D_2}$ is the design-based estimate for admin2 area j. Note that this metric only evaluates accuracy of the point estimate prediction with respect

to the design-based point estimate. We present this standard metric with caution because it may not be the best description of model performance, as it does not take the high uncertainty of $r_j^{D_2}$ into account. Second, we approximate the log score (or log-CPO) on the logit scale, which is defined as $\log p_{-j}\left(\text{logit}\left(r_j^{D_2}\right)\right)$, where p_{-j} is the predictive density based on the model which excludes area j data. This density does not have a closed form, but we can approximate it by

$$p_{-j}\left(\text{logit}\left(r_j^{D_2}\right)\right) \approx \frac{1}{B}\sum_{b=1}^{B}\phi\left(\text{logit}\left(r_j^{D_2}\right)) \mid \text{logit}\left(\hat{r}_{-j}^{(b)}\right), V_j^{D_2}\right)$$

where $\phi\left(\cdot \mid \text{logit}\left(\hat{r}_{-j}^{(b)}\right), V_j^{D_2}\right)$ is the density of a normal random variable with mean, $\text{logit}\left(\hat{r}_{-j}^{(b)}\right)$, and variance, $V_j^{D_2}$, and $\hat{r}_{-j}^{(b)}$ is the b^{th} draw from the posterior distribution of admin2-level prevalence in area j, based on the model which excludes area j data. This metric also evaluates the accuracy of the point estimate, but with respect to the whole distribution of the design-based estimate. Lastly, we calculate the interval score, a proper scoring rule (Gneiting and Raftery, 2007) which favors narrow intervals containing the true parameter value. Let (l_{-j}, u_{-j}) denote the 90% credible interval for admin2-level prevalence in area j, based on the model which excludes area j data. Then we define the interval score for administrative area j as

$$u_{-j} - l_{-j} + \frac{2}{\alpha}\max\{0, l_{-j} - \text{logit}(\hat{r}_{-j})\} + \frac{2}{\alpha}\max\{0, \text{logit}(\hat{r}_{-j}) - u_{-j,}\}$$

where we take $\alpha = 0.1$. We average each of these 3 metrics over the 87 admin2 areas and compare them across models in Table 2.2. The mean absolute error is very similar across all models, with the soft DABUL model have the lowest value (0.0122). The average log score is also quite similar across models, with the standard unit-level model having the highest value (0.440). The average interval score was lowest for the soft DABUL model (3.1) and highest for the standard unit-level model (4.2). This result shows that although the DABUL intervals are more conservative, this is offset by having better coverage, so, according to this proper scoring rule, the DABUL intervals are preferred over those of the standard unit-level

model.

Table 2.2. Summary of cross-validation metrics averaged over admin2 areas with valid design-based estimates (87 out of 115). The best scores are in bold.

	Standard unit-level (nested)	Soft DABUL	Hard DABUL
Mean absolute error	0.0125	**0.0122**	0.0126
Average log score	**0.440**	0.424	0.421
Average interval score	4.2	**3.1**	3.9

2.7 Discussion

We first observe that in the case of rare events and small sample size, when there are multiple nested areas, using a model with a nested structure can help reduce overshrinkage. We have demonstrated the value of a nested unit-level Bayesian model for estimating rare event prevalence which utilizes design-based estimates at a higher aggregation level. We presented two models: the hard DABUL model which imposes a hard benchmarking constraint, and the soft DABUL model which takes uncertainty of the benchmark into account, thus imposing a soft constraint. In our simulation study and application to Zambia DHS data, we observe that the DABUL models produce conservative credible intervals with much lower frequency of undercoverage, compared to the standard unit-level model. Although the hard DABUL model slightly underestimates the variance by not taking the uncertainty of the direct estimates into account, the resulting credible intervals are still fairly conservative. For this reason, and for the obvious benefit of exact agreement in aggregation, our results suggest that the hard DABUL model may be preferable over the soft DABUL model in some cases. However, our cross-validation study suggests, via average interval scores, that while both DABUL models provide credible intervals which are superior to that of the standard unit-level model, the soft DABUL model provides the credible intervals with the best performance. Further study is necessary to determine the costs and benefits of choosing between the two DABUL models, however, the soft DABUL model remains the more conservative, principled approach because of its proper accounting for the large uncertainty of

the benchmarks.

One criticism of the use of generalized linear unit-level models for complex survey data, is that they do not directly take sample design into account. Solutions have been developed for the case of linear models (You and Rao, 2002), but they do not extend to the generalized linear case (Wakefield et al., 2025). While the DABUL models do not completely solve this open problem, they are an improvement in the sense that they take design-based estimates at a higher aggregation level into account.

The DABUL models could also be extended to include covariates, which is encouraged whenever covariates with predictive power are available. We do not include covariates here because previous work has shown that available covariates in LMICs have little effect on predictive power in estimating NMR when spatial random effects are already included in the model (Golding et al., 2017,Wakefield et al., 2019). It is straightforward to include covariates in the DABUL models, though the aggregation step can introduce additional error.

Because the DABUL models, as derived, use a Negative Binomial sampling model, they require the outcome to be a rare event. While this model could theoretically be extended to estimate prevalence of non-rare events by using a Binomial or Beta-Binomial sampling model, these distributions do not enjoy the same additive properties as the Poisson and Negative Binomial distributions. As a result, the distributions of Y_{i+} and $\tilde{Y}_i \mid Y_{i+}$ quickly become unwieldy as they require summing over a complete enumeration of the state space. Conversely, this model could easily be extended to continuous outcomes because the Normal distribution does share these additive properties. The only necessarily changes to the model would be to replace (8) and (9) with the Gaussian distributions implied by the Gaussian likelihood.

Chapter 3

SAMPLING DISTRIBUTIONS FOR COMPLEX SURVEY DESIGN VARIANCE ESTIMATORS IN A FAY-HERRIOT MODEL

3.1 Introduction

The Fay-Herriot area-level model is a common choice for SAE because it accounts for the design by modeling survey-weighted estimates and enables borrowing of information across areas through covariates and random effects, which often increases precision and is particularly beneficial when areas have little or no data (Fay and Herriot, 1979). In the sampling model for weighted means, the variances of the weighted mean estimators are assumed to be known, but in practice, estimates of these variances are used. A common estimator of this variance is obtained via Taylor linearization (Wolter, 2007, chap. 6; Korn and Graubard, 2011, chap. 2.4; Särndal et al., 2003, chap. 5.5; Lohr, 2021, chap. 9.1). While such estimators are design-consistent, they can be imprecise, particularly when sample size is small. A less-considered detriment is the tendency for this variance estimator to have downward bias for small sample sizes (Särndal et al., 2003, p. 176; Lohr, 2021, p. 369). As a result, taking variance estimates as known in the Fay-Herriot model can result in interval estimates which are not only imprecise, but have systematic undercoverage for small sample sizes.

This work is motivated by the challenges posed by estimating small area means using survey data in LMICs. The major surveys conducted in LMICs, particularly the Demographic and Health Surveys (DHS) (Corsi et al., 2012) and Multiple Indicator Cluster Surveys (MICS) (UNICEF, 2026), use stratified two-stage cluster designs where at the first stage a fixed number of clusters are sampled from each stratum using probability proportional to size (PPS) sampling, and at the second stage a fixed number of households are sampled from each cluster. These strata are generally chosen to be the admin1 areas crossed with urban and rural classification, resulting in data which is powered to provide reliable survey-weighted estimates at the admin1 level. This design makes estimation for unplanned

domains, for example, admin2 areas, challenging because data are sparse.

As an example, Figure 3.1 maps the number of clusters sampled from admin1 and admin2 areas in the 2022 Kenya DHS (ICF International, 2022). Design-based estimation for the 47 admin1 areas is feasible, as they are planned domains with an average of 36 clusters sampled per area. On average, 1.76% of rural clusters and 3.8% of urban clusters in each admin1 area are sampled. While the number of urban and rural clusters sampled from each admin1 area is fixed under the survey design, the number of clusters sampled from each admin2 area is random. In this survey, 6 out of Kenya's 300 admin2 areas have no sampled clusters and 127 have fewer than five sampled clusters. As a result, there are several areas for which weighted mean estimates cannot be obtained and many for which estimates are imprecise or have unstable variance estimates.

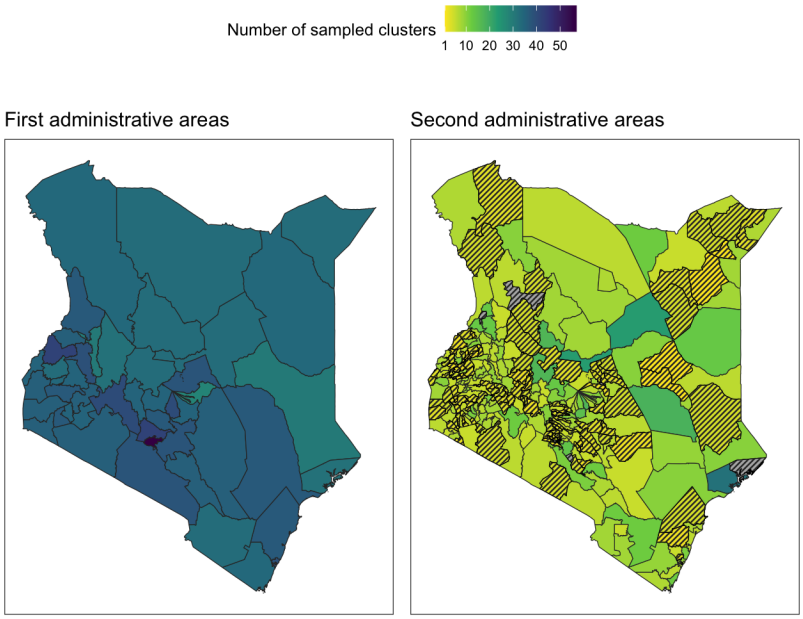


Figure 3.1. Number of sampled clusters in each admin1 and admin2 area in the 2022 Kenya DHS. Admin2 areas with fewer than 5 sampled clusters are marked by hatching and areas with no sampled clusters are filled in gray.

For example, consider the estimation of average height-for-age z-score (HAZ) for children under 5. The HAZ measures how many standard deviations a child's height is from the median height for children of the same sex and age, using median and standard deviations

from the WHO Multicentre Growth Reference Study Group (2006). This measure is useful to evaluate child nutritional status relative to the international population. Reducing the rate of stunting, defined as a HAZ below -2, is one of the Sustainable Development Goals (SDGs) targeting malnutrition (United Nations General Assembly, 2015). In Figure 3.2 we map the design-based mean and standard deviation estimates for HAZ for children under 5, using the Kenya 2022 DHS. Note that design-based variance estimates cannot be obtained for admin2 areas with only one sampled cluster. We observe notable between-area variation at the admin1 and admin2 levels. In this context it is particularly detrimental to use a Fay-Herriot model at the admin2 level which assumes the variances of the weighted estimates are known, as data sparsity makes variance estimates more unstable.

There is a vast literature concerning how variance estimates may be adjusted or have their uncertainty accounted for. When the outcome of interest is a proportion or total, one common strategy is to use a generalized variance function (GVF), which exploits the mean-variance relationship assumed under a binomial distribution (Valliant, 1987; Wolter, 2007, chap. 7; Korn and Graubard, 2011, chap. 5.6; Lohr, 2021, chap. 9.4). However, in the case of continuous outcomes, justifying any particular choice of GVF becomes difficult, as there is not a clear mean-variance relationship to exploit. Some work has been done by the US Bureau of Labor Statistics to use GVFs for time series analyses of continuous outcomes (Cho et al., 2002; McIllece, 2018), but this requires frequent, data-rich surveys.

Another common strategy is to jointly model the mean and variance, so that the uncertainty of the variance estimate is taken into account. This strategy requires specifying a sampling distribution for the design variance estimator. Previous work, including You and Chapman (2006), Maiti et al. (2014), Sugasawa et al. (2017), Erciulescu et al. (2019), and Parker et al. (2024), assume that the survey is a stratified simple random sample with each area of interest, i, being a stratum, and assume individual outcomes within each stratum are identically and independently distributed (IID) normal. These assumptions result in the variance estimator, $\hat{V}_i$, having the sampling distribution $V_i/\mathrm{df}_i \times \chi^2_{\mathrm{df}_i}$, where V_i is the true

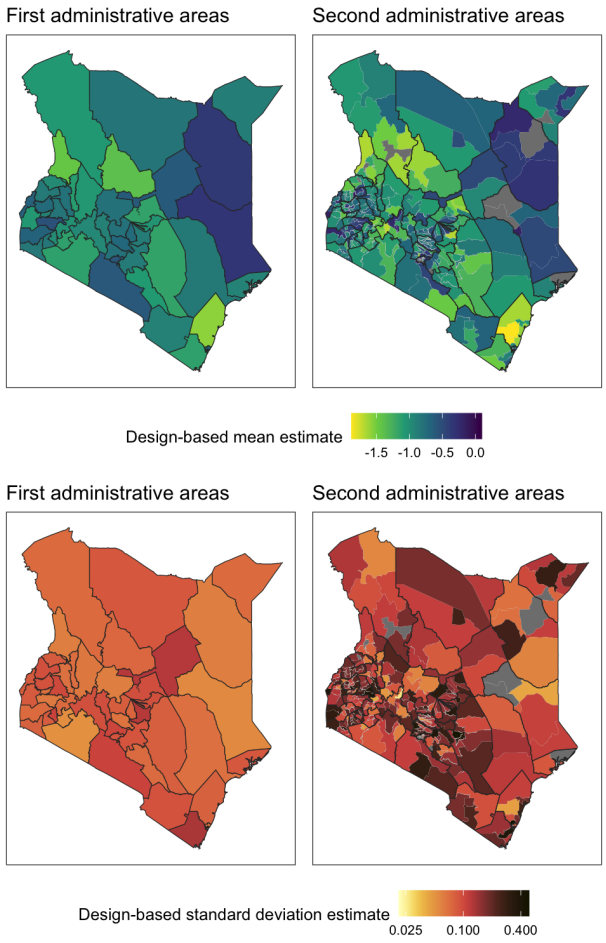


Figure 3.2. Design-based mean and standard deviation estimates for HAZ at the admin1 and admin2 levels using 2022 Kenya DHS. Areas for which design-based estimates cannot be obtained are filled in gray.

design variance and the degrees of freedom (df_i) are equal to the sample size minus one.

Gao and Wakefield (2023) modify this sampling distribution to account for a stratified cluster sampling design by using the appropriate nominal degrees of freedom, which is the number of sampled clusters minus the number of strata (Korn and Graubard, 2011, p. 34). Little attention has been paid to the underlying design and population assumptions implied by this sampling distribution and whether interval estimates of small areas means resulting from such a model are robust to violations of these assumptions. In this work we will thoroughly examine the assumptions which are intrinsic to the commonly used sampling distribution, introduce a sampling distribution which makes fewer design assumptions, and

assess their performance in a Fay-Herriot model with complex survey data.

We begin in Section 3.2 by introducing the two-stage cluster sampling design and defining the Fay-Herriot model using a superpopulation framework. In Section 3.3 we derive two sampling distributions for the design variance estimators under a stratified two-stage cluster sampling design and several design and superpopulation assumptions. In Section 3.4 we fully specify our variance smoothing Fay-Herriot model using each of the candidate sampling distributions under a Bayesian framework. In Section 3.5 we conduct a simulation study in which we compare these sampling distributions to the empirical distribution of the design variance estimator, as well as the model performance of a standard Fay-Herriot model (without variance smoothing) against Fay-Herriot models with variance smoothing using each of the proposed sampling distributions. In Section 3.6 we apply these models to HAZ data from the Kenya 2022 DHS and conclude with a discussion of our findings in Section 3.7.

3.2 Fay-Herriot model for stratified two-stage cluster samples

3.2.1 Notation and setup

Appendix B1 provides a summary of definitions and notation introduced throughout this chapter. Consider a population composed of strata $\mathcal{H} = \{1, ..., H\}$, where $H \geq 1$ is the number of strata. Stratum $h \in \mathcal{H}$, contains M_h clusters, the set of which is denoted U_h. Cluster $c \in U = \cup_{h \in \mathcal{H}} U_h$ contains N_c individuals. Let $y_{c\ell}$ denote the value of some continuous outcome for individual $\ell = 1, ..., N_c$ in cluster $c \in U$. Suppose we are interested in estimating the population means of this continuous outcome measure for a set of areas, $i = 1, ..., K$ wherein area i spans the subset, $\mathcal{H}_i \subseteq \mathcal{H}$. Let U_{hi} be the subset of clusters in U_h which are contained in area i, with $M_{hi} = |U_{hi}|$. It follows that $U_{hi} = \emptyset$ for $h \in \mathcal{H} \setminus \mathcal{H}_i$. We take as the parameters of interest, the area-specific population means,

$$\theta_i = \frac{\sum_{h \in \mathcal{H}_i} \sum_{c \in U_{hi}} \sum_{\ell=1}^{N_c} y_{c\ell}}{\sum_{h \in \mathcal{H}_i} \sum_{c \in U_{hi}} N_c}, \qquad i = 1, ..., K.$$

Now consider a survey design in which at the first stage, $m_h < M_h$ clusters are sampled with PPS sampling based on the population size of the cluster, for $h \in \mathcal{H}$, the set of which is denoted $S_h \subset U_h$, and in the second stage, $n_c < N_c$ individuals are sampled with equal probability from each cluster $c \in S = \cup_{h\in\mathcal{H}} S_h$. Let the indicator variable, ζ_{cl}, denote whether individual ℓ in cluster c was sampled. Note that this is negligibly different from the DHS design, which samples households at the second stage, rather than individuals. For the first stage, the sampling probability is determined by the listed size of cluster c, L_c, in a master sampling frame which is usually based on the most recent census. Thus the first stage sampling probability for cluster c in stratum h, is $p_{1,c} = m_h L_c / \sum_{c'\in U_h} L_{c'}$. In practice, the true size of cluster c, N_c, will likely be different from that which is listed in the sampling frame, so in the second stage, the sampled clusters are re-enumerated and the second stage sampling probability is $p_{2,c} = n_c/N_c$. The sampling weight for cluster c in stratum h is

$$w_c = \frac{1}{p_{1,c}p_{2,c}} = \frac{N_c \sum_{c'\in U_h} L_{c'}}{L_c m_h n_c}.$$

As a result of this design, each individual in the same cluster has the same sampling weight, w_c. We assume throughout that $n_c \ll N_c$ and $m_h \ll M_h$, therefore the finite population corrections are negligible.

Let S_{hi} be the subset of clusters in S_h which are contained in area i, with $m_{hi} = |S_{hi}|$, and let $m_{\cdot i} = \sum_{h\in\mathcal{H}_i} m_{hi}$ be the total number of sampled clusters in area i. Note that if area i is a planned domain, $S_{hi} = S_h$ for all $h \in \mathcal{H}_i$, whereas if area i is an unplanned domain, $S_{hi} \subsetneq S_h$, for $h \in \mathcal{H}_i$. Then the Hájek estimator (Hájek, 1971) for θ_i is

$$\hat{\theta}_i = \frac{\sum_{h\in\mathcal{H}_i} \sum_{c\in S_{hi}} w_c^\star \bar{y}_c}{\sum_{h\in\mathcal{H}_i} \sum_{c\in S_{hi}} w_c^\star} \tag{19}$$

where $\bar{y}_c = \left(\sum_\ell^{N_c} \zeta_{c\ell} y_{c\ell}\right)/n_c$ and $w_c^\star = w_c n_c$.

Following Equation 2.3-10 of Korn and Graubard (2011), we apply the Taylor linearization method to expression (19) and obtain a design-consistent estimator for $\mathrm{Var_D}(\hat{\theta}_i)$ condi-

tioning on cluster sampling weights and sample sizes,

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_h} w_c^\star\right)^2} \sum_{h\in\mathcal{H}_i} \frac{m_h}{m_h-1} \sum_{c\in S_h} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_h} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \tag{20}$$

where $w_c^\star = 0$ for all $c \in S_h \setminus S_{hi}$ and $h \in \mathcal{H}_i$, as these clusters are outside the area of interest. When the area is an unplanned domain we sum over the clusters in the larger planned domain to account for the additional variability introduced by the fact that the number of sampled clusters in the unplanned domain is random. See Appendix B2.1 for more details. Expression (20) is equivalent to the expression in SAS Institute Inc. (2016, p. 9282–9283), and yields estimates which are nominally the same as those obtained using the `survey` package (Lumley, 2024). If only one cluster is sampled from area i, $\hat{\theta}_i = \bar{y}_c$, so it follows that $\hat{V}_i = 0$. While adjustments have been proposed to allow for variance estimation in this limiting case (Wakefield et al., 2026), we do not explore them here and assume that $m_{.i} > 1$.

The standard Fay-Herriot model is

$$\hat{\theta}_i \mid \theta_i \sim \mathcal{N}(\theta_i, \hat{V}_i), \qquad \theta_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + b_i \tag{21}$$

where $\boldsymbol{X}_i$ is an area-specific auxiliary variable vector, $\boldsymbol{\beta}$ is a vector of fixed effect regression parameters, and $\boldsymbol{b} = (b_1, ..., b_K)^{\mathrm{T}}$ is a vector of the residual between-area variation, which are modeled as normal area-level random effects. In the original formulation of the Fay-Herriot model, the random effect is IID normal, but one may also include a spatially structured component in the random effect (Rao and Molina, 2015, chap. 4.4.4; Chung and Datta, 2020; Gao and Wakefield, 2023; Wakefield et al., 2025).

3.2.2 Variance smoothing Fay Herriot model

Now we will extend this framework to jointly model the design variances along with the means. It is common in the variance smoothing literature (You and Chapman, 2006; Maiti

et al., 2014; Sugasawa et al., 2017; Erciulescu et al., 2019; Parker et al., 2024) to use the sampling distribution,

$$\hat{V}_i \mid V_i \sim \frac{V_i}{\mathrm{df}_i}\chi^2_{\mathrm{df}_i} \tag{22}$$

where $\mathrm{df}_i = \left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} n_c\right) - 1$ and V_i is the true design variance. You and Chapman (2006) give V_i a diffuse inverse-gamma prior, which Maiti et al. (2014) and Sugasawa et al. (2017) note often leads to unstable estimation. The former modifies You and Chapman's model by taking an empirical Bayes approach, and the latter suggests using more informative hyperparameters and including covariates. Erciulescu et al. (2019) chooses a normal distribution for $\log(V_i)$ conditional on covariates. As Maiti et al. (2014) note, the sampling distribution expressed in (22) is only valid under simple random sampling of IID normal data. A reasonable modification of (22) for the complex survey case, to account for stratification and clustering, is to simply plug in the appropriate nominal degrees of freedom for a complex survey, $\mathrm{df}_i = \sum_{h\in\mathcal{H}_i}(m_{hi} - 1)$, as is done in Gao and Wakefield (2023).

In contrast to the above work which takes a design-based approach by directly modeling the design variance, V_i, we will use a superpopulation approach and model the superpopulation variance, σ_i^2, along the with the superpopulation mean, θ_i. In particular, we assume

$$y_{c\ell} \mid \theta_i, \gamma_h, \sigma_i^2 \sim \mathcal{N}(\theta_{hi}, \sigma_i^2), \qquad \text{for all } c \in U_{hi}, \tag{23}$$

where $\theta_{hi} = \theta_i + \gamma_h$ is the superpopulation mean for strata h and area i, $(\gamma_h)_{h\in\mathcal{H}}$ are stratum mean effects, and σ_i^2 is the within-stratum superpopulation variance for area i. These strata-specific means are well-defined under the Fay-Herriot model if $\gamma_h = 0$ for one $h \in \mathcal{H}_i$ and we use the proportion of the population that belongs to each strata as an auxiliary variable in the latent mean model, as we will specify in Section 3.4. It follows from (19) that, for fixed cluster sampling weights and sample sizes, the design variance under superpopulation

assumption (23) and the survey design described in Section 3.2.1 is

$$V_i^\star(\sigma_i^2) = \sigma_i^2 \frac{\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^{\star 2}/n_c}{(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star)^2}. \tag{24}$$

We choose a latent model for σ_i^2 which mirrors that of the latent mean model, namely,

$$\log(\sigma_i^2) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\eta} + e_i$$

where $\boldsymbol{Z}_i$ is an area-specific auxiliary variable vector, $\boldsymbol{\eta}$ is a vector of fixed effect regression parameters, and $\boldsymbol{e} = (e_1, ..., e_K)$ is a vector of residuals, modeled as normal area-level random effects. This formulation allows for leveraging of any existing spatial structure, via the random effect and covariates, and an intuitive interpretation of the random effects and covariates.

3.3 Sampling distributions for the complex design variance estimator

3.3.1 Simple sampling distribution

Here we derive a sampling distribution for (20) which is similar to the one used in Gao and Wakefield (2023), but follows a superpopulation framework. We impose five assumptions on the survey design which are sufficient to obtain this sampling distribution:

(i) Area i is a planned domain (i.e., $S_{hi} = S_h$ and $m_{hi} = m_h$ for all $h \in \mathcal{H}_i$).

(ii) The same number of clusters are sampled from each stratum in area i ($m_{hi} = m_{\cdot i}/|\mathcal{H}_i|$).

(iii) Every sampled cluster in area i has the same sample size, n_0.

(iv) The total number of individuals in stratum h, $\sum_{c\in U_h} N_c$, is equal for all strata $h \in \mathcal{H}_i$.

(v) Re-enumeration of cluster sizes as described in Section 3.2.1 is not needed for any cluster in area i.

Note that assumptions (iii), (iv), and (v) imply that all clusters in area i have the same sampling weight. Under these design assumptions the variance estimator (20) simplifies to

$$\hat{V}_i = \frac{1}{m_{\cdot i}(m_{\cdot i} - |\mathcal{H}_i|)} \sum_{h \in \mathcal{H}_i} \sum_{c \in S_{hi}} \left[\bar{y}_c - \frac{1}{m_{hi}} \sum_{c' \in S_{hi}} \bar{y}_{c'} \right]^2 . \tag{25}$$

See Appendix B2.2 for more details.

Under these design assumptions and superpopulation assumption (23), it follows that $\bar{y}_c \sim \mathcal{N}(\theta_{hi}, \sigma_i^2/n_0)$, for all $c \in S_{hi}$, and

$$\hat{V}_i \mid \sigma_i^2 \sim \frac{V_i^\dagger(\sigma_i^2)}{m_{\cdot i} - |\mathcal{H}_i|} \sum_{h \in \mathcal{H}_i} \chi^2_{m_{hi}-1} = \frac{V_i^\dagger(\sigma_i^2)}{m_{\cdot i} - |\mathcal{H}_i|} \chi^2_{m_{\cdot i} - |\mathcal{H}_i|}, \tag{26}$$

where

$$V_i^\dagger(\sigma_i^2) = \frac{\sigma_i^2}{\sum_{h \in \mathcal{H}_i} \sum_{c \in S_{hi}} n_c} \tag{27}$$

is the design variance of $\hat{\theta}_i$ under this model.

As a consequence of assuming cluster sample sizes are equal, we can relax the assumption of independence within clusters and allow for all clusters in area i to have the same within-cluster correlation. To illustrate this, suppose that individual outcomes in cluster $c \in U_{hi}$ are normally distributed with mean, θ_{hi}, variance σ_i^2, and a correlation of ρ_0 between all individuals in cluster c. Then it follows that

$$\mathrm{Var}(\bar{y}_c) = \sigma_i^2 \left(1 + \rho_0 \left(n_0 - 1\right)\right) / n_0,$$

so the within-cluster correlation is absorbed into the estimation of σ_i^2.

3.3.2 Survey-weighted sampling distribution

In this section we derive a sampling distribution for (20), under the survey design described in Section 3.2.1, thus requiring fewer design assumptions than the simple sampling distribution

presented in (26). From superpopulation assumption (23) and simple random sampling within each cluster it follows that

$$\bar{\boldsymbol{y}}_{\boldsymbol{i}} \mid \theta_i, \boldsymbol{\gamma}_{\boldsymbol{i}}, \sigma_i^2 \sim \mathcal{N}\left(\theta_i \mathbf{1} + \boldsymbol{\gamma}_{\boldsymbol{i}}, \sigma_i^2 \mathbf{D}_{\boldsymbol{i}}\right) \tag{28}$$

where $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ is composed of subvectors, $(\bar{\boldsymbol{y}}_h)_{h\in\mathcal{H}_i}$, of length m_h with elements $(\bar{y}_c)_{c\in S_h}$, $\boldsymbol{\gamma}_{\boldsymbol{i}}$ is a vector composed of subvectors, $(\gamma_h \mathbf{1}_{m_h})_{h\in\mathcal{H}_i}$, and $\mathbf{D}_{\boldsymbol{i}}$ is a a diagonal matrix with elements $(1/n_c)_{c\in\cup_{h\in\mathcal{H}_i} S_h}$.

Let $\mathbf{B}_{\boldsymbol{i}}$ be a block diagonal matrix with blocks $(\mathbf{B}_h)_{h\in\mathcal{H}_i}$ where $\mathbf{B}_h = \frac{m_h}{m_h-1}\left(\mathbf{I}_{m_h} - \frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)$ and $\mathbf{1}_{m_h}$ is an m_h-length column vector of 1s. Let $\boldsymbol{w_i}^\star$ be a vector composed of subvectors, $(\boldsymbol{w}_h^\star)_{h\in\mathcal{H}_i}$, with elements $(w_c^\star)_{c\in S_h}$ and define $\mathbf{T}_{\boldsymbol{i}} = \mathbf{W}_{\boldsymbol{i}}\left(\mathbf{I}_{m_{\cdot i}} - \frac{\mathbf{1}\boldsymbol{w_i}^{\star\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star}\right)$, where $\mathbf{W}_{\boldsymbol{i}} = \mathrm{diag}\left(\boldsymbol{w_i}^\star\right)$. Note that $(\bar{\boldsymbol{y}}_{\boldsymbol{i}}, \boldsymbol{\gamma}_{\boldsymbol{i}}, \mathbf{D}_{\boldsymbol{i}}, \mathbf{B}_{\boldsymbol{i}}, \boldsymbol{w_i}^\star)$ must maintain consistent ordering. Then we can express the weighted mean estimator (19) and its corresponding variance estimator (20) as,

$$\hat{\theta}_i = \frac{\boldsymbol{w_i}^{\star\mathrm{T}}\bar{\boldsymbol{y}}_{\boldsymbol{i}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star}, \qquad\qquad \hat{V}_i = \frac{1}{\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star\right)^2}\bar{\boldsymbol{y}_i}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}} \tag{29}$$

where $\mathbf{M}_{\boldsymbol{i}} = \mathbf{T}_{\boldsymbol{i}}^{\mathrm{T}}\mathbf{B}_{\boldsymbol{i}}\mathbf{T}_{\boldsymbol{i}}$. See Appendix B2.3 for more details regarding the equality of the form of the variance estimator expressed in (20) and the matrix form expressed in (29).

Let $r_i \leq m_{\cdot i}$ denote the rank of $\mathbf{D}_{\boldsymbol{i}}^{1/2}\mathbf{M}_{\boldsymbol{i}}\mathbf{D}_{\boldsymbol{i}}^{1/2}$, $\boldsymbol{q}_{\boldsymbol{i}}$ denote the r_i-dimensional vector of non-zero eigenvalues of $\mathbf{D}_{\boldsymbol{i}}^{1/2}\mathbf{M}_{\boldsymbol{i}}\mathbf{D}_{\boldsymbol{i}}^{1/2}$, and $\boldsymbol{v}_{\boldsymbol{ij}}$ denote the $m_{\cdot i}-$length eigenvector corresponding to eigenvalue q_{ij}, scaled such that $\boldsymbol{v}_{\boldsymbol{ij}}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{i}}\boldsymbol{v}_{\boldsymbol{ij}} = 1$. Then under (28), we obtain a sampling distribution for $\hat{V}_i$, as a sum of scaled chi-squares, namely,

$$\hat{V}_i \mid \boldsymbol{\gamma}_{\boldsymbol{i}}, \sigma_i^2 \sim \frac{\sigma_i^2}{\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star\right)^2}\sum_{j=1}^{r_i} q_{ij}\chi_1^2(\delta_{ij})$$

where $\chi_1^2(\delta_{ij})$ denotes a chi-square distribution with 1 degree of freedom and noncentrality parameter, $\delta_{ij} = \left(\boldsymbol{v_{ij}}^{\mathrm{T}}\boldsymbol{\gamma}_{\boldsymbol{i}}\right)^2/\sigma_i^2$. Recall from Equation (24) that under these superpopulation

assumptions, the design variance of $\hat{\theta}_i$ is

$$V_i^\star(\sigma_i^2) = \sigma_i^2 \frac{\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^{\star 2}/n_c}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star\right)^2} = \sigma_i^2 \frac{\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star}{\left(\mathbf{1}^\mathrm{T}\boldsymbol{w_i}^\star\right)^2},$$

so it follows that this sampling distribution for $\hat{V}_i$, which we will refer to as the 'survey-weighted' sampling distribution, can also be expressed as

$$\hat{V}_i \mid \boldsymbol{\gamma_i}, \sigma_i^2 \sim \frac{V_i^\star(\sigma_i^2)}{\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star} \sum_{j=1}^{r_i} q_{ij}\chi_1^2(\delta_{ij}). \tag{30}$$

For ease of implementation in the Fay-Herriot model, it is preferable to approximate the sum of non-central chi-square random variables of (30) with a single random variable. The Satterthwhaite approximation is a single scaled chi-square random variable which matches the first two moments of a sum of scaled chi-squares (Satterthwaite, 1946). The mean and variance of $\sum_j^{r_i} q_{ij}\chi^2(1,\delta_{ij})$ are $Q_1 = \sum_j^{r_i} q_{ij}(1+\delta_{ij})$ and $Q_2 = 2\sum_j^{r_i} q_{ij}^2(1+2\delta_{ij})$, respectively, so by applying this moment-matching approach, we obtain the Satterthwhaite-approximated survey-weighted (SASW) sampling distribution,

$$\hat{V}_i \mid \boldsymbol{\gamma_i}, \sigma_i^2 \sim \frac{V_i^\star(\sigma_i^2)}{\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star} \frac{Q_2}{2Q_1}\chi^2_{2Q_1^2/Q_2}. \tag{31}$$

In Appendix B3 we illustrate via several examples that this approximation has high accuracy. Note that $\hat{V}_i \mid \sigma_i^2, \boldsymbol{\gamma_i}$ is a biased estimator for $V_i^\star(\sigma_i^2)$ under this sampling distribution. We show in Appendix B4 that, under mild design assumptions, this bias is downward (i.e., $\mathbb{E}[\hat{V}_i \mid \sigma_i^2, \boldsymbol{\gamma_i}] - V_i^\star(\sigma_i^2) < 0$), which reflects the fact that while $\hat{V}_i$ is a consistent estimator for the design variance, it is generally a downwardly-biased estimator for finite samples (Särndal et al., 2003, p.176; Lohr, 2021, p. 369).

3.4 Variance smoothing Fay-Herriot model specification

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_K^2)$ denote the superpopulation means and within-stratum superpopulation variances for areas $i = 1, ..., K$. For each area i, we obtain a Hájek estimate for θ_i from expression (19), denoted $\hat{\theta}_i$, and an estimate for the design variance from expression (20), denoted $\hat{V}_i$. We define a Fay-Herriot model with variance smoothing

$$\begin{aligned} \hat{\theta}_i \mid \theta_i, \sigma_i^2 &\sim \mathcal{N}(\theta_i, v_i(\sigma_i^2)), & \hat{V}_i \mid \sigma_i^2, \gamma &\sim v_i(\sigma_i^2)\, c_i(\gamma, \sigma_i^2)\, \chi^2_{d_i(\gamma,\sigma_i^2)}, \\ \theta_i &= \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + p_i\gamma + b_i, & \log(\sigma_i^2) &= \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\eta} + e_i, \end{aligned} \tag{32}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are area-specific auxiliary variable vectors, $(\boldsymbol{\beta}, \gamma)$ and $\boldsymbol{\eta}$ are vectors of fixed effect regression parameters, p_i is the urban population proportion, and $\boldsymbol{b}$ and $\boldsymbol{e}$ are vectors of the residual between-area variation modeled as normal area-level random effects, described in more detail below. The functions $\mathrm{v}_i(\sigma_i^2)$, $c_i(\gamma, \sigma_i^2)$ and $d_i(\gamma, \sigma_i^2)$ are defined to attain the simple (26) or SASW sampling distribution (31). For the simple sampling distribution,

$$\mathrm{v}_i(\sigma_i^2) = V_i^{\dagger}(\sigma_i^2) = \sigma_i^2/(m_{\cdot i} n_0), \tag{33}$$

$$c_i(\gamma, \sigma_i^2) = 1/(m_{\cdot i} - |\mathcal{H}_i|), \qquad d_i(\gamma, \sigma_i^2) = m_{\cdot i} - |\mathcal{H}_i|,$$

where $n_0 = (\sum_{h \in \mathcal{H}_i} \sum_{c \in S_{hi}} n_c)/m_{\cdot i}$. For the SASW sampling distribution,

$$\mathrm{v}_i(\sigma_i^2) = V_i^{\star}(\sigma_i^2) = \sigma_i^2 \frac{\boldsymbol{w_i}^{\star\mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^{\star}}{(\mathbf{1}^{\mathrm{T}} \boldsymbol{w_i}^{\star})^2}, \tag{34}$$

$$c_i(\gamma, \sigma_i^2) = \frac{Q_2(\gamma, \sigma_i^2)}{2Q_1(\gamma, \sigma_i^2) \boldsymbol{w_i}^{\star\mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^{\star}}, \qquad d_i(\gamma, \sigma_i^2) = \frac{2Q_1(\gamma, \sigma_i^2)^2}{Q_2(\gamma, \sigma_i^2)},$$

where $Q_1(\gamma, \sigma_i^2)$ and $Q_2(\gamma, \sigma_i^2)$ are as defined in (31).

The area-level random effects, $\boldsymbol{b}$ and $\boldsymbol{e}$, can either be IID normal with precision hyperparameters τ_b and τ_e, respectively, or the BYM2 spatial model with hyperparameters (τ_b, ϕ_b) and (τ_e, ϕ_e), respectively. The BYM2 model, introduced by Riebler et al. (2016),

is a re-parameterized version of the Besag-York-Mollié (BYM) model (Besag et al., 1991), which includes both unstructured IID normal random effects and structured, intrinsically conditional autoregressive (ICAR), spatial random effects (Besag et al., 1991). The BYM2 model has two parameters: τ, indicating the overall precision of the random effect, and ϕ, indicating the proportion of variation that is explained by the structured component. The structured component has a sum-to-zero constraint to ensure identifiability and is scaled such that the marginal standard deviation of the random effect is approximately $1/\sqrt{\tau}$ (Riebler et al., 2016).

For $(\boldsymbol{\eta}, \boldsymbol{\beta}, \gamma)$ we use normal priors with fixed mean and variance. For the hyperparameters of $\boldsymbol{b}$ and $\boldsymbol{e}$, we use penalized complexity (PC) priors for τ_b and τ_e with hyperparameters u and α, which correspond to the prior belief that $P(1/\sqrt{\tau_b} > u) = \alpha$ (Simpson et al., 2017), and for ϕ_b and ϕ_e we use Beta priors, which encourage shrinkage towards the simpler, non-spatially structured model.

3.5 Simulation study

In this study we seek to compare the simple and SASW sampling distributions to the empirical distribution of the design variance estimator and assess the importance of accounting for the uncertainty of the design variance estimator by using each of these sampling distributions. We compare the performance of Fay-Herriot models with variance smoothing using the simple and SASW sampling distributions to the performance of the standard Fay-Herriot model with no variance smoothing. We also compare the performance of these variance smoothing models when informative covariates for the variance are unavailable and the variance is not spatially structured. The focus of our study is estimating means and their uncertainty at the admin2 level, which are unplanned domains.

3.5.1 Simulation design

To mimic our motivating context we use the geography of Kenya, which has 47 admin1 areas and $K = 300$ admin2 areas and choose simulation settings which mimic the sampling frame and design of the 2022 Kenya DHS (ICF International, 2022). For the sampling frame, we use the total number of clusters in each stratum h, M_h, provided in the 2022 Kenya DHS report. The number of clusters in the sampling frame of each admin2 area is not reported, so we group clusters within each stratum into their nested admin2 areas so that the number of clusters is proportional to its total population, for which we use WorldPop estimates (Tatem, 2017). Each admin2 area has the under-5 population estimated by WorldPop and each individual is assigned to a cluster in its respective area with equal probability.

The survey uses a stratified two-stage cluster design with $H = 92$ strata: an urban and rural stratum for each admin1 area, except for two areas which are solely urban. Let $\mathcal{H}^{\mathrm{U}} \subset \mathcal{H}$ and $\mathcal{H}^{\mathrm{R}} \subset \mathcal{H}$ denote the subsets of urban and rural strata, respectively. For each stratum h, we draw a fixed number of clusters, m_h, with PPS sampling, which is summarised in Figure 3.1 (except for one simulation setting, noted below, in which we increase the number of sampled clusters). We assume the listed size of each cluster in the sampling frame is correct (i.e., $L_c = N_c$) and, therefore, we do not require re-enumeration at the second stage. See Appendix B5.1 for an additional simulation setting which considers re-enumeration of cluster sizes.

While the DHS samples a fixed number of households from each cluster, we directly sample individuals from each cluster, with the distribution of sample sizes across clusters being chosen to closely emulate the sample sizes in the 2022 Kenya DHS. For each urban cluster, the number of sampled individuals, n_c, is drawn from a negative binomial distribution with size, 8, and mean, 9. For each rural cluster, the number of sampled individuals, n_c, is drawn from a negative binomial distribution with size, 4, and mean, 11. See Appendix B6 for more details. The sampling weight for an individuals in cluster $c \in S_h$ is the inverse of their sampling probability so, according to this design, $w_c = (\sum_{c' \in U_h} N_{c'})/(m_h n_c)$, where

$\sum_{c' \in U_h} N_{c'}$ is the population size in stratum h. Note that this simulation design breaks the assumptions on the simple sampling distribution that the domain is planned and all strata and clusters have equal sample size and sampling weights.

We define the superpopulation mean and within-stratum superpopulation variance for area i in stratum h as

$$\theta_{hi} = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \gamma_i \mathbb{I}\left(h \in \mathcal{H}^U\right) + b_i, \quad \log(\sigma_{hi}^2) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\eta} + \kappa_i \mathbb{I}\left(h \in \mathcal{H}^U\right) + e_i, \tag{35}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are each vectors of length 4 with the first element equal to 1 and the remaining elements having values generated from a standard normal distribution. We choose $\boldsymbol{\beta} = (-1, -0.15, -0.1, 0.25)^{\mathrm{T}}$, $\boldsymbol{\eta} = (0.5, -0.15, -0.1, 0.25)^{\mathrm{T}}$, and $\boldsymbol{\gamma}$ and $\boldsymbol{\kappa}$ are area-specific urban effects which vary across settings. Note that the simulated covariates are different for the mean and variance latent models. The between-area variation is quantified by area-level random effects, $\boldsymbol{b} = (b_1, ..., b_K)$ and $\boldsymbol{e} = (e_1, ..., e_K)$ which are generated from independent BYM2 models with parameters $(1/\sqrt{\tau_b} \approx 0.32, \phi_b = 0.75)$ and $(1/\sqrt{\tau_e} \approx 0.22, \phi_e = 0.75)$. Given these choices of $(\boldsymbol{\beta}, \boldsymbol{\eta}, \tau_b, \tau_e, \phi_b, \phi_e)$, approximately 50% of the between-area variability in the mean and within-stratum variance can be explained through the covariates, and 75% of the remaining 50% can be explained through the spatial structure.

Five simulation settings, summarized in Table 3.1, are used to assess how the models perform under correct specification, increased sample size, and misspecification of the latent models and population distribution. In setting 1, we simulate data to be correctly specified under superpopulation assumption (23) and the model defined in (32). Specifically, we choose $\kappa_i = 0$ and $\gamma_i = 1$ for each area i, which restricts the model to have equal within-stratum superpopulation variance within each area ($\sigma_{hi}^2 = \sigma_i^2$), and equal difference between stratum means among all areas ($\gamma_i = \gamma$). Then for each individual ℓ in cluster $c \in U_{hi}$, we generate $y_{c\ell} \sim \mathcal{N}(\theta_{hi}, \sigma_{hi}^2)$. In setting 1a we generate outcomes in the same manner, but sample five times as many clusters as in the 2022 Kenya DHS.

Table 3.1. Summary of simulation settings, where θ_{hi} and σ^2_{hi} are as defined in (35)

Setting	Description	Urban effects	Superpopulation distribution
1	correctly specified	$\gamma_i = 1; \kappa_i = 0$	$\mathcal{N}(\theta_{hi}, \sigma^2_{hi})$
1a	correctly specified with 5 times more sampled clusters	$\gamma_i = 1; \kappa_i = 0$	$\mathcal{N}(\theta_{hi}, \sigma^2_{hi})$
2	varying urban effects	$\gamma_i \sim \mathcal{N}(1, 1^2)$ $\kappa_i \sim \mathcal{N}(\log(1.5), 0.25^2)$	$\mathcal{N}(\theta_{hi}, \sigma^2_{hi})$
3	t-distributed data	$\gamma_i = 1; \kappa_i = 0$	t-distributed with mean θ_{hi}, variance σ^2_{hi}, and df= 5
4	within-cluster correlation	$\gamma_i = 1; \kappa_i = 0$	normal with mean θ_{hi}, variance σ^2_{hi}, and within-cluster correlation, $\rho = 0.25$

In setting 2, we test misspecification of the latent mean and variance models. We generate outcomes in the same manner as in setting 1, with the only difference being that we draw $\gamma_i \sim \mathcal{N}(1, 1^2)$ and $\kappa_i \sim \mathcal{N}(\log(1.5), 0.25^2)$, which implies that the within-stratum superpopulation variance in an area is 50% higher, on average, in the urban stratum than in the rural stratum. In settings 3 and 4, we test violations of superpopulation assumption (23). Both settings use the same latent models as in setting 1. In setting 3, we relax the normality assumption and generate independent outcomes from a t-distribution with 5 degrees of freedom, shifted and scaled so that the mean and variance is θ_{hi} and σ^2_{hi}, respectively. In setting 4, we relax the assumption of independence within clusters by generating outcomes from a normal distribution with mean θ_{hi}, variance σ^2_{hi}, and a correlation of $\rho = 0.25$ within each cluster. For each of the 5 settings we generate a fixed population surface and sample $G = 100$ datasets using the stratified two-stage cluster sampling design described above. A summary of admin2 area sample sizes is provided in Table 3.2.

3.5.2 Candidate models

For each dataset g sampled from the simulated population, we use the design-based mean estimates and variance estimates, $(\hat{\boldsymbol{\theta}}^{(g)}, \hat{\mathbf{V}}^{(g)})$, to fit four types of Fay-Herriot variance smoothing models, as defined in (32). For all models, $\boldsymbol{X}$ is the same matrix as in the data generat-

Table 3.2. Summary of admin2 sample sizes across 100 simulations for each setting.

	Settings 1-4 Mean (SD)	Setting 1a Mean (SD)
Number of sampled clusters:		
Urban	2.2 (2.0)	11.2 (8.2)
Rural	3.4 (2.7)	17.0 (11.3)
Number of sampled individuals:		
Urban	20.0 (19.4)	101.0 (75.6)
Rural	37.3 (31.9)	186.3 (126.9)
Number of areas (out of 300):		
With no sampled clusters	7.7 (2.3)	0.9 (0.6)
With < 5 sampled clusters	128.0 (5.8)	4.3 (1.0)

ing process, $\boldsymbol{\beta}$ is a vector of fixed effect regression coefficients, and $\boldsymbol{b}$ are BYM2 area-level random effects.

The first type of model variation is whether the simple or SASW sampling distribution is used and the second type is whether the latent variance model is structured. For the structured latent variance model, $\boldsymbol{Z}$ is the same matrix as in the data generating process, $\boldsymbol{\eta}$ is a vector of fixed effect regression coefficients, and $\boldsymbol{e}$ are BYM2 area-level random effects. For the unstructured latent variance model, $\boldsymbol{Z}$ is a vector of 1s, η is an intercept term, and $\boldsymbol{e}$ is an IID normal area-level random effect. Crossing these two types of model variations yields four models which we will call 'Simple-struct', 'Simple-unstruct', 'SASW-struct', and 'SASW-unstruct'. Note that whether the model is labeled structured or unstructured refers only to the latent variance model, as the latent mean model is structured in all variations. We choose these four models to not only compare the performance of the simple and SASW sampling distributions, but to observe how these results may change if information regarding the structure of the variance is not available, as is often the case in applied settings.

For comparison, we also fit two Fay-Herriot models which do not use variance smoothing. The first model, which we refer to as the 'standard' model, is specified in (21). The second model, which we refer to as the 'oracle' model, uses the sampling distribution,

$$\hat{\theta}_i^{(g)} \mid \theta_i \sim \mathcal{N}\left(\theta_i, V_i\right),$$

where V_i, the empirical design variance, is

$$V_i = \frac{1}{|G_i|} \sum_{g \in G_i} \left(\hat{\theta}_i^{(g)} - \frac{1}{|G_i|} \sum_{g' \in G_i} \hat{\theta}_i^{(g')} \right)^2 \tag{36}$$

and G_i is the subset of dataset indices for which area i has valid design-based mean and variance estimates ($m_{\cdot i} > 1$). We include this model to compare performance when the design variance is correctly estimated without smoothing, which cannot be reliably done in a real data setting. Both comparison models use the same latent mean model as the four variance smoothing models.

The same priors on regression parameters and hyperparameters are used for all models. For the intercept of $\log(\sigma_i^2)$, η_1, we choose a prior of $\mathcal{N}(0.5, 0.5^2)$, and for all other regression parameters we choose a prior of $\mathcal{N}(0, 2^2)$. For the PC priors on τ_b and τ_e we choose $u = 1$ and $\alpha = 0.01$. For ϕ_b and ϕ_e we choose a shrinkage prior of Beta$(0.5, 1)$ which has mean, 0.33, and standard deviation, 0.3.

We use the `Stan` software (Stan Development Team, 2024) for all models, and estimate the posterior distribution using 4 chains of 1000 iterations after 1000 iterations of burn-in. `Stan` diagnostics indicate convergence and efficient sampling for all models and simulations. For settings 1-4, the average run times across 100 simulations for the 'Simple-struct', 'Simple-unstruct', 'SASW-struct', and 'SASW-unstruct' models were 26, 20, 63, and 51 seconds, respectively. Due to increased sample size in setting 1a, models had longer run times: 'Simple-struct', 'Simple-unstruct', 'SASW-struct', and 'SASW-unstruct' models had average run times of 34, 27, 184, and 184 seconds, respectively.

3.5.3 Results

Assessment of simple and SASW sampling distributions We first evaluate how well the theoretical design variance under each model, $V_i^\dagger(\sigma_i^2)$ for the simple sampling distribution, defined in (33), and $V_i^\star(\sigma^2)$ for the SASW sampling distribution, defined in (34), approximate

the true design variance, V_i, obtained empirically through simulation, defined in (36). We assume that the Monte Carlo error in the empirical estimate of V_i is negligible. We do not use model estimates of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\gamma}$, but plug in the correct values, which are defined in the simulation design, so that we may assess accuracy of these sampling distributions when parameters are correctly specified. We will address parameter estimation under each model in the next subsection.

For each area i, we evaluate the average theoretical-to-truth design variance ratios,

$$\frac{1}{|G_i|}\sum_{g\in G_i}\frac{V_i^{(g)\dagger}(\sigma_i^2)}{V_i} \quad \text{and} \quad \frac{1}{|G_i|}\sum_{g\in G_i}\frac{V_i^{(g)\star}(\sigma_i^2)}{V_i}, \tag{37}$$

for the simple and SASW sampling distributions, respectively. In Figure 3.3 we observe that, on average, the theoretical design variance under each of the models underestimates the design variance, conditional on correct specification of the within-stratum superpopulation variance. This underestimation of the variance is a result of conditioning on the sampling weights and sample sizes of the sampled clusters which are, in fact, random. The simple sampling distribution underestimates the design variance by a larger magnitude because the additional design assumptions, as noted in Section 3.3.1, result in underestimation of the variability imposed by the complex design. Thus we observe that, conditional on the within-stratum superpopulation variance being correctly specified, both models produce estimates of the design variance which shrink towards 0, with more shrinkage under the simple sampling distribution than the SASW sampling distribution.

In Appendix B5.2 we compare the simple and SASW sampling distributions to the empirical sampling distribution of the design variance estimator and observe that the SASW sampling distribution is a slightly better approximation, while the simple sampling distribution tends to underestimate the mean of the design variance estimator.

Variance estimation Here we evaluate how well the within-stratum superpopulation variance, σ_i^2, is estimated for each area i under each variance smoothing model, and how this

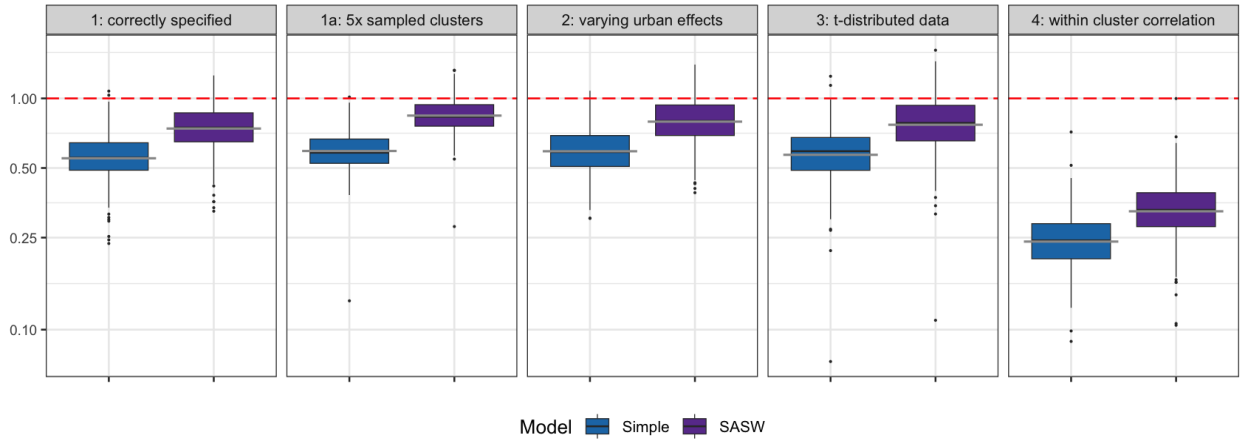


Figure 3.3. Average theoretical-to-truth design variance ratio (37) for each model and area, across 100 simulations for each of 5 settings. The gray lines indicate the average ratio across areas and the red line indicates the point where the theoretical variance is equal to the true variance. Values below the red line indicate underestimation.

propagates to estimation of the design variance, V_i. For each area i, we evaluate the average estimate-to-truth ratio for the within-stratum superpopulation variance,

$$\frac{1}{G}\sum_{g=1}^{G}\frac{\widetilde{\sigma}_i^{2(g)}}{\sigma_i^2} \tag{38}$$

for each of the variance smoothing models, where $\widetilde{\sigma}_i^{2(g)}$ denotes the posterior mean estimate for σ_i^2 using sample dataset g. For each area i we also evaluate the average estimate-to-truth ratio for the design variance

$$\frac{1}{|G_i|}\sum_{g\in G_i}\frac{\mathrm{v}_i^{(g)}(\widetilde{\sigma}_i^{2(g)})}{V_i} \tag{39}$$

for each model, where $\mathrm{v}_i^{(g)}(\sigma^2) = V_i^{\dagger(g)}(\sigma^2)$ for the simple variance smoothing models, $\mathrm{v}_i^{(g)}(\sigma^2) = V_i^{\star(g)}(\sigma^2)$ for the SASW variance smoothing models, and $\mathrm{v}_i^{(g)}(\sigma^2) = \hat{V}_i^{(g)}$ for the standard model. Note that these ratio measures differ from (37) because we plug in the estimate of σ_i^2 under each model.

In Figure 3.4 we present the average estimate-to-truth ratios for the within-stratum superpopulation variance and design variance, for each area i, model, and setting, across $G = 100$ sampled datasets. In Figure 3.4A we observe that the within-stratum superpop-

ulation variance, σ_i^2, is overestimated, on average, for all models. This is probably due to the sparsity and subsequent noise in the data, as we see the overestimation decreases as sample size increases. In Figure 3.4B, we observe that the design variance, V_i, is also overestimated, on average, for all smoothing models, but to a lesser degree than the within-stratum superpopulation variance. We find that the shrunken theoretical design variance under each model, observed in Figure 3.3, counteracts the inflated within-stratum superpopulation variance estimates. Moreover, the simple sampling distribution counteracts this inflated within-stratum superpopulation variance more strongly because its theoretical design variance underestimates the design variance by a larger magnitude. As sample size increases, inflation of the within-stratum superpopulation variance and underestimation of the theoretical design variance both decrease, resulting in more accurate design variance estimates. In the presence of within-cluster correlation, estimation of σ_i^2 is poor because, as previously stated, it is misspecified, but the estimation of design variance is comparable to the other simulation settings because the correlation is absorbed into the estimation of σ_i^2. As expected, the design variance estimates without smoothing underestimate the true design variance, except for when sample size is increased.

Mean estimation Now we compare the performance of the point estimates and credible intervals for the area-level means under each model using four metrics. Let $\tilde{\theta}_i^{(g)}$ denote the posterior mean estimate for θ_i, for area i, using sample dataset g, and $\left(l_i^{(g)}, u_i^{(g)}\right)$ denote its corresponding 90% credible interval. First, we evaluate the root mean squared error (RMSE) for each area i,

$$\sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\tilde{\theta}_i^{(g)} - \theta_i\right)^2}.$$

Next, we evaluate the coverage and average width of the 90% credible intervals for each area i,

$$\frac{1}{G}\sum_{g=1}^{G} I\left(\theta_i \in \left[l_i^{(g)}, u_i^{(g)}\right]\right), \qquad \frac{1}{G}\sum_{g=1}^{G}\left(u_i^{(g)} - l_i^{(g)}\right).$$

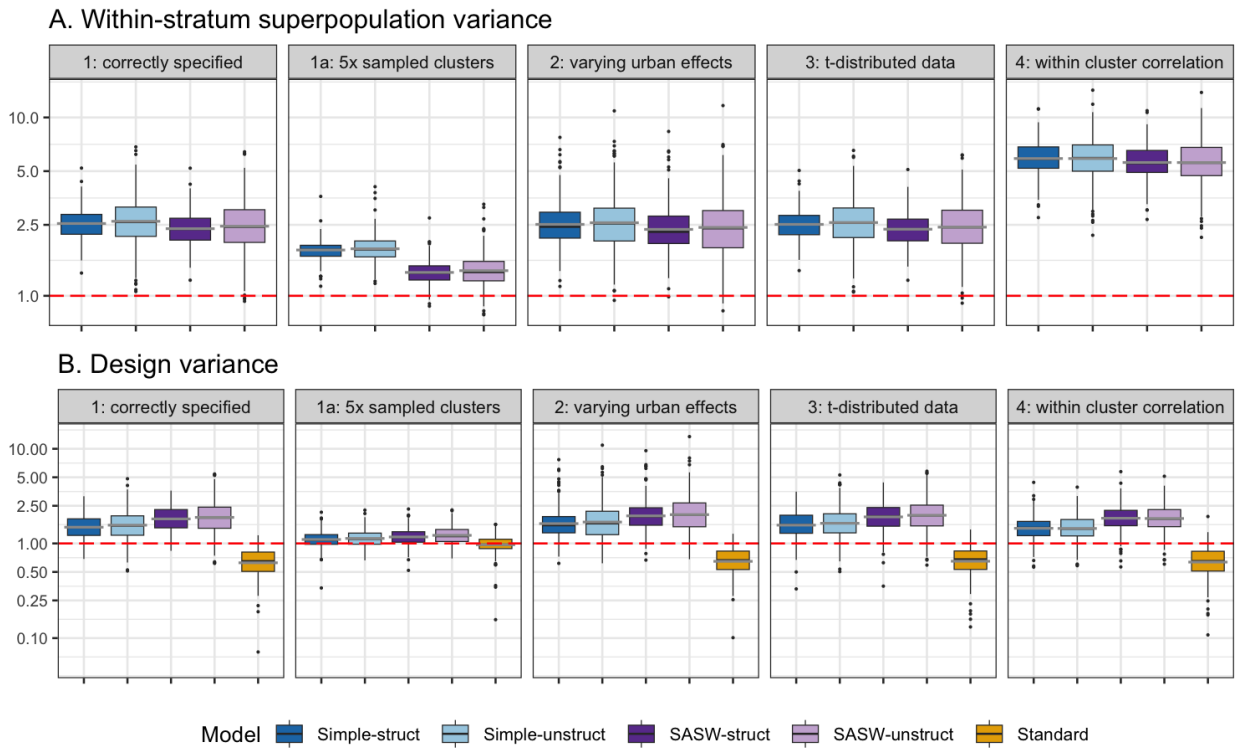


Figure 3.4. Estimate-to-truth ratios for within-stratum superpopulation variance (38) and design variance (39) for each area i and model across 100 simulations, for each of 5 settings. The gray lines indicate the average ratio across areas, and the red line indicates the point where the estimate is equal to the truth. Values above the red line indicate overestimation.

Finally, we evaluate the average interval score, a proper scoring rule discussed in Gneiting and Raftery (2007), which favors narrow intervals containing the true parameter value. The average interval score for 90% credible intervals for area i is

$$\frac{1}{G}\sum_{g=1}^{G}\left(u_i^{(g)} - l_i^{(g)} + \frac{2}{\alpha}\max\left\{0, l_i^{(g)} - \theta_i\right\} + \frac{2}{\alpha}\max\left\{0, \theta_i - u_i^{(g)}\right\}\right),$$

where we take $\alpha = 0.1$. In Figure 3.5 we present the RMSE and the coverage, average width, and average interval score of 90% credible intervals, per area, for each model, across $G = 100$ sampled datasets for each setting.

In Figure 3.5A we observe that all variance smoothing models have comparable RMSE to the oracle model, on average, which is lower than that of the standard model, except for when sample size is increased, at which point the standard model performs comparably. In

Figures 3.5B and 3.5C we observe that, on average, all variance smoothing models produce more conservative credible intervals and the standard model produces credible intervals with undercoverage. This follows directly from the findings in Section 3.5.3 that the design variance is overestimated, on average, for all variance smoothing models, and underestimated, on average, for the standard model. The simple variance smoothing models produce intervals which are slightly less conservative as a result of its theoretical design variance underestimating the true design variance by a larger magnitude. We observe that when sample size increases, all credible intervals are closer to the nominal rate, on average. In Figure 3.5D, we observe that the variance smoothing models produce credible intervals which are more favorable than those under the standard model, with the simple variance smoothing models being slightly preferred over the SASW smoothing models.

Summary of findings From this simulation study, we conclude that, under the tested settings, the standard model without variance smoothing underestimates uncertainty of the area-level means leading to undercoverage, so using either variance smoothing model is preferred over the standard model. While within-cluster correlation yields poor estimation of the within-stratum superpopulation variance due to misspecification, the design variance is well estimated, yielding point and interval estimates which have comparable performance to those when the model is correctly specified. Moreover, the simple variance smoothing model is slightly preferred over the SASW variance smoothing model because its theoretical design variance has stronger shrinkage towards 0, which is beneficial in counteracting the inflated estimation of the within-stratum superpopulation variance, intrinsic to cases of sparse data. We observe that using a structured latent variance model has very little influence on model performance; the choice of sampling distribution for the design variance estimator is more influential.

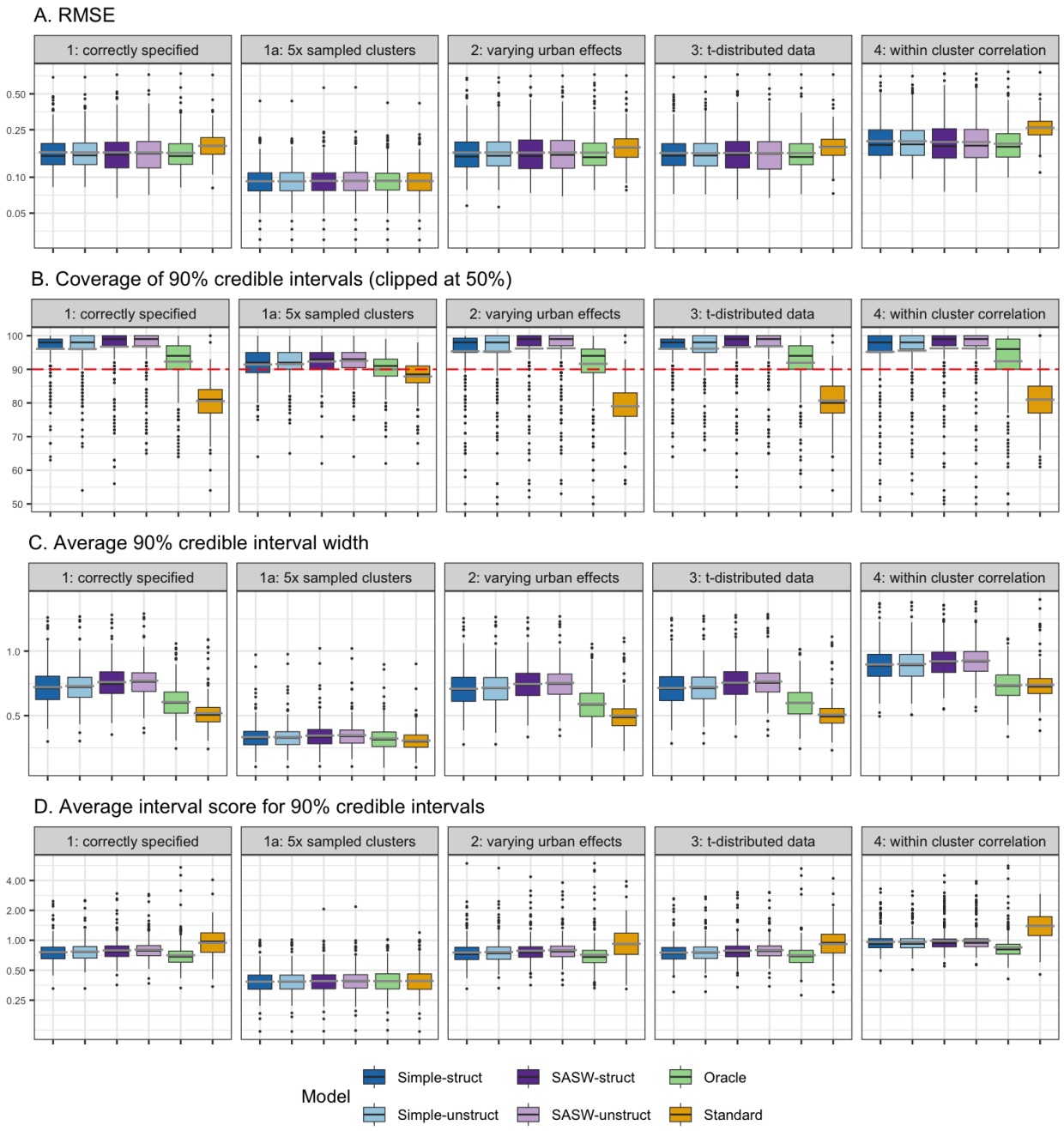


Figure 3.5. RMSE of area-level point estimates and coverage, average width, and average interval score of 90% credible intervals, for each area and model, across 100 simulations, for each of 5 settings. The gray lines indicate the mean across areas.

3.6 Data example: height-for-age z-score in Kenya

We return to the example of HAZ at the admin2 level in the 2022 Kenya DHS, introduced in Section 3.1. We obtain design-based estimates, $(\hat{\theta}_i, \hat{V}_i)$, for each admin2 area i that has

at least two sampled clusters (288 out of 300 areas). We use $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{V}})$ to fit two variance smoothing Fay-Herriot models, as defined in Section 3.4,

$$\hat{\theta}_i \mid \theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, v_i(\sigma_i^2)), \qquad \hat{V}_i \mid \sigma_i^2, \gamma \sim v_i(\sigma_i^2)\, c_i(\gamma, \sigma_i^2)\, \chi^2_{d_i(\gamma,\sigma_i^2)},$$
$$\theta_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} + b_i, \qquad \log(\sigma_i^2) = \eta + e_i,$$

where $\boldsymbol{X}_i$ is an area-specific auxiliary variable vector further defined below, $\boldsymbol{\beta}$ is a vector of fixed effect regression coefficients, η is an intercept term, and $\boldsymbol{b}$ and $\boldsymbol{e}$ are vectors of the residual between-area variation, modeled as area-level BYM2 random effects. The same priors and hyperpriors are used as in the simulation study. The functions $\mathrm{v}_i(\sigma_i^2)$, $c_i(\gamma, \sigma_i^2)$ and $d_i(\gamma, \sigma_i^2)$, defined in (33) and (34), are chosen to attain the simple (26) or SASW (31) sampling distributions. For comparison, we also fit a standard Fay-Herriot model as defined in (21), where $\boldsymbol{X}_i$, $\boldsymbol{\beta}$, and $\boldsymbol{b}$ retain the same definitions as in the variance smoothing models.

For the area-specific auxiliary variables in the latent mean model we use the urban under-5 population proportion, population density, yearly average temperature, yearly total precipitation, elevation, motorized travel time to healthcare, and yearly average nighttime lights. See Appendix B7 for more details. We do not use auxiliary variables for the latent variance model because, in practice, a principled approach for choosing covariates for the latent variance is less straightforward than for the latent mean, and in the simulation study we found their inclusion was not influential on model performance, even when informative covariates are chosen. We use the `Stan` software (Stan Development Team, 2024) for all models, and estimate the posterior distribution using 4 chains of 1000 iterations after 1000 iterations of burn-in. The run times for the standard, simple variance smoothing, and SASW variance smoothing models were 26, 29, and 69 seconds, respectively. `Stan` diagnostics indicate convergence and efficient sampling for all models.

Figure 3.6 presents the mean estimates and width of 90% credible intervals under the standard model and each of the variance smoothing models. We observe that the mean estimates are fairly similar across models, with slightly more shrinkage towards the average in

the variance smoothing models. The credible interval widths among the variance smoothing models are slightly higher on average than for the standard model, but there are several areas for which the standard model yields higher uncertainty than the variance smoothing models.

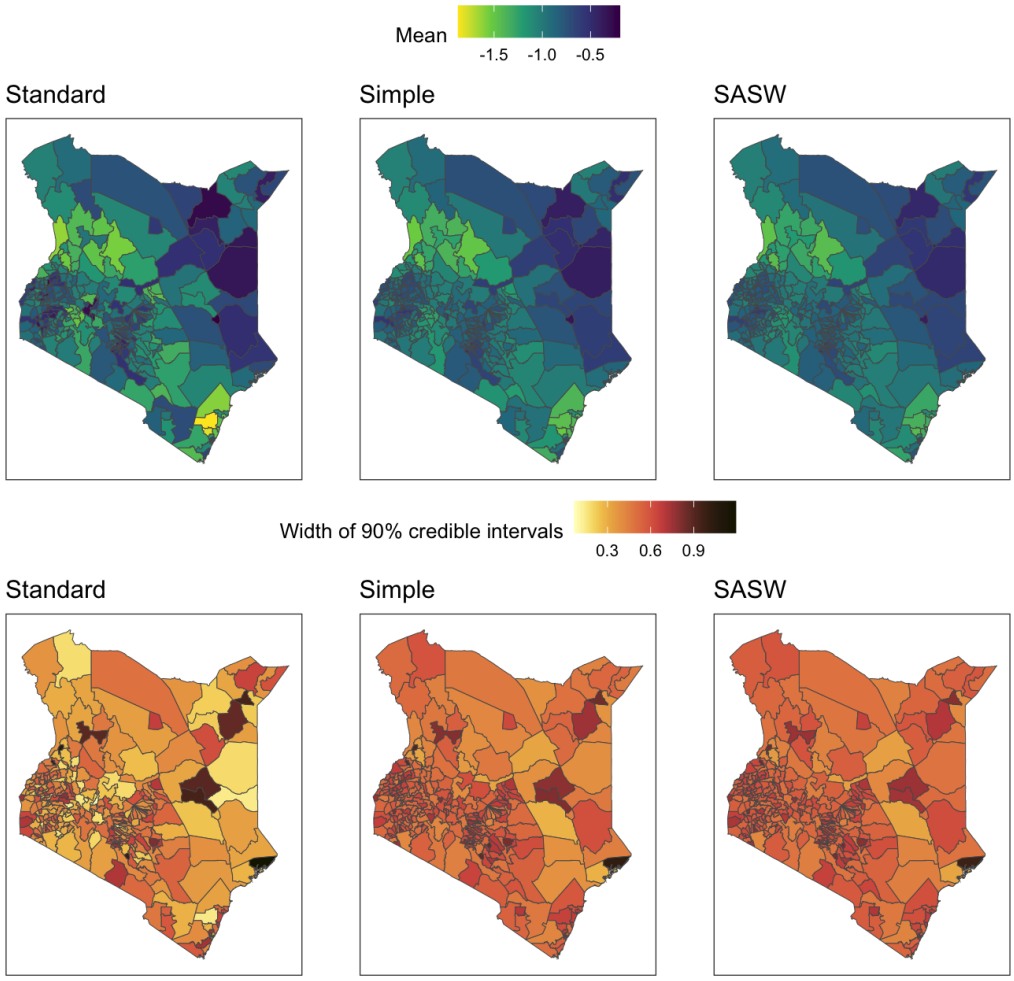


Figure 3.6. Mean estimates and width of 90% credible intervals for HAZ using 2022 Kenya DHS, under a standard Fay-Herriot model and two types of variance smoothing Fay-Herriot models.

Figure 3.7 more clearly illustrates the effect of the variance smoothing models on the credible intervals, where admin2 areas that have extremely narrow or wide intervals under the standard model are indicated in purple and orange, respectively. We observe that the variance smoothing models produce more reasonable uncertainty estimates. As observed in the simulation study, the simple and SASW variance smoothing models perform similarly

and, on average, produce more conservative intervals than the standard model. In Appendix B8 we use scatter plots to compare the mean and standard deviation estimates under each model to the design-based estimates.

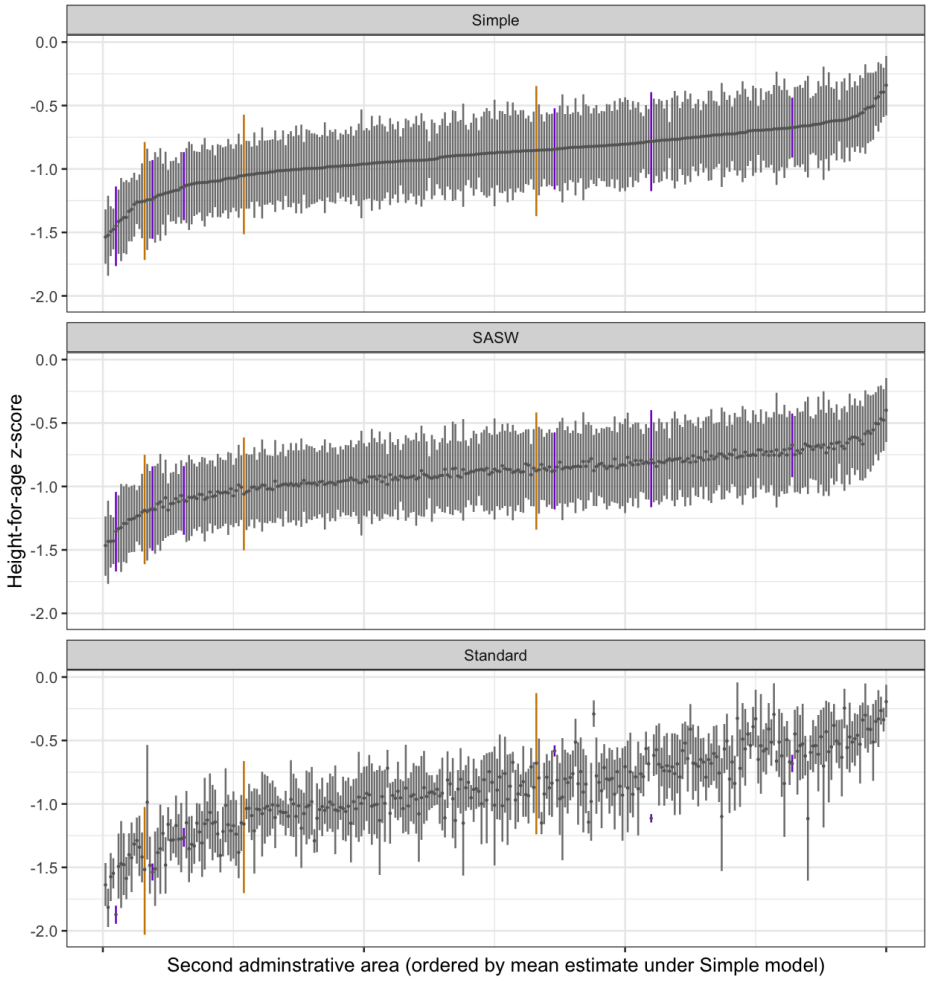


Figure 3.7. Mean estimates with 90% credible intervals for HAZ using 2022 Kenya DHS, under a standard Fay-Herriot model and two types of variance smoothing Fay-Herriot models. admin2 areas which have extremely narrow or wide credible intervals under the standard model are indicated in purple and orange, respectively.

In Figure 3.8 we visualize the uncertainty of estimates under each model by mapping the posterior probability that a given area has a HAZ which is in the lowest 10% and 25% of areas. The posterior probability that area i is in the lowest p% of areas is calculated by,

$$\frac{1}{4000}\sum_{s=1}^{4000}\mathbb{I}\left\{\frac{1}{K}\sum_{j=1}^{K}\mathbb{I}\left(\tilde{\theta}_i^{(s)} > \tilde{\theta}_j^{(s)}\right) \leq p\right\}$$

where $\tilde{\theta}_i^{(s)}$ and $\tilde{\theta}_j^{(s)}$ are the s^{th} posterior draws for θ_i and θ_j, respectively. We observe that the variance smoothing models produce very similar results to each other and, although uncertainty is higher on average for these models, the posterior probabilities are still quite informative in comparison to that of the standard model.

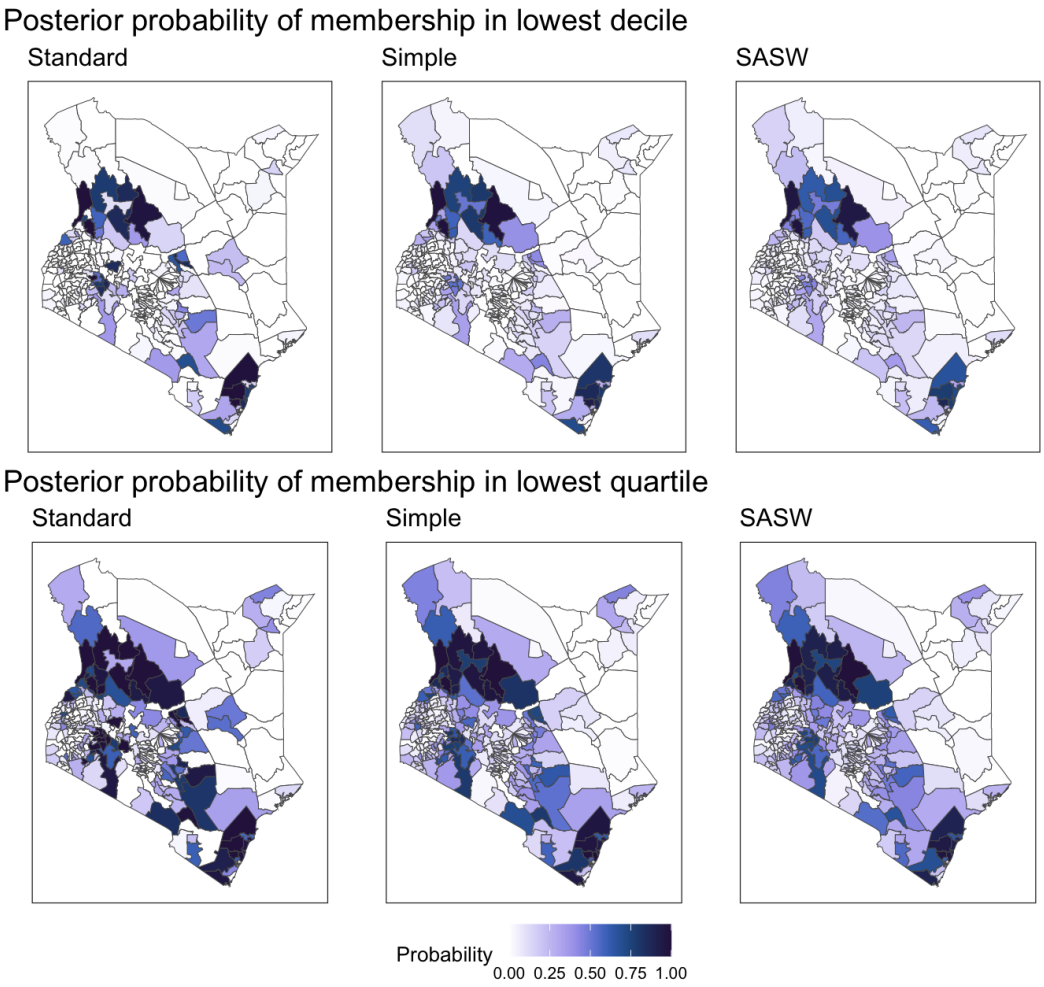


Figure 3.8. Posterior probabilities of membership in lowest decile and quartile for HAZ using 2022 Kenya DHS, under a standard Fay-Herriot model and two types of variance smoothing Fay-Herriot models.

3.7 Discussion

We began by establishing the limitations of assuming the design variance is known in a Fay-Herriot model, particularly when data are sparse, and motivating the importance of jointly modeling the design variance along with the mean. While joint smoothing Fay-Herriot

models have been employed in previous literature, their choices of sampling distribution make strong implicit design and population assumptions. This work thoroughly considers the choice of sampling distribution for the variance estimator in a complex survey setting and the assumptions made on the superpopulation and survey design.

We considered two sampling distributions for the design variance estimator under superpopulation assumption (23), one which arises from the common recommendation of using df = #clusters − #strata, and another which makes less strict assumptions on the survey design by allowing for unplanned domains and varying sampling weights and sample sizes. We illustrated through a simulation study that while the latter is, in fact, a closer approximation of the empirical distribution of the design variance estimator, using the simple sampling distribution in the Fay-Herriot model does not hinder model performance. In fact, the additional assumptions of the simple sampling distribution results in more shrinkage of the design variance towards 0 which can be helpful in counteracting the observational noise inherent to sparse data. We also found that while the SASW sampling distribution requires independence of outcomes within clusters, variance smoothing models using both sampling distributions are robust to violations of this assumption. In our simulation study we verified that the standard Fay-Herriot model can produce intervals with undercoverage in cases of sparse data and, although variance smoothing models produce more conservative intervals, they are preferred over those of the standard model, according to proper scoring rules.

We applied these models to the estimation of HAZ using the 2022 Kenya DHS and showed that, although the variance smoothing models produce more conservative intervals than the standard model, they can still provide informative posterior probabilities, which is particularly useful in policy and evaluation settings. From these results we conclude that both variance smoothing models are preferred over the absence of variance smoothing, and the simple variance smoothing model is a suitable choice because of its shrinkage properties and more straightforward implementation.

While this work is motivated by data collected using a two-stage stratified cluster sam-

pling design, the derivations and results hold generally for survey designs with cluster sampling where each individual in the cluster has the same sampling probability. These models could be extended for multi-stage sampling designs in which individuals in a cluster have different sampling probabilities, but further study is required to assess model performance in this setting. To extend these variance smoothing models for estimation of area-level proportions, we would need to assume that the distribution of the cluster-level proportions can be approximated by a normal distribution. If the conditions for approximating the binomial distribution of the cluster-level proportions with a normal distribution are met (i.e. non-rare events and sufficient sample size) these variance smoothing models may be a viable option. In the DHS setting, cluster sample sizes are generally too small for normality of proportions to be a reasonable assumption, however further study is required to determine how robust the model would be to violations of this assumption.

Chapter 4

MULTIVARIATE AREAL SPATIAL MODELING USING COMPLEX SURVEY DATA

4.1 Introduction

There is a rich literature on multivariate areal spatial models in disease mapping, however, there has been much less development for such models in the context of SAE. SAE is a framework of statistical methods which use survey data to estimate outcome measures for small subregions. As described in Chapter 1, surveys such as the Demographic Health Surveys (DHS; Corsi et al., 2012) and Multiple Indicator Cluster Surveys (MICS; UNICEF, 2026) are key sources of subnational health monitoring in LMICs for local and federal governments, as well as international organizations such as the World Health Organization (WHO). Multivariate modeling can be especially useful when estimating survey outcomes with high sampling variance, which is particularly common when an outcome is a rare event or has a high rate of missingness. In these cases, jointly modeling the outcome of interest with auxiliary outcomes allows us to leverage shared spatial variation to improve estimation of the primary outcome.

The most commonly used area-level model for SAE is the Fay-Herriot (FH) model (Fay and Herriot, 1979), which accounts for the survey design by using survey-weighted area-level estimates as model input, rather than the raw data. Suppose we have m areas and are interested in estimating outcome measure η_i, for areas $i = 1, ..., m$. We obtain design-based estimates, $(\hat{\eta}_1, ..., \hat{\eta}_m)$, by calculating survey-weighted means of the outcome observations. Let $\theta_i = g(\eta_i)$ and $\hat{\theta}_i = g(\hat{\eta}_i)$ where g is some link function which converts the outcome measure to the real line. The univariate FH model links a latent model for $(\theta_1, ..., \theta_m)$ with the observation model, $\hat{\theta}_i \mid \theta_i \sim \mathcal{N}(\theta_i, \hat{v}_i)$, for each area i, where $\hat{v}_i$ is a design-based estimate of the sampling variance of $\hat{\theta}_i$, $v_i = \text{Var}\left(\hat{\theta}_i \mid \theta_i\right)$. In practice, $\hat{v}_i$ is usually estimated from the survey data and treated as fixed in the observation model (Fay and Herriot, 1979;

Pfeffermann, 2013; Rao and Molina, 2015; You and Hidiroglou, 2023; McGovern et al., 2026a). Analogously, the multivariate FH model, which jointly models R outcome measures from the same survey, links a latent model for $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m)$ with the observation model, $\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \sim \mathcal{N}_R\left(\boldsymbol{\theta}_i, \hat{\boldsymbol{V}}_i\right)$, for each area i, where $\hat{\boldsymbol{V}}_i$ is a design-based estimate of the sampling covariance matrix of $\hat{\boldsymbol{\theta}}_i$. As in the univariate case, this sampling covariance matrix is usually estimated from the data and assumed to be known in the multivariate FH model (Datta et al., 1998; Huang and Bell, 2004; Erciulescu and Opsomer, 2022; Porter et al., 2015). This matrix is generally non-diagonal and encodes the correlated sampling error between outcome measures. Using a fixed covariance matrix in the observation model raises distinct considerations beyond what has been studied in the disease mapping context. In particular, we would like to assess how different choices of latent models interact with this observation model to affect model performance.

Datta et al. (1998) introduced the most commonly used multivariate FH model (Datta et al., 1999; Huang and Bell, 2004; Arima et al., 2017; Erciulescu and Opsomer, 2022) with a latent model consisting of independent and identically distributed (IID) random area-level effects with between-outcome correlation within each area. This formulation models common spatial variation between outcomes within each area, but does not leverage spatial dependence between areas. The absence of spatial modeling is particularly detrimental in the context of SAE because when data are sparse and high-quality covariates are not always available, leveraging spatial structure can be especially useful in obtaining more reliable and precise estimates.

In the disease mapping literature, several areal spatial models have been extended for multivariate modeling by including correlation structure, including the multivariate conditional autoregressive (MCAR) model (Gelfand and Vounatsou, 2003) and the multivariate Besag-York-Mollie (MBYM) model (MacNab, 2011). With the exception of Porter et al. (2015), which uses the MCAR (Jin et al., 2005), these correlated spatial models have been seldom studied in the context of multivariate FH models. Another way to specify the latent

model arises from the shared component model (SCM) framework in the disease mapping literature. Introduced by Knorr-Held and Best (2001), the SCM follows the paradigm that correlation between outcomes is driven by underlying, unmeasured confounding. The random effects for each outcome are modeled as a linear combination of underlying random effects which are shared across outcomes. SCMs have also rarely been used in FH models, with the exception of Schumacher and Wakefield (2025) which incorporates SCMs in a bivariate FH model.

While correlated areal spatial models and SCMs have been compared in the context of disease mapping (MacNab, 2010; Martinez-Beneito, 2013; Retegui et al., 2023), their theoretical properties and performance when incorporated into a multivariate FH model has not been thoroughly studied. There is a growing literature within the multivariate disease mapping context regarding optimizing computation for high-dimensional data (Botella-Rocamora et al., 2015; Vicente et al., 2023), however, this work is of less practical use in the SAE context because of the sparsity of complex survey data. Conversely, the study of multivariate FH models concerns how to improve estimation within the confines of data sparsity and survey design.

This chapter makes two primary contributions at the intersection of spatial modeling and SAE. First, we adapt multivariate areal spatial models from the disease mapping literature to define a class of generalized multivariate BYM2 (GMBYM2) spatial models which encompass both correlated BYM2 models and SCMs with BYM2 components. The univariate BYM2 model, introduced by Riebler et al. (2016), is a re-parameterized version of the Besag-York-Mollié (BYM) model (Besag et al., 1991), which includes both unstructured IID normal random effects and structured, intrinsically conditional autoregressive (ICAR) spatial random effects. Second, we study theoretical properties of the multivariate FH model under a Bayesian framework, with a particular focus on how the fixed sampling covariance matrix in the observation model interacts with the covariance of the latent random effects distribution. We integrate four GMBYM2 models into a multivariate FH model framework

and evaluate how well each of them estimate a primary outcome, compared to a univariate FH model.

We begin in Section 4.2 with a review of the univariate FH model and apply this model to the estimation of neonatal mortality rate (NMR) in admin1 areas of Kenya using the 2022 Kenya DHS (ICF International, 2022), which demonstrates the appeal of a multivariate FH model. In Section 4.3 we review the literature on multivariate FH models, introduce the GMBYM2 model class, and propose four models from this class. In Section 4.4 we derive a general form of the conditional posterior distribution under the multivariate FH model, which allows us to investigate shrinkage properties and precision of the proposed GMBYM2-FH models as a function of variance and correlation at the observation and latent levels. In Section 4.5 we conduct a simulation study where we compare the proposed GMBYM2 FH models with a univariate FH model to validate our theoretical findings and assess model performance for a wider range of parameter settings. In Section 4.6 we return to our motiviating example and use the four proposed GMBYM2-FH models to jointly estimate neonatal mortality rate (NMR) in Kenya with two, less noisy, auxiliary outcomes. We conclude in Section 4.7 with a discussion of our findings and model selection recommendations.

4.2 Univariate Fay-Herriot model

Consider a population divided into H sampling strata, the set of which is denoted by $\mathcal{H}$. Each strata, $h \in \mathcal{H}$, is composed of M_h clusters, the set of which is denoted by U_h, and each cluster, $c \in U_h$, contains N_c individuals. Let $y_{c\ell r}$ denote the value of some outcome r for individual $\ell = 1, ..., N_c$ in cluster $c \in U_h$ and $h \in \mathcal{H}$. Suppose we are interested in the population quantity

$$\eta_{ir} = \frac{\sum_{h\in\mathcal{H}} \sum_{c\in U_{hi}} \sum_{\ell=1}^{N_c} y_{c\ell r}}{\sum_{h\in\mathcal{H}} \sum_{c\in U_{hi}} N_c}$$

for areas $i = 1, ..., m$, where $U_{hi} \subseteq U_h$ is the set of clusters in stratum h which are contained in area i.

We can then use survey data to obtain a design-based estimate for η_{ir}, denoted $\hat{\eta}_{ir}$, for

each area i, by computing the survey-weighted mean. For example, consider the stratified two-stage cluster sampling design used in the DHS and MICS, wherein the first stage, for each stratum $h \in \mathcal{H}$, m_h clusters are sampled from U_h with probability proportional to size, the set of which is denoted by S_h, and in the second stage, n_c individuals are sampled from cluster $c \in \cup_{h\in\mathcal{H}} S_h$ with equal probability. Let $S_{hi} \subseteq S_h$ be the set of sampled clusters in stratum h which are contained in area i. Then we obtain a design-based estimator (Hájek, 1971) for η_{ir},

$$\hat{\eta}_{ir} = \frac{\sum_{h\in\mathcal{H}} \sum_{c\in S_{hi}} w_c \sum_{\ell}^{N_c} \zeta_{c\ell} y_{c\ell r}}{\sum_{h\in\mathcal{H}} \sum_{c\in S_{hi}} w_c n_c}, \tag{40}$$

where w_c is the sampling weight for individuals in cluster $c \in S$ and $\zeta_{c\ell}$ denotes whether individual ℓ in cluster $c \in S$ was sampled. This estimator is design-consistent and asymptotically normal (Breidt and Opsomer, 2017) with some unknown sampling variance, $v_{\eta,ir} = \text{Var}(\hat{\eta}_{ir} \mid \eta_{ir})$, which is often estimated via Taylor linearization (Wolter, 2007, chap. 6; Korn and Graubard, 2011, chap. 2.4; Särndal et al., 2003, chap. 5.5; Lohr, 2021, chap. 9.1).

Let $\theta_{ir} = g_r(\eta_{ir})$ and $\hat{\theta}_{ir} = g_r(\hat{\eta}_{ir})$, where g_r is a link function which converts the measure for outcome r to the real line. For example, if outcome r is binary, we may take g_r to be the logit function. Then a univariate FH model (Fay and Herriot, 1979) for outcome r is

$$\hat{\theta}_{ir} \mid \theta_{ir} \sim \mathcal{N}\left(\theta_{ir}, \hat{v}_{ir}\right), \qquad \theta_{ir} = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_r + b_{ir} \qquad i = 1, ..., m \tag{41}$$

where $\boldsymbol{X}_i$ is the ith column of $\boldsymbol{X}$, a $p \times m$ matrix of area-specific auxiliary variables, $\boldsymbol{\beta}_r$ is a vector of fixed effect regression parameters, $(b_{1r}, ..., b_{mr})^{\mathrm{T}}$ is a vector of normal area-level random effects, and $\hat{v}_{ir}$ is a design-based estimate of $v_{ir} = \text{Var}(\hat{\theta}_{ir} \mid \theta_{ir})$, obtained by applying the delta method to the estimator of $v_{\eta,ir}$. The sampling variance is most commonly treated as known in the FH model (Fay and Herriot, 1979; Pfeffermann, 2013; Rao and Molina, 2015), though FH models which jointly model the mean and variance have been studied (You and Chapman, 2006; Sugasawa et al., 2017; Gao and Wakefield, 2023; You and Hidiroglou, 2023; McGovern et al., 2026a). When outcome r is continuous, $\hat{v}_{ir}$, follows a chi-square distribution

asymptotically under simple random sampling, but for a complex survey, the distribution of the estimator is unknown and can be approximated by a chi-square distribution with degrees of freedom adjusted for survey design (Korn and Graubard, 2011, p. 34; McGovern et al., 2026a). For example, under a stratified two-stage cluster sampling design, the degrees of freedom can be approximated by the number of sampled clusters minus the number of strata (Korn and Graubard, 2011, p. 34; McGovern et al., 2026a).

In the original formulation of the FH model, the random effect vector is IID normal, but one may instead use a spatially structured component in the random effect to leverage spatial dependence (Rao and Molina, 2015, chap. 4.4.4; Chung and Datta, 2020; Wakefield et al., 2025). We will model the random effects using a BYM2 spatial model (Riebler et al., 2016). We define $(b_{1r}, ..., b_{mr})^{\mathrm{T}} \mid \sigma_r^2, \phi_r \sim \mathrm{BYM2}(\sigma_r^2, \phi_r)$ by

$$(b_{1r}, ..., b_{mr})^{\mathrm{T}} \mid \sigma_r^2, \phi_r \sim \mathcal{N}_m \left(\mathbf{0}, \sigma_r^2 \left(\phi_r \boldsymbol{Q}_*^- + (1 - \phi_r)\mathbf{I}_m\right)\right) \tag{42}$$

where σ_r^2 indicates the overall variability of the random effect, ϕ_r indicates the proportion of variation that is explained by the structured component, and $\boldsymbol{Q}_*^-$ is the generalized inverse of the scaled neighborhood structure matrix $\boldsymbol{Q}_*$, as defined in Riebler et al. (2016). The structured component is scaled such that the marginal variance of the random effect is approximately equal to σ_r^2 (Riebler et al., 2016). In computations we work with the singular precision matrix for the structured component and must enforce the sum-to-zero constraint, which is implied by the Moore-Penrose generalized inverse.

4.2.1 Univariate FH estimates using 2022 Kenya DHS data

The 2022 Kenya DHS is powered to provide estimates at the admin1-level with an average of 36 clusters sampled from each of 47 admin1 areas. Despite this design, estimation of some outcomes at the admin1 level may not be reliable, especially when the outcome is rare or has a high rate of missingness. One such example is the rate of neonatal deaths, which is defined

as death of a live born infant in the first 28 days of life. The Sustainable Development Goals (SDGs; United Nations General Assembly, 2015) provide a target of no more than 12 neonatal deaths per 1000 live births by 2030. Estimation of neonatal mortality rate (NMR) at the subnational level is important because it allows government officials at the country or regional level to evaluate which regions need more targeted interventions. The 2022 Kenya DHS records an average of 9.2 deaths in each admin1 area across the 2018-2022 period, with 10 out of the 47 areas observing 5 or fewer deaths (ICF International, 2022). Due to this data sparsity, design-based estimates can be extremely noisy, resulting in FH estimates with extreme shrinkage.

In Figure 4.1 we map the design-based and univariate FH estimates and their corresponding standard deviation estimates for NMR and two other outcomes: average height-for-age z-score (HAZ) and proportion of mothers who had at least 4 antenatal care visits. We obtain the design-based estimates using the `survey` package (Lumley, 2024) in `R`. We use the `Stan` software (Stan Development Team, 2024) for the univariate FH models, and estimate the posterior distribution using 4 chains of 2000 iterations after 1000 iterations of burn-in. We use a normal prior for β_r with mean μ_r and standard deviation 2. We choose $\mu_1 = -3$ to reflect that NMR is a rare outcome and for the two auxiliary outcomes we choose $\mu_r = 0$. The choice of prior standard deviation corresponds to a fairly uninformative prior for each outcome on its natural scale because, recall, the binary outcomes are transformed to the logit scale and the continuous outcome is a z-score. For each σ_r we use a penalized complexity (PC) prior with hyperparameters $u = 1$ and $\alpha = 0.01$, which corresponds to the prior belief that $P(\sigma_r > 1) = 0.01$ (Simpson et al., 2017), and for each ϕ_r we use a Beta$(0.5, 1)$ prior which has mean, 0.33, and standard deviation, 0.3, and mildly encourages shrinkage towards the simpler, non-spatially structured model.

In the first column of Figure 4.1 we observe that the design-based estimates for NMR have very high uncertainty in most areas and the univariate FH point estimates have extreme shrinkage. Conversely, we observe in the second and third columns that the univariate FH

estimates for the other two outcomes do not exhibit overshrinkage and have slightly lower uncertainty than the design-based estimates. The raw correlation between the design-based estimates for NMR and average HAZ is 0.248 and the raw correlation between the design-based estimates for NMR and proportion of mothers who had at least 4 antenatal care visits is -0.125. While these raw correlations are not extremely high, this could be a result of the noisiness of the design-based estimates, particularly that of NMR. These findings suggest that jointly estimating NMR with the other, less noisy, outcomes may improve estimation by leveraging shared spatial structure between the outcomes.

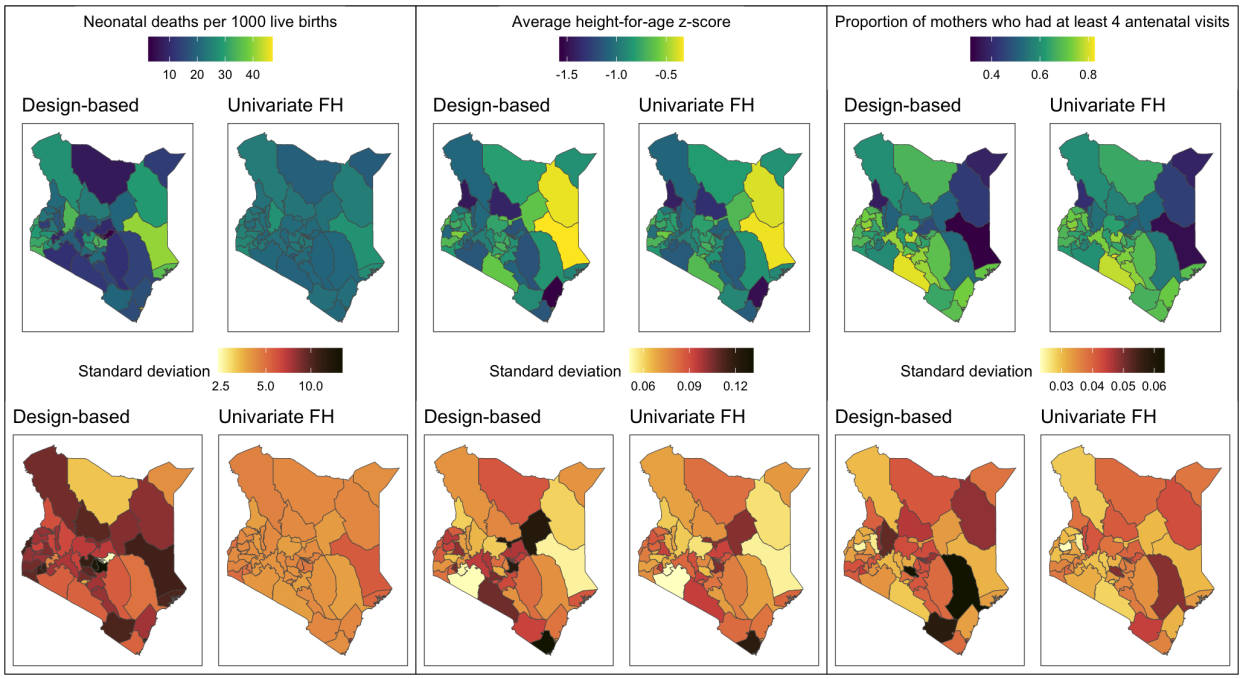


Figure 4.1. Design-based and univariate FH estimates of 2018-2022 neonatal mortality rate, average height-for-age z-score, and proportion of mothers who had ≥ 4 antenatal visits, using the 2022 Kenya DHS.

4.3 Multivariate Fay-Herriot model

Now we extend the univariate FH model to jointly model R outcomes. We use expression (40) to obtain design-based estimates, $\hat{\boldsymbol{\eta}}_i = (\hat{\eta}_{i1}, ..., \hat{\eta}_{iR})$, for $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{iR})$, for areas $i = 1, .., m$. As in the univariate case, $\hat{\boldsymbol{\eta}}_i$ is design-consistent and asymptotically

multivariate normal (Breidt and Opsomer, 2017) with some unknown sampling covariance, $\boldsymbol{V}_{\eta,i} = \text{Cov}(\hat{\boldsymbol{\eta}}_i \mid \boldsymbol{\eta}_i)$, which is often estimated via Taylor linearization.

Let $\hat{\boldsymbol{\theta}}_i = (g_1(\hat{\eta}_{i1}), ..., g_R(\hat{\eta}_{iR}))^{\text{T}}$, $\boldsymbol{\theta}_i = (g_1(\eta_{i1}), ..., g_R(\eta_{iR}))^{\text{T}}$, and $\boldsymbol{b}_i = (b_{i1}, ..., b_{iR})^{\text{T}}$ for each area i. The multivariate FH model is

$$\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \sim \mathcal{N}_R\left(\boldsymbol{\theta}_i, \hat{\boldsymbol{V}}_i\right), \qquad \theta_{ir} = \boldsymbol{X}_i^{\text{T}}\boldsymbol{\beta}_r + b_{ir} \qquad i = 1, ..., m, \tag{43}$$

where $\boldsymbol{b} = (\boldsymbol{b}_1^{\text{T}}, ..., \boldsymbol{b}_m^{\text{T}})^{\text{T}}$ is a vector of random effects modeled by a mean-zero multivariate normal distribution with some covariance $\boldsymbol{\Psi}$, and $\hat{\boldsymbol{V}}_i$ is a design-based estimate of $\boldsymbol{V}_i = \text{Cov}(\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i)$. Note that the sampling covariance, $\boldsymbol{V}_i$, for area i can also be expressed as $\boldsymbol{V}_i = \text{diag}(\boldsymbol{v}_i^{1/2})\boldsymbol{\Omega}_i\text{diag}(\boldsymbol{v}_i^{1/2})$ where $\boldsymbol{v}_i = (v_{i1}, ..., v_{iR})$ and $\boldsymbol{\Omega}_i$ is an $R \times R$ sampling correlation matrix. Analogous to the univariate case, we apply the multivariate delta method (Rao, 1973) to an estimate of $\boldsymbol{V}_{\eta,i}$ to obtain $\hat{\boldsymbol{V}}_i$, the distribution of which is unknown under complex sampling, but can be approximated by an inverse-Wishart with degrees of freedom adjusted for survey design. As in the univariate FH model, the sampling covariance matrix is generally treated as fixed in the multivariate FH model (Fay, 1987; Datta et al., 1998; Porter et al., 2015; Erciulescu and Opsomer, 2022). Observe that the multivariate FH model differs from the univariate FH model by accounting for between-outcome correlation at the observation level and allowing the borrowing of information across outcomes in the latent model, depending on the choice of random effects distribution.

Fay (1987) and Datta et al. (1991) first introduced the multivariate FH model with independent outcome-specific random effects,

$$b_{ir} \mid \sigma_r^2 \sim \mathcal{N}\left(0, \sigma_r^2\right), \qquad i = 1, ..., m \text{ and } r = 1, ..., R, \tag{44}$$

for hyperparameters $(\sigma_1^2, ..., \sigma_R^2)$. This formulation does not model shared latent spatial variability across outcomes or leverage latent spatial structure. It is most common in the multivariate FH literature to use an extension of (44), originally implemented in Datta et al.

(1998),

$$\boldsymbol{b}_i \mid \boldsymbol{\Sigma} \sim \mathcal{N}_R\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \qquad i = 1, ..., m, \tag{45}$$

where $\boldsymbol{\Sigma}$ is an $R \times R$ covariance matrix (Datta et al., 1999; Huang and Bell, 2004; Arima et al., 2017; Erciulescu and Opsomer, 2022). This formulation allows for correlation between outcomes at the latent level, but still assumes independence between areas.

Multivariate FH models which leverage spatial structure are largely understudied with two notable exceptions. Porter et al. (2015) substitutes a multivariate CAR model (Gelfand and Vounatsou, 2003) into the Datta et al. (1998) multivariate FH model to obtain,

$$\boldsymbol{b} \mid \boldsymbol{\Sigma}_m, \boldsymbol{\Sigma}_R \sim \mathcal{N}_{mR}\left(\mathbf{0}, \boldsymbol{\Sigma}_m \otimes \boldsymbol{\Sigma}_R\right), \tag{46}$$

where $\boldsymbol{\Sigma}_m$ is an $m \times m$ covariance matrix characterized by a CAR model, $\boldsymbol{\Sigma}_R$ is an unstructured $R \times R$ matrix, and $\otimes$ denotes the Kronecker product. This formulation requires CAR hyperparameters to be equal for all outcomes. Alternatively, Schumacher and Wakefield (2025) use a bivariate SCM,

$$b_{i1} = w_{i1} + \gamma w_{i2} \qquad i = 1, ..., m$$

$$b_{i2} = w_{i2} \qquad i = 1, ..., m$$

where $(w_{1k}, ..., w_{mk})^{\mathrm{T}} \mid \sigma_k^2, \phi_k \sim \text{BYM2}(\sigma_k^2, \phi_k)$ for $k = 1, 2$ and $\gamma \in \mathbb{R}$. While useful in the bivariate case, extending this model for $R > 2$ is nontrivial and requires incorporating constraints in the correlation structure.

4.3.1 The Generalized Multivariate BYM2 model

Here we introduce a class of spatial models we will call Generalized Multivariate BYM2 (GMBYM2) models. We will study several models within this class and discuss their relation to existing models in the SAE and disease mapping literature. Under the GMBYM2 model,

the random effects, $\boldsymbol{b} = (\boldsymbol{b}_1^{\mathrm{T}}, ..., \boldsymbol{b}_m^{\mathrm{T}})^{\mathrm{T}}$, are modeled by a mean-zero multivariate normal distribution with $mR \times mR$ covariance matrix,

$$\boldsymbol{\Psi} = \boldsymbol{Q}_*^- \otimes \boldsymbol{\psi}_u + \mathbf{I}_m \otimes \boldsymbol{\psi}_v, \tag{47}$$

for some $R \times R$ covariance matrices, $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$, where $\boldsymbol{Q}_*^-$ is the Moore-Penrose generalized inverse of the scaled neighborhood structure matrix $\boldsymbol{Q}_*$, as defined in Riebler et al. (2016). This form is a BYM2 analog of the general modeling framework introduced by Martinez-Beneito (2013), which uses the MCAR spatial model (Gelfand and Vounatsou, 2003). In contrast to the Martinez-Beneito (2013) framework, this covariance form is the sum of two distinct Kronecker products because the BYM2 model has both a structured and an unstructured component. The covariance parameters, $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi_v}$, model the between-outcome covariance within the structured and unstructured variance components, respectively. Thus, the pairwise correlation between outcomes r and r' implied by the structured and unstructured variance components are characterized by $\mathrm{Cor}(\boldsymbol{\psi}_u)_{rr'}$ and $\mathrm{Cor}(\boldsymbol{\psi}_v)_{rr'}$, respectively. Analogous to Martinez-Beneito (2013), we will show that the GMBYM2 model encompasses both correlated spatial models and shared component models which use BYM2 random effects. The models proposed in the remainder of this section are summarized in Table 4.1.

MV-Sep The first model we will study in this class is a BYM2 analog of the original multivariate FH model in (44),

$$(b_{1r}, ..., b_{mr})^{\mathrm{T}} \mid \sigma_r^2, \phi_r \sim \mathrm{BYM2}(\sigma_r^2, \phi_r) \qquad r = 1, ..., R,$$

or, equivalently,

$$\boldsymbol{b} \mid \boldsymbol{\Sigma}, \boldsymbol{\Phi} \sim \mathcal{N}_{mR}\left(\mathbf{0}, \boldsymbol{Q}_*^- \otimes \boldsymbol{\Phi}\boldsymbol{\Sigma} + \mathbf{I}_m \otimes (\mathbf{I}_R - \boldsymbol{\Phi})\boldsymbol{\Sigma}\right) \tag{48}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, ..., \sigma_R^2)$ and $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1, ..., \phi_R)$. As in (44), latent correlation between outcomes is assumed to be 0, as the latent model is composed of R independent latent

models. Therefore, when implemented into a multivariate FH model, the MV-Sep model only accounts for correlation between outcomes through the sampling covariance matrix in the observation model.

MV-Cor The next model we will study in the GMBYM2 class is an extension of the most commonly used multivariate FH model (45). The area-level random effects for each outcome r are modeled by a BYM2(σ_r^2, ϕ_r) model, and within each area, the latent correlation between outcomes is estimated by an $R \times R$ correlation matrix, $\mathbf{\Lambda}$. Precisely, the MV-Cor model is

$$\boldsymbol{b} \mid \mathbf{\Sigma}, \mathbf{\Phi} \sim \mathcal{N}_{mR}\left(\mathbf{0}, \boldsymbol{Q}_*^- \otimes (\mathbf{\Phi\Sigma})^{1/2}\mathbf{\Lambda}(\mathbf{\Phi\Sigma})^{1/2} + \mathbf{I}_m \otimes ((\mathbf{I}-\mathbf{\Phi})\mathbf{\Sigma})^{1/2}\mathbf{\Lambda}((\mathbf{I}-\mathbf{\Phi})\mathbf{\Sigma})^{1/2}\right) \quad (49)$$

where $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ are as defined for the MV-Sep model. This formulation leverages spatial dependence between areas and borrows information across outcomes through shared spatial variation. This latent model is a BYM2 analog of the MCAR (Gelfand and Vounatsou, 2003) and MBYM models (MacNab, 2011).

MV-Sym Now we propose a symmetric SCM model and show how it can be characterized within the GMBYM2 model class. We choose

$$b_{ir} = \gamma_r z_{0,i} + z_{r,i}, \qquad i = 1, ..., m \text{ and } r = 1, ..., R \quad (50)$$

where $\boldsymbol{z}_k = (z_{k,1}, ..., z_{k,m})^{\mathrm{T}}$ is a BYM2 random effect vector with hyperparameters $\sigma_{z,k}^2$ and $\phi_{z,k}$, for $k = 0, 1, ..., R$, and $\gamma_r \in \mathbb{R}$ determines the relative contribution of the shared random effect to outcome r. This formulation implies that all drivers of correlation, which are modeled as $\boldsymbol{z}_0$, are shared across all outcomes. When this model is used in disease mapping, the constraint $\sum_{r=1}^{R} \log(\gamma_r) = 0$ is usually applied to maintain identifiability (Held et al., 2005). This constraint implies that all correlation between outcomes must be positive. This constraint may be reasonable in some disease mapping contexts where comorbidity

can be assumed, but this is a rather strong assumption for general outcomes. There are many cases where outcomes could have negative pairwise correlation or we do not want to assume directionality of the correlation. Therefore, we propose maintaining identifiability by enforcing the constraints, $\sigma^2_{z,0} = 1$ and $\gamma_{r^*} > 0$ for some fixed $r^* \in [1, ..., R]$.

To express with MV-Sym model as a member of the GMBYM2 model class, we must define the forms of $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$. Let $\boldsymbol{z}^* = \left(\boldsymbol{z}_0^{\mathrm{T}}, \boldsymbol{z}_1^{\mathrm{T}}, ..., \boldsymbol{z}_R^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\boldsymbol{P}^* = \begin{pmatrix} \gamma_1 & 1 & 0 & ... & 0 \\ \gamma_2 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ \gamma_R & 0 & 0 & ... & 1 \end{pmatrix},$$

where $\boldsymbol{z}_k$ and γ_r are as defined in (50). It follows that, under the MV-Sym model, $\boldsymbol{b} = (\mathbf{I}_m \otimes \boldsymbol{P}^*)\boldsymbol{z}^*$, so, using the mixed-product property of Kronecker products (i.e $(A \otimes B)(C \otimes D) = AC \otimes BD$), the MV-Sym model can be expressed as a GMBYM2 model (47) with

$$\boldsymbol{\psi}_u = \boldsymbol{P}^* \boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}} \qquad \text{and} \qquad \boldsymbol{\psi}_v = \boldsymbol{P}^* (\mathbf{I}_{R+1} - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}} \tag{51}$$

where $\boldsymbol{\Sigma}_{z^*} = \mathrm{diag}(\sigma^2_{z,0}, \sigma^2_{z,1}, ..., \sigma^2_{z,R})$ and $\boldsymbol{\Phi}_{z^*} = \mathrm{diag}(\phi_{z,0}, \phi_{z,1}, ..., \phi_{z,R})$. See Appendix C1.1 for derivation. In Appendix C1.1 we also show that under the MV-Sym model, the pairwise correlations implied by the structured and unstructured components are characterized by

$$\mathrm{Cor}(\boldsymbol{\psi}_u)_{rr'} = \frac{\gamma_r}{\left(\gamma_r^2 + \frac{\phi_{z,r}}{\phi_{z,0}} \sigma^2_{z,r}\right)^{1/2}} \frac{\gamma_{r'}}{\left(\gamma_{r'}^2 + \frac{\phi_{z,r'}}{\phi_{z,0}} \sigma^2_{z,r'}\right)^{1/2}}$$

and

$$\mathrm{Cor}(\boldsymbol{\psi}_v)_{rr'} = \frac{\gamma_r}{\left(\gamma_r^2 + \frac{(1-\phi_{z,r})}{(1-\phi_{z,0})} \sigma^2_{z,r}\right)^{1/2}} \frac{\gamma_{r'}}{\left(\gamma_{r'}^2 + \frac{(1-\phi_{z,r'})}{(1-\phi_{z,0})} \sigma^2_{z,r'}\right)^{1/2}},$$

for $r \neq r'$, respectively. Observe that as the variance of the outcome-specific surface, $\sigma^2_{z,r}$,

increases with respect to the variance contribution from shared surface, γ_r^2, the correlation converges to 0. This intuitively implies that correlation between outcomes is stronger when more of each outcome's variability is shared. Moreover, we observe that the pairwise correlation implied by the structured and unstructured components are distinct, unless $\phi_{z,0} = \phi_{z,r}$ for $r = 1, ..., R$. In particular, when the variability in the shared surface is more spatially structured relative to the other surfaces (i.e. $\phi_{z,0} > \phi_{z,r}$), the correlation increases towards 1 in the structured component and decreases towards 0 in the unstructured component. These forms will be useful in Section 4.4.2 when we examine theoretical properties of the MV-Sym model as a function of correlation.

MV-Asym Lastly, we propose an asymmetric SCM model which focuses more explicitly on improving estimation of the primary outcome, $r = 1$, by using auxiliary outcomes $r = 2, ..., R$. We choose

$$\begin{aligned} b_{i1} &= z_{1,i} + \sum_{r=2}^{R} \gamma_r z_{i,r}, \qquad i = 1, \ldots, m \\ b_{ir} &= z_{r,i}, \qquad i = 1, \ldots, m \text{ and } r = 2, \ldots, R \end{aligned} \tag{52}$$

where $\boldsymbol{z}_r = (z_{r,1}, ..., z_{r,m})^{\mathrm{T}}$ is a BYM2 random effect vector with hyperparameters $\sigma_{z,r}^2$ and $\phi_{z,r}$, for $r = 1, ..., R$, and $\gamma_r \in \mathbb{R}$ for $r = 2, ..., R$ determines the relative contribution of the random effect for each auxiliary outcome $r > 1$ to the random effect for the primary outcome, $r = 1$. Note that the MV-Asym model is equivalent to a measurement error model (Fuller, 2009), where the auxiliary outcomes are treated as covariates with sampling errors which are correlated with that of each other and the primary outcome. In contrast to the MV-Sym model, this formulation only models pairwise correlation with the primary outcome and does not model pairwise correlation between auxiliary outcomes.

To derive the forms of $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$ under the MV-Asym model, let $\boldsymbol{z} = \left(\boldsymbol{z}_1^{\mathrm{T}}, ..., \boldsymbol{z}_R^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\boldsymbol{P} = \begin{pmatrix} 1 & \gamma_2 & ... & \gamma_R \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix},$$

where $\boldsymbol{z}_r$ and γ_r are as defined in (52). It follows that, under the MV-Asym model, $\boldsymbol{b} = (\mathbf{I}_m \otimes \boldsymbol{P})\boldsymbol{z}$, so, using the mixed-product property of Kronecker products, the MV-Asym model can be expressed as a GMBYM2 model (47) with

$$\boldsymbol{\psi}_u = \boldsymbol{P}\boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}} \qquad \text{and} \qquad \boldsymbol{\psi}_v = \boldsymbol{P}(\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}} \tag{53}$$

where $\boldsymbol{\Sigma}_z = \mathrm{diag}(\sigma^2_{z,1}, ..., \sigma^2_{z,R})$ and $\boldsymbol{\Phi}_z = \mathrm{diag}(\phi_{z,1}, ..., \phi_{z,R})$. See Appendix C1.2 for derivation. In Appendix C1.2 we also show that under the MV-Asym model, the pairwise correlations implied by the structured and unstructured components are characterized by

$$\mathrm{Cor}(\boldsymbol{\psi}_u)_{1r'} = \frac{\gamma_{r'}\sqrt{\phi_{z,r'}}\sigma_{z,r'}}{\left(\phi_{z,1}\sigma^2_{z,1} + \sum_{\ell=2}^{R}\gamma_\ell^2\phi_{z,\ell}\sigma^2_{z,\ell}\right)^{1/2}}$$

and

$$\mathrm{Cor}(\boldsymbol{\psi}_v)_{1r'} = \frac{\gamma_{r'}\sqrt{(1-\phi_{z,r'})}\sigma_{z,r'}}{\left((1-\phi_{z,1})\sigma^2_{z,1} + \sum_{\ell=2}^{R}\gamma_\ell^2(1-\phi_{z,\ell})\sigma^2_{z,\ell}\right)^{1/2}}$$

for $r' > 1$, and $\mathrm{Cor}(\boldsymbol{\psi}_u)_{rr'} = \mathrm{Cor}(\boldsymbol{\psi}_v)_{rr'} = 0$ for $r > r' > 1$, respectively. Observe that as the variance contribution to the primary outcome from the surface for outcome r', $\gamma^2_{r'}\sigma^2_{z,r'}$, decreases relative to the other variance contributions, the correlation between the primary outcome and outcome r' converges to 0. As with the MV-Sym model, we observe that the pairwise correlation implied by the structured and unstructured components are distinct, unless $\phi_{z,r}$ are equal for $r = 1, ..., R$. In particular, when the variability of the surface for

outcome r' is more structured than that of the other surfaces, the correlation between the primary outcome and outcome r' increases towards 1 in the structured component and decreases towards 0 in the unstructured component. These forms will be useful in Section 4.4.2 when we examine theoretical properties of the MV-Asym model as a function of correlation.

Observe that the MV-Sym and MV-Asym models are less complex than the MV-Cor model because they make assumptions about the correlation structure between outcomes.

	Form	Pairwise correlation
MV-Sep		
$\boldsymbol{\psi}_u$	$\boldsymbol{\Phi\Sigma}$	0
$\boldsymbol{\psi}_v$	$(\mathbf{I}-\boldsymbol{\Phi})\boldsymbol{\Sigma}$	0
MV-Cor		
$\boldsymbol{\psi}_u$	$(\boldsymbol{\Phi\Sigma})^{1/2}\boldsymbol{\Lambda}(\boldsymbol{\Phi\Sigma})^{1/2}$	$\boldsymbol{\Lambda}_{rk}$
$\boldsymbol{\psi}_v$	$((\mathbf{I}-\boldsymbol{\Phi})\boldsymbol{\Sigma})^{1/2}\boldsymbol{\Lambda}((\mathbf{I}-\boldsymbol{\Phi})\boldsymbol{\Sigma})^{1/2}$	$\boldsymbol{\Lambda}_{rk}$
MV-Sym		
$\boldsymbol{\psi}_u$	$\boldsymbol{P}^*\boldsymbol{\Phi}_{z^*}\boldsymbol{\Sigma}_{z^*}\boldsymbol{P}^*$	$\frac{\gamma_r\gamma_{r'}}{\sqrt{\left(\gamma_r^2+\frac{\phi_{z,r}}{\phi_{z,0}}\sigma^2_{z,r}\right)\left(\gamma_{r'}^2+\frac{\phi_{z,r'}}{\phi_{z,0}}\sigma^2_{z,r'}\right)}}$
$\boldsymbol{\psi}_v$	$\boldsymbol{P}^*(\mathbf{I}-\boldsymbol{\Phi}_{z^*})\boldsymbol{\Sigma}_{z^*}\boldsymbol{P}^*$	$\frac{\gamma_r\gamma_{r'}}{\sqrt{\left(\gamma_r^2+\frac{(1-\phi_{z,r})}{(1-\phi_{z,0})}\sigma^2_{z,r}\right)\left(\gamma_{r'}^2+\frac{(1-\phi_{z,r'})}{(1-\phi_{z,0})}\sigma^2_{z,r'}\right)}}$
MV-Asym		
$\boldsymbol{\psi}_u$	$\boldsymbol{P\Phi}_z\boldsymbol{\Sigma}_z\boldsymbol{P}$	$\frac{\gamma_r\sqrt{\phi_{z,r}}\sigma_{z,r}}{\sqrt{\phi_{z,r'}\sigma^2_{z,r'}+\sum_{\ell=2}^R\gamma_\ell^2\phi_{z,\ell}\sigma^2_{z,\ell}}}$, $r'=1$ 0, otherwise
$\boldsymbol{\psi}_v$	$\boldsymbol{P}(\mathbf{I}-\boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\boldsymbol{P}$	$\frac{\gamma_r\sqrt{(1-\phi_{z,r})}\sigma_{z,r}}{\sqrt{(1-\phi_{z,r'})\sigma^2_{z,r'}+\sum_{\ell=2}^R\gamma_\ell^2(1-\phi_{z,\ell})\sigma^2_{z,\ell}}}$, $r'=1$ 0, otherwise

Table 4.1. Summary of proposed GMBYM2 models. The first column defines the variance components which specify the model. The second column is the pairwise correlation between outcomes r and r', where $r \neq r'$, implied by each variance component.

4.4 Theoretical properties of the GMBYM2 FH models

We derive closed form expressions for the conditional posterior mean and variance under a multivariate FH model and use them to compare theoretical properties of the univariate FH model with that of the GMBYM2 FH models proposed in Section 4.3.1.

4.4.1 Conditional posterior distribution under the multivariate FH model

Let $\boldsymbol{\Theta} = \left(\boldsymbol{\theta}_1^{\mathrm{T}}, ..., \boldsymbol{\theta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$, $\hat{\boldsymbol{\Theta}} = \left(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}, ..., \hat{\boldsymbol{\theta}}_m^{\mathrm{T}}\right)^{\mathrm{T}}$, and $\hat{\boldsymbol{V}}$ be a block diagonal matrix with elements $\left(\hat{\boldsymbol{V}}_1, ..., \hat{\boldsymbol{V}}_m\right)$. Then a multivariate FH model as defined in (43) can also be expressed as

$$\hat{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta} \sim \mathcal{N}_{mR}\left(\boldsymbol{\Theta}, \hat{\boldsymbol{V}}\right) \qquad \boldsymbol{\Theta} \mid \boldsymbol{\beta}, \boldsymbol{\Psi} \sim \mathcal{N}_{mR}\left(\left(\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m\right)\boldsymbol{\beta}, \boldsymbol{\Psi}\right) \tag{54}$$

for some $mR \times mR$ covariance matrix $\boldsymbol{\Psi}$, where $\mathbf{X}$ is as defined in (41) and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, ..., \boldsymbol{\beta}_R^{\mathrm{T}}\right)^{\mathrm{T}}$. It follows that the posterior mean and variance of $\boldsymbol{\Theta}$, conditional on hyperparameters $(\boldsymbol{\beta}, \boldsymbol{\Psi})$, are

$$\mathbb{E}\left[\boldsymbol{\Theta} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{V}}\right] = \left(\hat{\boldsymbol{V}}^{-1} + \boldsymbol{\Psi}^{-1}\right)^{-1}\left(\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\Theta}} + \boldsymbol{\Psi}^{-1}(\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta}\right) \tag{55}$$

and

$$\mathrm{Var}\left(\boldsymbol{\Theta} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{V}}\right) = \left(\hat{\boldsymbol{V}}^{-1} + \boldsymbol{\Psi}^{-1}\right)^{-1}, \tag{56}$$

respectively. We can also write the conditional posterior mean of $\boldsymbol{\Theta}$ as

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Theta} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{V}}\right] &= \left(\hat{\boldsymbol{V}}^{-1} + \boldsymbol{\Psi}^{-1}\right)^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}\left(\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\Theta}} + \boldsymbol{\Psi}^{-1}(\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta}\right) \\
&= \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}\hat{\boldsymbol{V}}^{-1}\right)^{-1}\left(\boldsymbol{\Psi}\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{\Theta}} + (\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta} + \hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}\right) \\
&= \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}\hat{\boldsymbol{V}}^{-1}\right)^{-1}\left((\boldsymbol{\Psi}\hat{\boldsymbol{V}}^{-1} + \mathbf{I}_{mR})\hat{\boldsymbol{\Theta}} + (\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta} - \hat{\boldsymbol{\Theta}}\right) \\
&= \hat{\boldsymbol{\Theta}} - \mathcal{S}\left(\hat{\boldsymbol{\Theta}} - (\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta}\right)
\end{aligned}$$

where $\mathcal{S} = \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}\hat{\boldsymbol{V}}^{-1}\right)^{-1}$ is the 'shrinkage matrix' which determines how far the posterior mean will deviate from the design-based estimates. This parameterization is quite useful as we can compare $\mathcal{S}$ across models to better understand the theoretical properties of the conditional posterior means under each model as the sampling variance, $\hat{\boldsymbol{V}}$, and the latent variance, $\boldsymbol{\Psi}$, vary. Specifically, if $\tilde{\boldsymbol{\Theta}}^{(1)}$ and $\tilde{\boldsymbol{\Theta}}^{(2)}$ are the conditional posterior means under two models and $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ are their corresponding shrinkage matrices, then

$$\tilde{\boldsymbol{\Theta}}^{(1)} - \tilde{\boldsymbol{\Theta}}^{(2)} = \left(\mathcal{S}^{(2)} - \mathcal{S}^{(1)}\right)\left(\hat{\boldsymbol{\Theta}} - (\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta}\right).$$

Therefore, if the shrinkage matrices for two models are equal, then the conditional posterior means under those two models are equivalent. Furthermore, to compare the conditional posterior means for one outcome, r, across models, we can use the property

$$\tilde{\boldsymbol{\Theta}}_r^{(1)} - \tilde{\boldsymbol{\Theta}}_r^{(2)} = \left(\mathcal{S}_r^{(2)} - \mathcal{S}_r^{(1)}\right)\left(\hat{\boldsymbol{\Theta}} - (\boldsymbol{X}^{\mathrm{T}} \otimes \mathbf{I}_m)\boldsymbol{\beta}\right), \tag{57}$$

where $\tilde{\boldsymbol{\Theta}}_r$ is the outcome-specific subvector of $\tilde{\boldsymbol{\Theta}}$ and $\mathcal{S}_r$ is the $m \times mR$ submatrix of $\mathcal{S}$ containing the rows corresponding to outcome r. Note that we can also use this parameterization to specify the conditional mean under the univariate FH model with $\mathcal{S}^{\mathrm{Univ}} = \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}^{\mathrm{Sep}}\mathrm{diag}(\hat{\boldsymbol{V}})^{-1}\right)^{-1}$, where $\mathrm{diag}(\hat{\boldsymbol{V}})$ is a diagonal matrix with elements composed of vectors, $(\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_m)$.

4.4.2 Illustration of conditional posterior distribution properties

Now we assess how the conditional posterior distribution under the multivariate FH models differ from that under the univariate FH model, and how this varies as a function of the covariance parameters at the observation and latent levels. Here we will compare the GMBYM2 FH models proposed in Section 4.3.1 to a univariate BYM2 model, but we show in Appendix C2 that the findings in this section are generalizable to the broader class of multivariate FH models defined in (54). For this illustration, we use 2 auxiliary outcomes

for a total of $R = 3$ and $m = 64$ areas with a square grid spatial structure. For the BYM2 hyperparameters, we choose $\phi_r = 0.5$ and $\sigma_r^2 = 1$ for all outcomes. We use two configurations for the sampling variance. In the first configuration, we choose $\hat{v}_{ir} = 1$ for all areas and outcomes so that the signal-to-noise ratio is equal across outcomes, which reflects that all outcomes have equal levels of uncertainty. In the second configuration we choose $\hat{v}_{i1} = 1$ and $\hat{v}_{ir} = 0.25^2$ for $r = 2, 3$ so that the signal-to-noise ratio is higher for the auxiliary outcomes, which reflects that the the auxiliary outcomes have less uncertainty than the primary outcome.

For ease of comparison between models, we assume that the pairwise sampling correlation, $\text{Cor}(\hat{\boldsymbol{V}}_i)_{rr'} = \omega$, for some $\omega \in [0, 1)$, for all areas and outcomes, and all nonzero pairwise latent correlation is equal to $\rho \in [0, 1)$. To apply this equicorrelation assumption to each model, we can define $\boldsymbol{\Lambda}$ (for the MV-Cor model) and $\boldsymbol{\gamma}$ (for the MV-Sym and MV-Asym models) such that the nonzero pairwise correlations are all equal to ρ. This equicorrelation assumption also implies that $\phi_r = \phi^*$ for all $r = 1, 2, 3$ because this is the only way to ensure $\text{Cor}(\boldsymbol{\psi}_u) = \text{Cor}(\boldsymbol{\psi}_v)$ in the MV-Sym and MV-Asym models. Under the equicorrelation assumption, the conditional posterior distributions under the MV-Sym and MV-Cor models are equivalent. As a result of fixing the hyperparameters $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ in this illustration, for the MV-Asym model, we must bound ρ above by $\frac{1}{\sqrt{R-1}}$ to ensure that $\sigma_{z,1}^2 \geq 0$. See Appendix C3 for more details.

In Figure 4.2, we plot the normalized Frobenius distance between the shrinkage submatrices corresponding to the primary outcome under the GMBYM2 and univariate FH models,

$$\frac{|| \mathcal{S}_1^{\text{MV}} - \mathcal{S}_1^{\text{Univ}} ||_2}{|| \mathcal{S}_1^{\text{Univ}} ||_2}, \tag{58}$$

for $\mathcal{S}_1^{\text{MV}} \in \left\{ \mathcal{S}_1^{\text{Sep}}, \mathcal{S}_1^{\text{Cor}}, \mathcal{S}_1^{\text{Sym}}, \mathcal{S}_1^{\text{Asym}} \right\}$, along with the conditional posterior variance (56) of the primary outcome under each model, as a function of ρ, for several values of ω. It follows from (57) that when the Frobenius distance between the shrinkage matrices is 0, the conditional posterior means are equal. We observe that for the MV-Cor, MV-Sym, and

MV-Asym models, there is some ρ^* for which the conditional posterior mean and variance are equivalent to that under the univariate FH model, and at this point the conditional posterior variance is maximized. Across all settings, we observe that the MV-Cor, MV-Sym, and MV-Asym models are very similar when $\rho \leq \rho^*$, but when $\rho > \rho^*$, the MV-Asym model deviates more strongly from the univariate FH model, both in terms of mean and variance. For the MV-Cor and MV-Sym models, ρ^* equals ω when the signal-to-noise ratio is equal across all outcomes. We prove this result analytically in Appendix C4 where we show this is a result of $\boldsymbol{V}_i$ being proportional to $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$. Lastly, we observe that when the primary outcome is noisier than the auxiliary outcomes, ρ^* is less than ω. This insight is useful because it implies that when the primary outcome is noisier than the auxiliary outcomes, the latent correlation does not have to be as strong for there to be potential benefits of using a multivariate FH model.

We conclude that the conditional posterior mean and variance of an outcome under a multivariate FH models differs most from its univariate counterpart when sampling correlation is lower than latent correlation or the other outcomes its jointly modeled with are less noisy. In Appendix C2 we show that changes in σ_r^2, $\hat{v}_{ir}$, and ϕ_r modify overall levels, but the same patterns shown in Figure 4.2 persist. Crucially, these results all condition on correct hyperparameter specification, which cannot be assumed in practice, especially when using sparse data.

4.5 Simulation Study

Through this simulation study we evaluate how the theoretical differences between the univariate FH and multivariate FH models examined in Section 4.4 affect estimation of the primary outcome in practice. Specifically, we assess the accuracy and shrinkage of point estimates and the coverage and width of 90% credible intervals under each GMBY2-FH model, compared to that under the univariate FH model, for different choices of latent correlation and sampling covariance parameters.

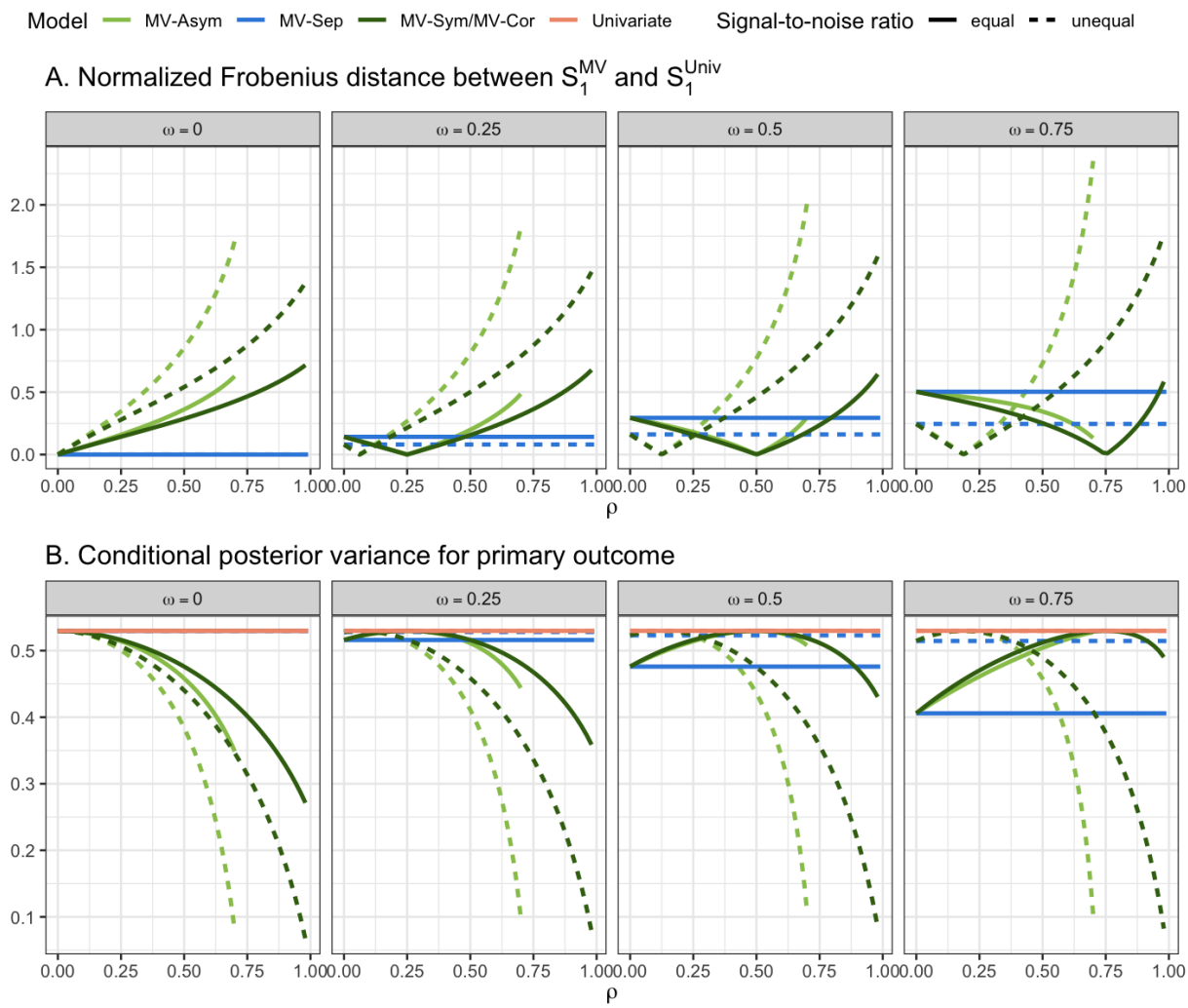


Figure 4.2. Properties of the conditional posterior distribution under GMBYM2-FH models compared to that under the univariate FH model, as a function of pairwise latent correlation, ρ, for different values of pairwise sampling correlation, ω. The solid lines indicate that signal-to-noise ratio is equal for all outcomes, and the dotted lines indicate that the primary outcome has a lower signal-to-noise ratio than the other outcomes.

4.5.1 Simulation design

We use one primary and two auxiliary outcomes for a total of $R = 3$ and $m = 64$ areas with a square grid spatial structure. For each realization, $j = 1, ..., 500$, we generate $\boldsymbol{b}^{(j)} \in \mathbb{R}^{mR}$ under the MV-Cor latent model (49). We choose $\boldsymbol{\Sigma}$ to be a 3×3 diagonal matrix with elements equal to 1, $\boldsymbol{\Phi}$ to be a 3×3 diagonal matrix with elements equal to 0.5, and $\boldsymbol{\Lambda}$ to be a correlation matrix with all off-diagonal elements equal to ρ, which will vary across

settings. This construction implies that each outcome-specific random effect vector follows a BYM2 distribution with $\sigma_r^2 = 1$ and $\phi_r = 0.5$, and that latent pairwise correlation between each outcome is equal to some ρ. Recall from Section 4.4.2 that the MV-Cor and MV-Sym models are equivalent in this equicorrelation case when conditioning on $(\boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{\Lambda})$, so this simulation design does not favor one model over the other. This latent model is chosen to reflect the real-world situation where pairwise correlations are present that the MV-Asym and MV-Sep models do not account for.

We choose the outcome-specific intercepts to be equal to 0 and do not include covariates, so it follows that $\boldsymbol{\theta}_i^{(j)} = \boldsymbol{b}_i^{(j)}$ for area $i = 1, ..., m$. In contrast to Section 4.4.2, we want to generate more realistic sampling covariance matrices, $\boldsymbol{V}_i^{(j)}$, which reflect the fact that in real survey data the sampling correlation can be quite different across areas. We define $\boldsymbol{V}_0 = \text{diag}(\boldsymbol{v}^{1/2})\boldsymbol{\Omega}\text{diag}(\boldsymbol{v}^{1/2})$ where $\boldsymbol{v} = (v_1, v_2, v_3)$ varies across settings and $\boldsymbol{\Omega}$ is a 3×3 correlation matrix with off-diagonal elements equal to some ω which also varies across settings. Then we generate $\boldsymbol{V}_i^{(j)}$ from an inverse-Wishart distribution with mean $\boldsymbol{V}_0$, and 20 degrees of freedom. By generating the sampling covariance matrix in this manner we can allow for differences across areas, while evaluating model performance for different average levels of sampling correlation, ω. Note that this construction allows for sampling correlation between outcomes to be negative. Then, for areas $i = 1, ..., m$, we use $\left(\boldsymbol{\theta}_i^{(j)}, \boldsymbol{V}_i^{(j)}\right)$ to generate design-based estimates,

$$\hat{\boldsymbol{\theta}}_i^{(j)} \sim \mathcal{N}_R\left(\boldsymbol{\theta}_i^{(j)}, \boldsymbol{V}_i^{(j)}\right).$$

Recall from Section 4.3 that the distribution of the design-based sampling covariance estimator, $\hat{\boldsymbol{V}}_i$, can be approximated by an inverse-Wishart distribution with the degrees of freedom adjusted for sampling design. Thus we generate $\hat{\mathbf{V}}_i^{(j)}$ from an inverse-Wishart distribution with mean $\mathbf{V}_i^{(j)}$ and degrees of freedom equal to 34, which mirrors the survey design and average sample sizes in the 2022 Kenya DHS (ICF International, 2022).

We choose simulation settings which vary the latent correlation parameter, ρ, and the sampling variance, $\boldsymbol{V}_i$. For the latent correlation parameter, ρ, we use four settings: $\rho =$

$0, 0.25, 0.5$, and 0.75. For the sampling covariance matrix we use 3 settings. For the first setting we choose $v_1 = 1$, $v_2 = v_3 = 0.25^2$, and $\omega = 0$; for the second setting we choose $v_1 = 1$, $v_2 = v_3 = 0.25^2$, and $\omega = 0.5$; and for the third setting we choose $v_1 = v_2 = v_3 = 1$, and $\omega = 0$. The first two settings allow us to assess how the average level of sampling correlation affects model performance, while the first and third settings allow us to assess how model performance changes when the auxiliary outcomes have less uncertainty relative to the primary outcome. We cross the 3 choices of sampling covariance matrix with the 4 choices of latent correlation for a total of 12 simulation settings.

4.5.2 Model Implementation

For each setting and realization, j, we use $\left(\hat{\boldsymbol{\Theta}}^{(j)}, \hat{\boldsymbol{V}}^{(j)}\right)$ to fit the MV-Sep (48), MV-Cor (49), MV-Sym (50), and MV-Asym (52) FH models. For comparison, we also fit a univariate FH model (41) for the primary outcome with the univariate BYM2 latent model defined in (42).

For all models and outcomes $r = 1, 2, 3$, we use a normal prior for β_r with mean 0 and standard deviation 2, a penalized complexity (PC) prior for σ_r with hyperparameters $u = 2$ and $\alpha = 0.01$, which corresponds to the prior belief that $P(\sigma_r > 2) = 0.01$ (Simpson et al., 2017), and a Beta$(0.5, 1)$ prior for ϕ_r which has mean, 0.33, and standard deviation, 0.3, and mildly encourages shrinkage towards the simpler, non-spatially structured model. For the MV-Cor model, we use a Lewandowski-Kurowicka-Joe (LKJ) prior for $\boldsymbol{\Lambda}$, with shape parameter equal to 2. Note that LKJ(1) corresponds to a uniform prior over valid correlation matrices, while a larger shape parameter value concentrates mass toward the identity matrix. We choose $\boldsymbol{\Lambda} \sim$ LKJ(2) to provide mild shrinkage towards the identity matrix, which regularizes the correlation estimates away from the boundary of the parameter space. For the MV-Sym and MV-Asym models, we use a normal prior for each γ_r with mean 0 and standard deviation 2. To increase stability of the MV-Sym model, we choose the outcome, r^*, for which $\gamma_{r^*} > 0$ to be the outcome with the lowest sampling variance because it has the least noise. In practice, γ_{r^*} must be bounded away from 0 because when it gets too

close to 0, sign identifiability is lost and the posterior distributions of each γ_r can become bimodal, causing convergence issues. Therefore, we bounded $\gamma_{r^*} > 0.25$ to prevent these convergence issues, making the prior for γ_{r^*} a truncated normal distribution. This lower bound affects the interpretation of the MV-Sym model because the relationship of each outcome to the shared surface is not precisely symmetric.

We use the `Stan` software (Stan Development Team, 2024) for all models, and estimate the posterior distribution using 4 chains of 2000 iterations after 1000 iterations of burn-in. `Stan` diagnostics indicate convergence and efficient sampling for all models and simulations. The average run times across 500 simulations for the univariate, MV-Sep, MV-Cor, MV-Sym, and MV-Asym FH models were 8, 102, 92, 102, and 108 seconds, respectively. The average effective sample sizes across 500 simulations for the univariate, MV-Sep, MV-Cor, MV-Sym, and MV-Asym FH models were 2304, 2298, 2021, 2332, and 2237, respectively.

4.5.3 Metrics

We compare the point estimates and 90% credible intervals for the primary outcome measure $r = 1$ under each model using four metrics. Let $\tilde{\theta}_{i1}^{(j,g)}$ denote the posterior mean estimate for $\theta_{i1}^{(j)}$ under model g, and $\left(l_{i1}^{(j,g)}, u_{i1}^{(j,g)}\right)$ denote its corresponding 90% credible interval. First, for each realization j, we evaluate the percent change in root mean squared error (RMSE) for each GMBYM2-FH model, compared to the univariate FH model,

$$\frac{\text{RMSE}^{(j,MV)} - \text{RMSE}^{(j,Univ)}}{\text{RMSE}^{(j,Univ)}} \times 100\%,$$

where

$$\text{RMSE}^{(j,g)} = \frac{1}{64}\sqrt{\sum_{i=1}^{64}\left(\tilde{\theta}_{i1}^{(j,g)} - \theta_{i1}^{(j)}\right)^2}$$

for model g. Next, for each realization j, we evaluate the ratio of the between-area variability of the point estimates under model g and the between-area variability of the true parameters,

which we will refer to as the 'shrinkage factor',

$$\sqrt{\frac{\sum_{i=1}^{64}\left(\tilde{\theta}_{i1}^{(j,g)}-\frac{1}{64}\sum_{k=1}^{64}\tilde{\theta}_{k1}^{(j,g)}\right)^2}{\sum_{i=1}^{64}\left(\theta_{i1}^{(j)}-\frac{1}{64}\sum_{k=1}^{64}\theta_{k1}^{(j)}\right)^2}}.$$

The shrinkage factor measures the between-area shrinkage of the point estimates, where a value less than 1 indicates shrinkage.

Finally, for each realization j, we evaluate average coverage of the 90% credible intervals across areas under each model, g,

$$\frac{1}{64}\sum_{i=1}^{64} I\left(\theta_{i1}^{(j)} \in \left[l_{i1}^{(j,g)}, u_{i1}^{(j,g)}\right]\right),$$

and the percent change in average interval width for each of the GMBYM2-FH models, compared to the univariate FH model,

$$\frac{\sum_{i=1}^{64}\left(u_{i1}^{(j,MV)}-l_{i1}^{(j,MV)}\right)-\sum_{i=1}^{64}\left(u_{i1}^{(j,Univ)}-l_{i1}^{(j,Univ)}\right)}{\sum_{i=1}^{64}\left(u_{i1}^{(j,Univ)}-l_{i1}^{(j,Univ)}\right)}\times 100\%.$$

4.5.4 Results

In Figure 4.3 we present the percent change in RMSE of the point estimates under each GMBYM2-FH model, relative to that under the univariate FH model, and the shrinkage factor for each model, across 500 realizations for each simulation setting. The columns indicate different sampling covariance settings and the rows indicate different latent correlation settings. We observe that the MV-Cor, MV-Sym, and MV-Asym models perform very similarly, with RMSE and shrinkage decreasing relative to the univariate FH model as latent correlation increases. This decrease is more pronounced when sampling correlation is lower and auxiliary outcomes have less uncertainty. As for the MV-Sep model, when both sampling and latent correlations are high, the model performs worse than the univariate FH

model, exhibiting higher RMSE and an increase in shrinkage. When either or both levels of correlation are low, the MV-Sep model performs very similarly to the univariate FH model.

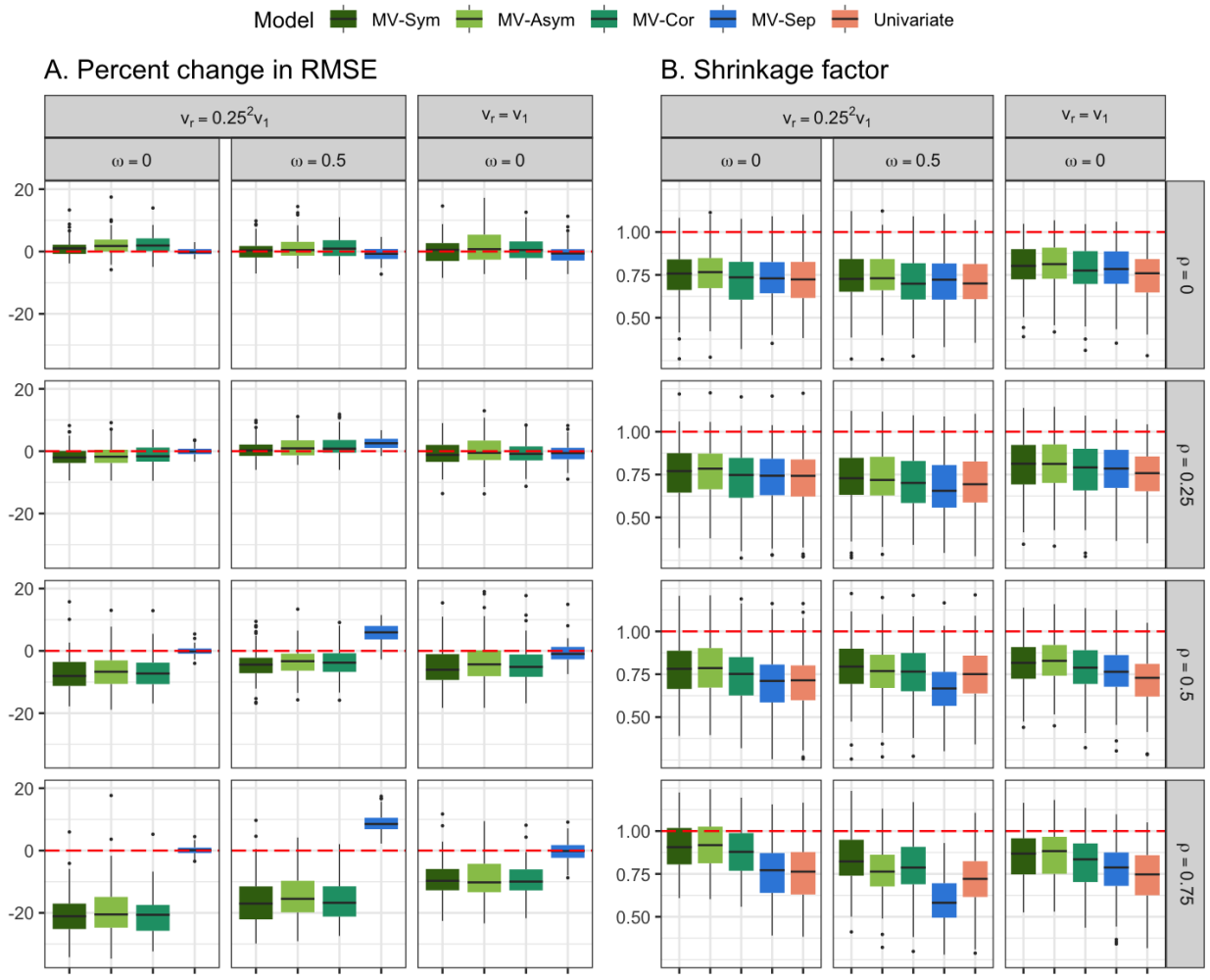


Figure 4.3. Percent change in RMSE of point estimates under each GMBYM2-FH model, relative to that under the univariate FH model, and shrinkage factor for the univariate and each GMBYM2 FH model, for 500 realizations for each simulation setting, which varies latent correlation, ρ, and average sampling correlation, ω.

In Figure 4.4 we present the coverage of the 90% credible intervals under each model and percent change in average interval width under each GMBYM2-FH model, relative to that under the univariate FH model, for 500 realizations for each simulation setting. In agreement with the point estimates, we observe that the MV-Cor, MV-Sym, and MV-Asym models perform similarly, with coverage similar to the univariate FH model and average interval width decreasing relative to the univariate FH model as latent correlation increases.

As in Figure 4.3, this decrease is more pronounced when sampling correlation is lower and auxiliary outcomes have less uncertainty. We observe that when latent correlation is low, the MV-Cor produces slightly more anti-conservative intervals, which are more narrow and have lower coverage. Also in agreement with Figure 4.3, we observe that the MV-Sep model performs similarly to the univariate FH model, except for when the sampling and latent correlation are high, at which point the MV-Sep model performs worse than the univariate FH model, with intervals that have low coverage.

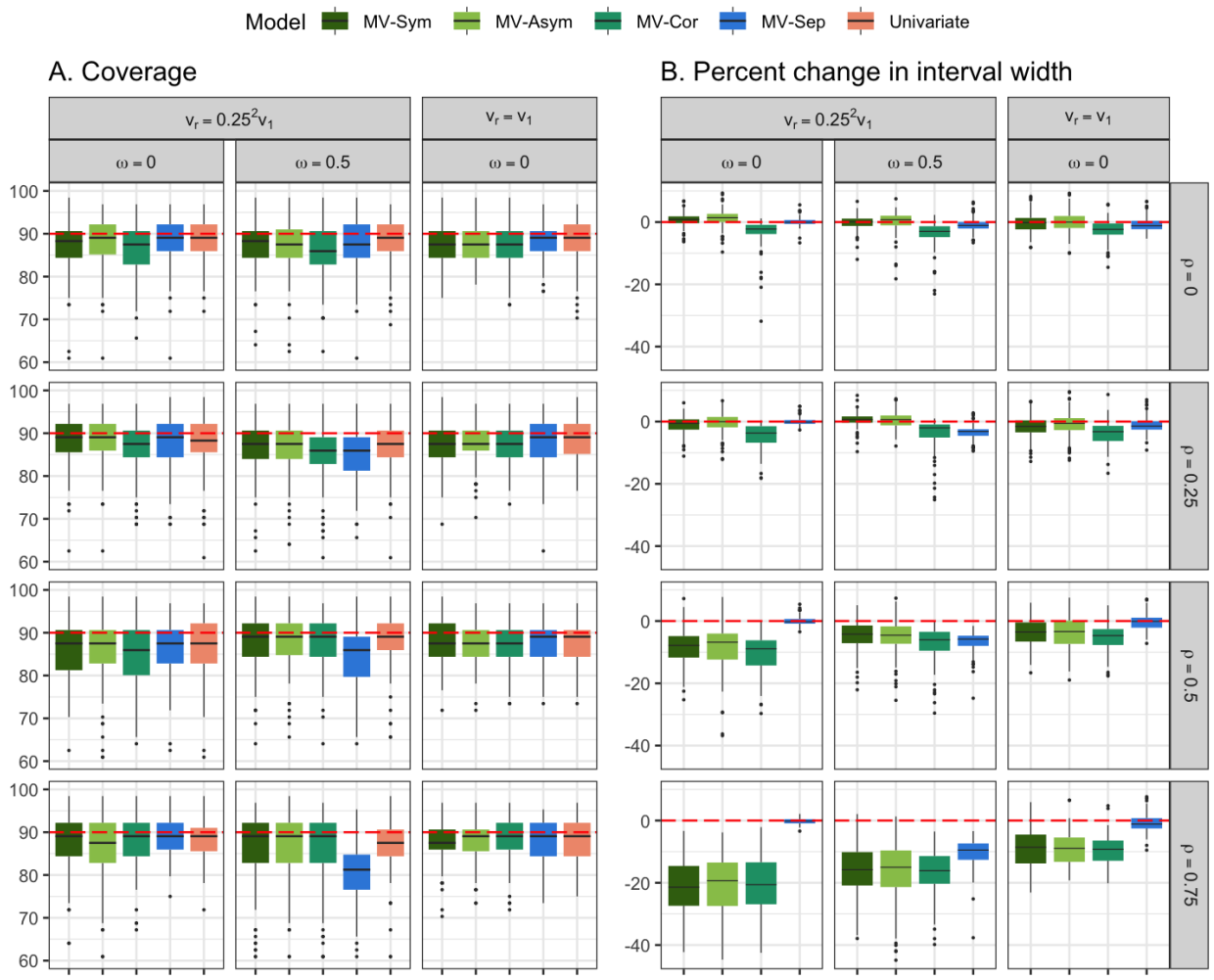


Figure 4.4. Coverage of 90% credible intervals for the univariate and each GMBYM2 FH model and percent change in average credible interval width for each GMBYM2-FH model, relative to univariate FH model, for 500 realizations under each simulation setting, which varies latent correlation, ρ, and average sampling correlation, ω.

Through this simulation study we conclude that when correlation is low, the GMBYM2-

FH models perform similarly to the univariate FH model, and when correlation is high the MV-Sep model performs worse than the univariate FH model, while the MV-Cor, MV-Sym, and MV-Asym models outperform the univariate FH model in terms of RMSE, shrinkage, and efficiency of credible intervals. Moreover, this improvement in estimation is more dramatic when the auxiliary outcomes have lower uncertainty than the primary outcome. An important consequence of our findings is that, under the tested settings, GMBYM2-FH models do not perform significantly worse than the univariate FH model when correlation is low, which implies that using a multivariate FH model when we are unsure of the true correlation structure would not negatively impact estimation.

4.6 Data Example

We return to the estimation of admin1-level NMR using 2022 Kenya DHS data introduced in Section 4.2.1 and use the proposed GMBYM2-FH models to jointly model NMR with two auxiliary outcomes: average HAZ and proportion of mothers who had at least 4 antenatal visits. We obtain design-based estimates for $\boldsymbol{\eta_i} = (\eta_{i1}, \eta_{i2}, \eta_{i3})^{\mathrm{T}}$, denoted $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} = (\hat{\eta}_{i1}, \hat{\eta}_{i2}, \hat{\eta}_{i3})^{\mathrm{T}}$, and an estimate of its sampling covariance matrix, $\mathrm{Cov}\left(\hat{\boldsymbol{\eta}}_i \mid \boldsymbol{\eta}_i\right)$, denoted $\hat{\boldsymbol{V}}_{\eta,i}$, for $i = 1, ..., 47$, using expression (40) and the `survey` package (Lumley, 2024) in `R`. We define $\theta_{ir} = g_r(\eta_{ir})$ and $\hat{\theta}_{ir} = g_r(\hat{\eta}_{ir})$ where g_r is the identity function for the continuous outcome and the logit function for the binary outcomes. Then for area i, we obtain an estimate of the sampling covariance matrix of $\hat{\boldsymbol{\theta}}_i$, denoted $\hat{\boldsymbol{V}}_i$, by applying the multivariate delta method (Rao, 1973) to $\hat{\boldsymbol{V}}_{\eta,i}$. We observe that, across areas, the pairwise sampling correlation between NMR and average HAZ is -0.03 on average (SD = 0.22) and the pairwise sampling correlation between NMR and proportion of mothers who had at least 4 antenatal care visits is 0.025 on average (SD = 0.23). Recall from the simulation study that the GMBYM2-FH models (except for MV-Sep) perform better, on average, when average sampling correlation across areas is low.

We use $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{V}})$ to fit the four proposed GMBYM2-FH models and compare them to

the univariate FH (41) estimates for NMR presented in Section 4.2.1. For all models and outcomes $r = 1, 2, 3$ we use the same priors on $(\beta_r, \sigma_r^2, \phi_r)$ as in Section 4.2.1. We choose the same priors for the correlation parameters as in the simulation study, and for the MV-Sym model, we choose average HAZ to be the outcome for which γ_{r^*} is nonnegative and restrict $\gamma_{r^*} > 0.1$ to maintain sign-identifiability. We use the `Stan` software (Stan Development Team, 2024) for all models, and estimate the posterior distribution using 4 chains of 2000 iterations after 1000 iterations of burn-in. `Stan` diagnostics indicate convergence and efficient sampling for all models. The run times for the univariate, MV-Sep, MV-Cor, MV-Sym, and MV-Asym FH models were 6, 86, 80, 91, and 88 seconds, respectively. The effective sample sizes for the univariate, MV-Sep, MV-Cor, MV-Sym, and MV-Asym FH models were 2248, 1927, 2243, 1983, and 1974, respectively.

In Figure 4.5 we visualize the posterior distributions of the latent pairwise correlations under each model. Recall from Section 4.3.1 that this latent pairwise correlation is defined by $\mathbf{\Lambda}_{1r}$ for $r = 2, 3$ under the MV-Cor model, and is defined by a function of $(\boldsymbol{\gamma}, \boldsymbol{\sigma}, \boldsymbol{\phi})$ for the structured and unstructured components under the MV-Sym and MV-Asym models. We observe that the pairwise correlation between NMR and average HAZ is estimated similarly across models, though the distributions have different shapes. We observe that under the MV-Sym model the pairwise correlation between NMR and proportion of women who had at least 4 antenatal care visits is estimated to very close to 0, in contrast to the other models which estimate negative correlation between the two outcomes. This discrepancy is likely a result of the construction of the MV-Sym model: because the two auxiliary outcomes are not strongly correlated (as we observe in the bottom row), the shared surface estimated under the MV-Sym model cannot simultaneously capture the correlation NMR has with each of the auxiliary outcomes.

In Figure 4.6 we present NMR estimates and width of the corresponding 90% credible intervals for admin1 areas in Kenya, using the univariate and GMBYM2 FH models. In Appendix C5 we also present estimates and credible interval widths for the auxiliary outcomes,

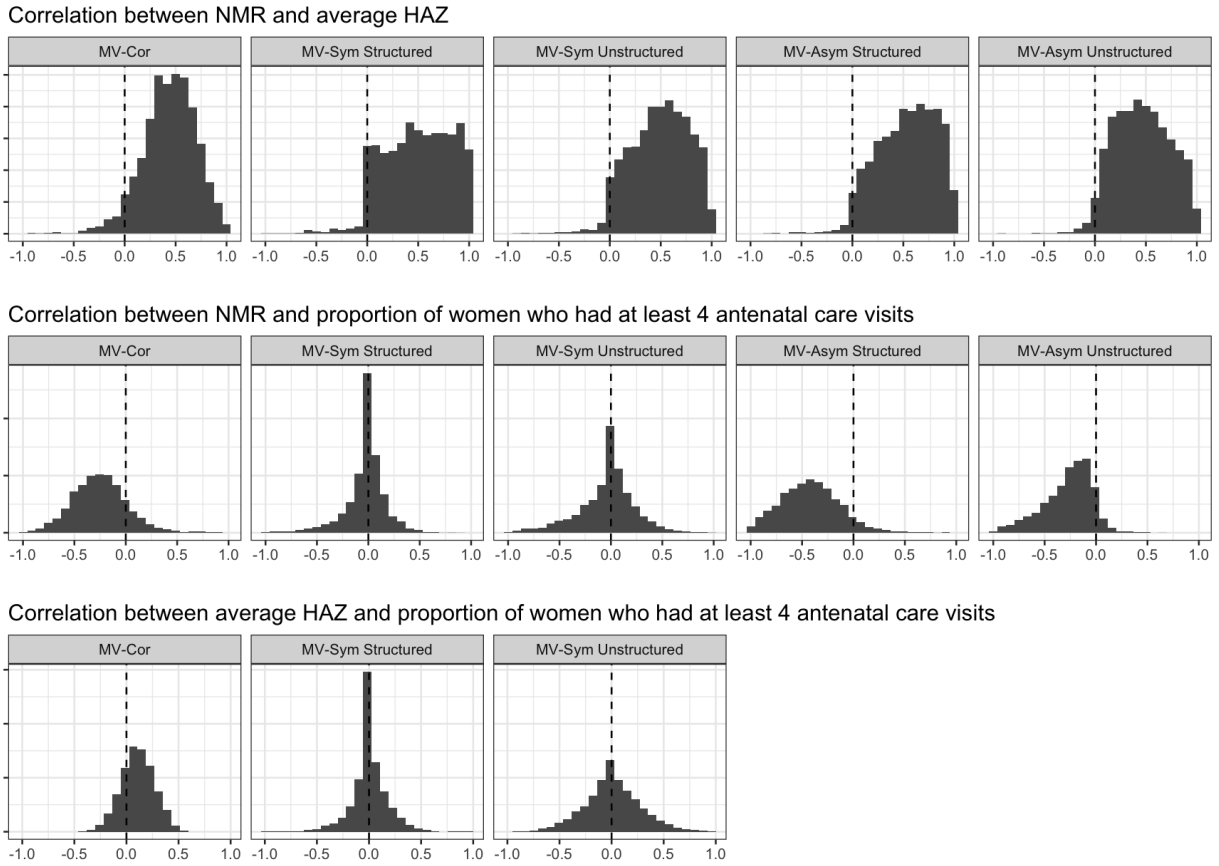


Figure 4.5. Posterior distributions of latent pairwise correlation under MV-Cor, MV-Sym, and MV-Asym FH models, using 2022 Kenya DHS data.

under each of the models. In agreement with the simulation study, we observe that the NMR estimates under the MV-Sep and univariate FH models are very similar and that the NMR estimates under the MV-Cor, MV-Sym, and MV-Asym models exhibit reduced shrinkage. Also in agreement with the simulation study, we observe that for most areas the intervals under the MV-Cor model are narrower than the those under MV-Sym and MV-Asym models. There are a few areas for which the MV-Cor, MV-Sym, and MV-Asym models yield wider intervals than the univariate FH model, which is likely the result of these models having less shrinkage. This effect is particularly clear in some of the eastern admin1 areas, where we observe an increase in both the NMR estimates and their corresponding interval widths. Because of the reduction in shrinkage, the MV-Cor, MV-Sym, and MV-Asym models retain more spatial variability which can be advantageous in identifying extreme areas.

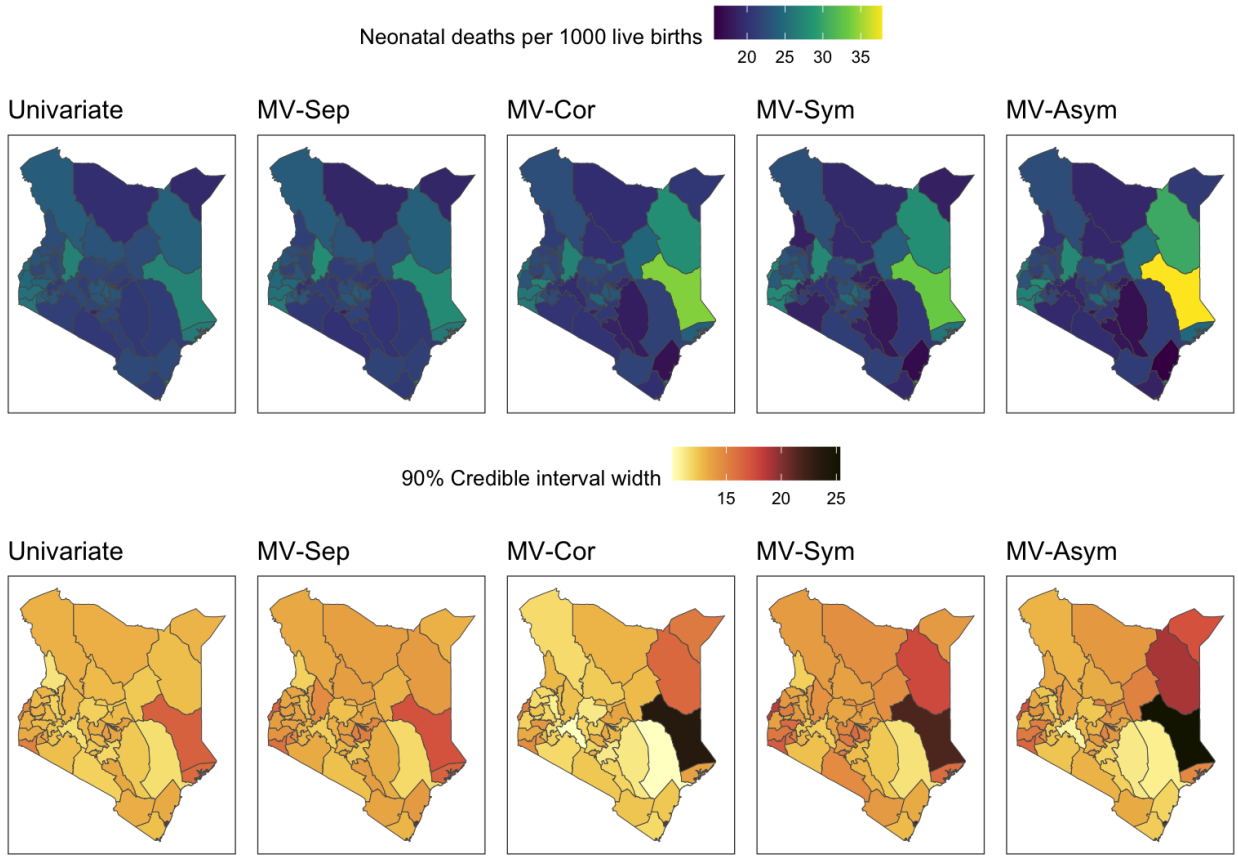


Figure 4.6. 2018-2022 NMR estimates and widths of 90% credible intervals for admin1 areas in Kenya under the univariate and GMBYM2 FH models, using 2022 Kenya DHS data.

4.7 Discussion

We began by establishing interest in using multivariate areal spatial models to improve estimation of a noisy primary outcome from complex survey data, and introduced the distinct considerations relevant to the multivariate FH model, particularly those resulting from the fixed covariance matrix in the observation model. We reviewed the existing multivariate FH literature, the majority of which does not utilize spatial structure, and introduced a class of multivariate latent models which use a BYM2 spatial model. Four models in this class were proposed: MV-Sep, which models independent latent distributions for each outcome; MV-Cor, which uses a correlated spatial model; and MV-Sym and MV-Asym, which utilize the SCM framework.

A general form was derived for the conditional posterior distribution under the multivariate FH models and followed by an examination of how the conditional posterior mean and variance for the primary outcome change under each model as a function of sampling and latent covariance. The MV-Cor, MV-Sym, and MV-Asym models all have some value of latent correlation, ρ^*, for which the conditional posterior mean and variance are equivalent to that under the univariate FH model. As latent correlation, ρ, increases relative to sampling correlation, ω, the conditional posterior variance decreases and the conditional posterior mean deviates further from that under the univariate FH model. This inflection point, ρ^* is lower when sampling correlation, ω, is low or the auxiliary outcomes are less noisy than the primary outcome.

In a simulation study, we assessed how the proposed GMBYM2-FH models performed in practice, in comparison to a univariate FH model. When latent correlation is high, the MV-Cor, MV-Sym, MV-Asym models all performed better than the univariate FH model in terms of RMSE, shrinkage, and interval efficiency. In agreement with our theoretical findings, the superior performance of the GMBYM2-FH model over the univariate FH model is amplified when sampling correlation is lower than latent correlation or the auxiliary outcomes have lower uncertainty than the primary outcome. The MV-Sep model performs similar to or worse than the univariate FH model, depending on strength of correlation. This is likely because the MV-Sep model introduces additional noise through the sampling correlation, but does not gain power by leveraging correlation at the latent level. Our application of the GMBYM2-FH models to the estimation of NMR in admin1 areas in Kenya validated our simulation study by yielding estimates from the MV-Sep model which were similar to that under the univariate FH model, and estimates from the other GMBYM2-FH models which had reduced shrinkage as compared to the univariate FH model.

While the MV-Cor, MV-Sym, and MV-Asym models performed similarly in our simulation study, there are a few key distinctions to be made when choosing between them. First, the MV-Sym model imposes the assumption that all underlying drivers of correlation

must be present in all outcomes which, as observed in the data example in Section 4.6, can be misspecified when auxiliary outcomes are correlated with the primary outcome, but not with each other. Moreover, the constraints imposed on the hyperparameters to allow for negative correlations limit the interpretability of the MV-Sym model, as whichever outcome has the constraint is favored in the shared surface. The MV-Cor model produces slightly more anti-conservative intervals, on average, when latent correlation is low, and is also the most complex of the three candidate models. The MV-Asym model, on the other hand, is more parsimonious because it only models correlation with the primary outcome. For these reasons, we have a slight preference towards the MV-Asym model, though we conclude that, under the tested settings, all three models are preferred over the univariate or MV-Sep FH model.

Chapter 5

CONCLUSION

This dissertation has focused on Bayesian spatial modeling frameworks for SAE which incorporate design-based estimates when data are too sparse for a standard Fay-Herriot (FH) model.

In Chapter 2 we introduced the DABUL model, a unit-level Bayesian hierarchical model for rare event prevalence which incorporates higher-level design-based direct estimates from the same data source, as either a soft or hard benchmarking constraint. The hard benchmarking constraint imposes exact agreement in aggregation, while the soft benchmarking constraint accounts for the uncertainty of the higher-level design-based estimates. In a simulation study we found that the soft DABUL model significantly reduces discrepancy in aggregation and that both DABUL models produce more conservative intervals with lower frequency of undercoverage for extreme areas than a standard unit-level model. In a cross-validation study for admin2-level NMR estimation using the 2013 Zambia DHS, we found that the soft DABUL model performed best in terms of average absolute error and interval score. This work has been published as McGovern et al. (2026b).

In Chapter 3 we proposed two sampling distributions for the complex design variance estimator and compared their performance in a variance smoothing FH model. The simpler sampling distribution arises from the common recommendation of using a chi-squared distribution with degrees of freedom equal to number of sampled clusters minus number of sampled strata, while the other sampling distribution makes less strict assumptions on the survey design and, as we validate through simulations, is a closer approximation of the empirical sampling distribution. In a simulation study, we found that the two variance smoothing models perform similarly to each other and produce better intervals than the FH model without variance smoothing, according to proper scoring rules. Based on the simulation study, we recommend the use of the simpler sampling distribution because, aside from its

easier implementation, we found that the additional assumptions of the simpler sampling distribution results in more shrinkage of the design variance towards 0, which helps to counteract the observational noise inherent in sparse data. We also applied these models to the estimation of height-for-age z-score using the 2022 Kenya DHS, which validated our simulation study findings. We provide a vignette for implementing the simple variance smoothing model `here`. A paper is available as McGovern et al. (2026a).

In Chapter 4 we studied theoretical properties of the multivariate FH model and adapted multivariate areal spatial models from the disease mapping literature and incorporated them into multivariate FH models. We introduced a class of multivariate latent models which use a BYM2 spatial model, and proposed four models in this class: MV-Sep, which models independent latent distributions for each outcome; MV-Cor, which uses a correlated spatial model; and MV-Sym and MV-Asym, which utilize the SCM framework. We derived a general form for the conditional posterior distribution under the multivariate FH models and studied how the conditional posterior mean and variance for the primary outcome change under each model as a function of sampling and latent covariance. In a simulation study, we found that when latent correlation is high, the MV-Cor, MV-Sym, MV-Asym models all performed better than the univariate FH model in terms of RMSE, shrinkage, and interval efficiency, and that this superior performance is amplified when sampling correlation is lower than latent correlation or the auxiliary outcomes have lower uncertainty than the primary outcome. Our application of the GMBYM2-FH models to the estimation of NMR in admin1 areas in Kenya validated our simulation study findings. While the MV-Cor, MV-Sym, and MV-Asym models performed similarly, we recommend the MV-Asym model for improving estimation of a primary outcomes because it's more parsimonious than the MV-Cor model and less restrictive than the MV-Sym model.

Future directions include extending the study of sampling distributions of the design variance estimator to accommodate binary outcomes and multivariate modeling. The sampling distributions studied in Chapter 3 only hold for binary data if we may assume the

cluster-level proportions are normally distributed. Further study is required to determine how robust a variance smoothing FH model for binary outcomes would be to violations of this assumption, or if there is another more suitable sampling distribution to consider. There is also potential to merge the models introduced in Chapters 3 and 4 by incorporating an inverse-Wishart sampling distribution into a multivariate variance smoothing FH model, which may make the joint modeling of a higher number of outcomes more practical.

References

AbouZahr, C., de Savigny, D., Mikkelsen, L., Setel, P. W., Lozano, R., Nichols, E., Notzon, F., and Lopez, A. D. (2015). Civil registration and vital statistics: Progress in the data revolution for counting and accountability. *The Lancet*, 386:1373–1385.

Adin, A., Krainski, E. T., Lenzi, A., Liu, Z., Martínez-Minaya, J., and Rue, H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, 62.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180:1191–1209.

Astfalck, L., Sen, D., Patra, S., Cripps, E., and Dunson, D. (2018). Posterior projection for inference in constrained spaces. *arXiv preprint arXiv:1812.05741*.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

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.

Bell, W. R., Datta, G. S., and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100:189–202.

Berg, E. and Fuller, W. A. (2018). Benchmarked small area prediction. *Canadian Journal of Statistics*, 46:482–500.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation

with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8:657–676.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36:192–225.

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.

Botella-Rocamora, P., Martinez-Beneito, M., and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine*, 34:1548–1559.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32:190–205.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, 90:2499–2523.

Burris, K. C. and Hoff, P. D. (2020). Exact adaptive confidence intervals for small areas. *Journal of Survey Statistics and Methodology*, 8:206–230.

Cho, M. J., Eltinge, J. L., Gershunskaya, J., and Huff, L. (2002). Evaluation of generalized variance function estimators for the US current employment survey. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 534–539.

Chung, H. C. and Datta, G. S. (2020). Bayesian hierarchical spatial models for small area estimation. Technical report, Washington, DC: Center for Statistical Research & Methodology, U.S. Census Bureau.

Corsi, D. J., Neuman, M., Finlay, J. E., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology*, 41:1602–1613.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65:1254–1261.

Datta, G., Fay, R., and Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In *Proceedings of Bureau of the Census 1991 Annual Research Conference*, pages 63–79. US Bureau of the Census Washington DC.

Datta, G. and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science*, 27:95–114.

Datta, G. S., Day, B., and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75:269–279.

Datta, G. S., Day, B., and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 60:344–362.

Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test*, 20:574–588.

De Nicolò, S., Fabrizi, E., and Gardini, A. (2024). Mapping non-monetary poverty at multiple geographical scales. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187:1096–1119.

Diggle, P. J. and Giorgi, E. (2019). *Model-Based Geostatistics for Global Public Health: Methods and Applications*. Chapman and Hall/CRC.

Dong, Q., Wu, Y., Li, Z. R., and Wakefield, J. (2026). Toward a principled workflow for prevalence mapping using household survey data. *Journal of Survey Statistics and Methodology*, 14:209–237.

Dong, T. Q. and Wakefield, J. (2021). Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine*, 39:2584–2594.

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, 59:286–295.

Earth Observation Group, NOAA National Centers for Environmental Information (2022). VIIRS nighttime lights version 2 annual composites. `https://eogdata.mines.edu/products/vnl/`.

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.

Erciulescu, A. L. and Opsomer, J. D. (2022). A model-based approach to predict employee compensation components. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71:1503–1520.

Fay, R. E. (1987). Application of multivariate regression to small domain estimation. *Small Area Statistics*, pages 91–102.

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.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37:4302–4315.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Gao, P. A. and Wakefield, J. (2023). A spatial variance-smoothing area level model for small area estimation of demographic rates. *International Statistical Review*, 91:493–510.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4:11–15.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., et al. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the sustainable development goals. *The Lancet*, 390:2171–2182.

Hájek, J. (1971). Discussion of "An essay on the logical foundations of survey sampling, part I", by D. Basu. In *V. Godambe and D. Sprott (Eds.), Foundations of Statistical Inference*. Holt, Rinehart and Winston.

Held, L., Natário, I., Fenton, S. E., Rue, H., and Becker, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research*, 14:61–82.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.

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.

Huang, E. T. and Bell, W. R. (2004). An empirical study on using ACS supplementary survey data in SAIPE state poverty models. In *2004 Proceedings of the American Statistical Association*, pages 3677–3684. US Bureau of the Census Washington DC.

ICF International (2014). Zambia demographic and health survey 2013–14 [dataset].

ICF International (2022). Kenya demographic and health survey 2022 [dataset].

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized Hierarchical multivariate CAR models for areal data, second edition. *Biometrics*, 61:950–961.

Kawano, S., Parker, P. A., and Li, Z. R. (2026). On data thinning for model validation in small area estimation. *arXiv preprint arXiv:2604.04141*.

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 164:73–85.

Korn, E. L. and Graubard, B. I. (2011). *Analysis of health surveys*. John Wiley & Sons.

Li, Z. R., Martin, B. D., Dong, T. Q., Fuglstad, G.-A., Godwin, J., Paige, J., Riebler, A., Clark, S., and Wakefield, J. (2024). Space-time smoothing of demographic and health indicators using the R package SUMMER. *arXiv preprint arXiv:2007.05117*.

Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.

Lumley, T. (2024). Survey: Analysis of complex survey samples. Version 4.4.

MacNab, Y. C. (2010). On Bayesian shared component disease mapping and ecological regression with errors in covariates. *Statistics in Medicine*, 29:1239–1249.

MacNab, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, 20:49–68.

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.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22:1649–1660.

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100:539–553.

McGovern, A., Fuglstad, G.-A., and Wakefield, J. (2026a). Sampling distributions for complex design variance estimators in a Fay-Herriot model. *arXiv preprint arXiv:2604.23029*.

McGovern, A., Wilson, K., and Wakefield, J. (2026b). Direct-assisted Bayesian unit-level modeling for small area estimation of rare event prevalence. *Journal of Survey Statistics and Methodology*, 14:178–208.

McIllece, J. J. (2018). On generalized variance functions for sample means and medians. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*.

Merfeld, J. D., Newhouse, D. L., Weber, M., and Lahiri, P. (2022). Combining survey and geospatial data can significantly improve gender-disaggregated estimates of labor market outcomes. *Institute of Labor Economics*.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49:65–82.

Märtens, K. (2017). NUTS. `https://github.com/kasparmartens/NUTS`.

Nandram, B. and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37:137–152.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2024). Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25:1–35.

Okonek, T. and Wakefield, J. (2024). A computationally efficient approach to fully Bayesian benchmarking. *Journal of Official Statistics*, 40:283–316.

Osgood-Zimmerman, A. and Wakefield, J. (2023). A statistical review of template model builder: A flexible tool for spatial modelling. *International Statistical Review*, 91:318–342.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2022). 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.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431.

Parker, P. A., Holan, S. H., and Janicki, R. (2024). Conjugate modeling approaches for small area estimation with heteroscedastic structure. *Journal of Survey Statistics and Methodology*, 12:1061–1080.

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

Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9:73–84.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015). Small area estimation via multivariate Fay–Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, 57:15–29.

Pratt, J. W. (1963). Shorter confidence intervals for the mean of a normal distribution with known variance. *Annals of Mathematical Statistics*, 34:574–586.

Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley New York.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Retegui, G., Etxeberria, J., Riebler, A., and Ugarte, M. D. (2023). Predicting cancer incidence in regions without population-based cancer registries using mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186:874–889.

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.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71:319–392.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

SAS Institute Inc. (2016). *Statistical Analysis Software (SAS) User's Guide, Version 9.4*.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10:1677–1708.

Schumacher, A. E. and Wakefield, J. (2025). Small area estimation methods for multivariate health and demographic outcomes using complex survey data. *arXiv preprint arXiv:2511.15917*.

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

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

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.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80:219–238.

Stan Development Team (2024). The Stan Core Library. Version 2.35.0.

Stefan, M. and Hidiroglou, M. A. (2021). Small area benchmarked estimation under the basic unit level model when the sampling rates are non-negligible. *Survey Methodology*, 47:123–150.

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. (2017). Worldpop, open data for spatial demography. *Scientific data*, 4:1–4.

UNICEF (1995–2026). Multiple indicator cluster surveys (MICS). Technical report, United Nations Children's Fund.

United Nations General Assembly (2015). Transforming our world: the 2030 agenda for sustainable development.

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

Vicente, G., Adin, A., Goicoa, T., and Ugarte, M. D. (2023). High-dimensional order-free multivariate spatial disease mapping. *Statistics and Computing*, 33:104.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 28:2614–2634.

Wakefield, J., Gao, P., Fuglstad, G.-A., and Li, Z. R. (2025). The two cultures of prevalence mapping: Small area estimation and model-based geostatistics. *Statstical Science*. to appear.

Wakefield, J., Jiang, J., and Wu, Y. (2026). Automatic variance adjustment for small area estimation. *arXiv preprint arXiv:2602.14387*.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review*, 88:398–418.

Wang, J., Fuller, W. A., and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34:29.

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.

WHO Multicentre Growth Reference Study Group (2006). *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. World Health Organization.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer New York.

Wu, Y. and Wakefield, J. (2024). Modelling urban/rural fractions in low-and middle-income countries. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187:811–830.

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

You, Y. and Hidiroglou, M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, 39:571–590.

You, Y. and Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30:431–439.

You, Y. and Rao, J. N. K. (2003). Pseudo hierarchical bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111:197–208.

Yu, C. and Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105:319–335.

Zhang, J. L. and Bryant, J. (2020). Fully Bayesian benchmarking of small area estimation models. *Journal of Official Statistics*, 36:197–223.

Appendix A

APPENDIX TO CHAPTER 2

A1 Definitions and notation

$\odot$ denotes the Hadamard product, i.e., for 2 $m-$length vectors, $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}\odot\mathbf{y}$ is an $m-$length vector with elements $(\mathbf{x}\odot\mathbf{y})_i = \mathbf{x}_i\mathbf{y}_i$.

Element	Description	Corresponding vector(s)
m_k	number of k^{th} administrative areas	
Δ	set containing all clusters	
$\delta_k(j)$	set containing all clusters in k^{th} admin area j	
(δ_U, δ_R)	set of all urban and rural clusters, respectively	
γ_c	indicator variable denoting whether cluster c was sampled	$\boldsymbol{\gamma} = \{\gamma_c : c \in \Delta\}$ $\boldsymbol{\gamma}_i = \{\gamma_c : c \in \delta_1(i)\}$
n_c	number of births in the sampled households in cluster c	$\mathbf{n} = \{n_c : c \in \Delta\}$
N_c	number of births in all households in cluster c	$\mathbf{N} = \{N_c : c \in \Delta\}$ $\mathbf{N}_i = \{N_c : c \in \delta_1(i)\}$ $\mathbf{N}_i^{(s)} = \{N_c : \gamma_c = 1\}$
$N_{i+} = \sum_{c\in\delta_1(i)} N_c$	number of births in admin1 area i	$\mathbf{N}_+ = \{N_{i+} : i = 1, ..., m_1\}$
Z_c	number of neonatal deaths in the sampled households in cluster c	$\mathbf{Z} = \{Z_c : c \in \Delta\}$ $\mathbf{Z}_i = \{Z_c : c \in \delta_1(i)\}$
Y_c	number of neonatal deaths in all households in cluster c	$\mathbf{Y} = \{Y_c : c \in \Delta\}$ $\mathbf{Y}_i = \{Y_c : c \in \delta_1(i)\}$ $\mathbf{Y}^{(s)} = \{Y_c : \gamma_c = 1\}$ $\mathbf{Y}_i^{(s)} = \mathbf{Y}_i \cap \mathbf{Y}^{(s)}$
$Y_{i+} = \sum_{c\in\delta_1(i)} Y_c$	number of neonatal deaths in admin1 area i	$\mathbf{Y}_+ = \{Y_{i+} : i = 1, ..., m_1\}$
$Y_{i+}^{(s)} = \sum_{\substack{c\in\delta_1(i)\\ \gamma_c=1}} Y_c$	number of neonatal deaths in the sampled households in admin1 area i	$\mathbf{Y}_+^{(\mathbf{s})} = \{Y_{i+}^{(s)} : i = 1, ..., m_1\}$
r_c	risk of neonatal death in cluster c	$\mathbf{r} = \{r_c : c \in \Delta\}$ $\mathbf{r}_i = \{r_c : c \in \delta_1(i)\}$
$r_i^{D_k}$	design-based prevalence estimate for k^{th} administrative area i	$\mathbf{r}^{\mathbf{D_k}} = \{r_i^{D_k} : i = 1, ..., m_k\}$
$V_j^{D_k}$	design-based variance of logit$(\hat{r}_i^{D_k})$	$\mathbf{V}^{\mathbf{D_k}} = \{V_j^{D_k} : i = 1, ..., m_k\}$

A2 Comparison of spatial models for Zambia NMR

We compare NMR estimates and uncertainty under model (6) using IID and BYM2 spatial effects. From these figures, we observe that, in this example, choice of spatial effect model makes little difference, so we choose BYM2 as it provides the option of more structure (by including an ICAR component in addition to an IID component) and has the added benefit of allowing more informative estimation in areas without observed data. Note that uncertainty of estimates under the BYM2 spatial effect is lower on average, and those for which it is higher are areas which have both low sample size and surrounding neighbors with low sample size.

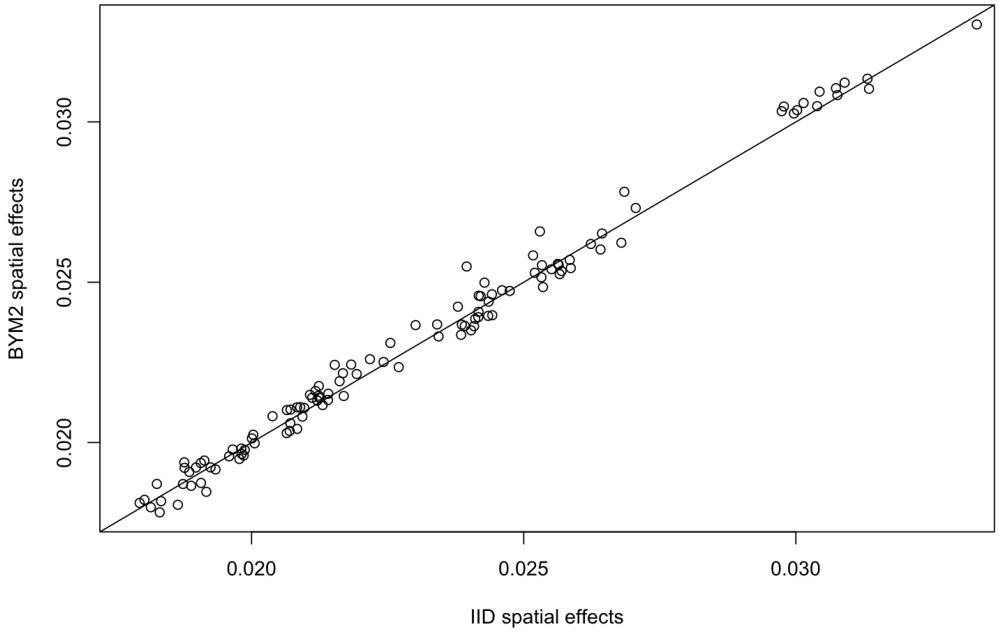


Figure A1. Admin2-level NMR estimates

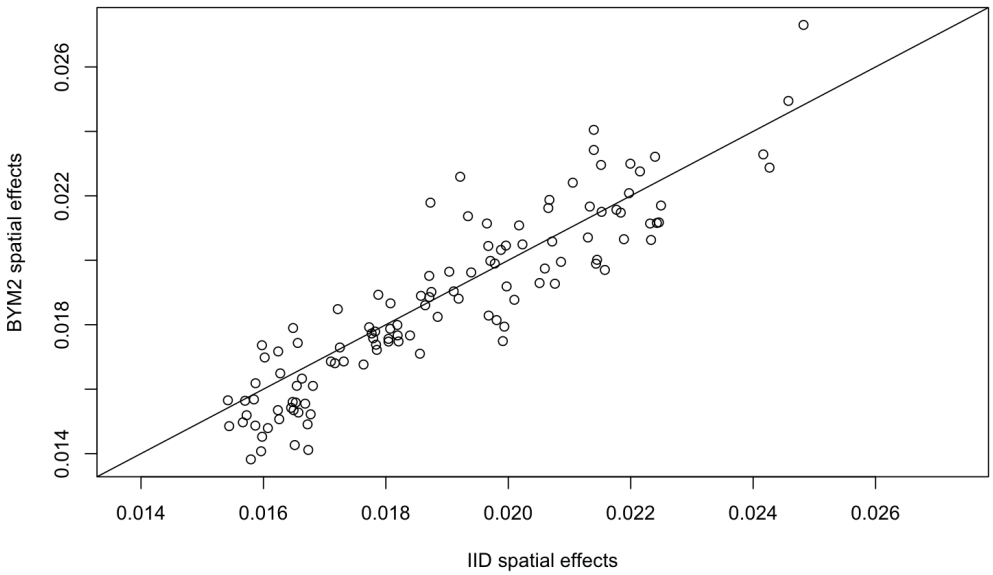


Figure A2. Width of 90% credible intervals for admin2-level NMR estimates

A3 Priors on regression and overdispersion parameters

The priors on the regression parameters $(\alpha_U, \alpha_R, \boldsymbol{\beta})$ are chosen with the goal of being relatively diffuse on the exponential scale, while still accounting for the prior belief that the outcome measure is a rare event. Figure A3 displays how variability of the Normal distribution is translated on the exponential scale in the case of rare events. Note that when the variance on the linear scale is large, as in the panel on the right, the distribution on the exponential scale has very high probability near 0, so the variability on the desired scale actually *decreases*. From this figure we can deduce that the chosen prior on (α_U, α_R), $\mathcal{N}(-3.5, 3^2)$, is in fact a diffuse prior in this context. The chosen prior for the fixed effects $\boldsymbol{\beta}$, $\mathcal{N}(0, 1)$, has a smaller variance to avoid a prior on the linear predictor with such high variability that the implied prior on the exponential scale shrinks towards 0.

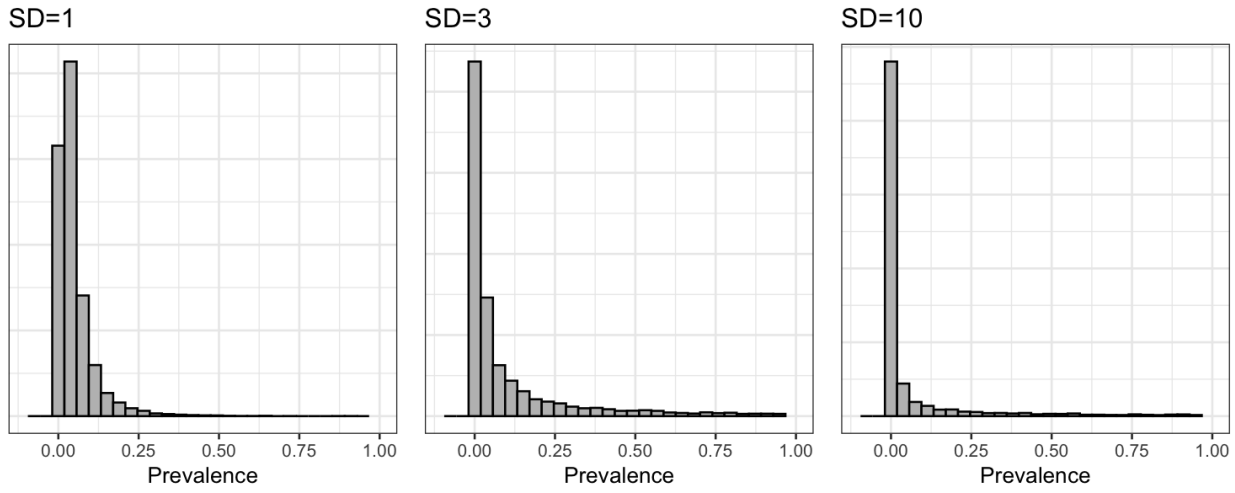


Figure A3. Exponential transformations of Normal distributions with mean -3.5 and variances 1^2, 3^2, and 10^2, respectively.

The prior on the overdispersion parameter, Exp(1), is chosen to reflect the belief that the variance of the outcome is likely to be less than two times its mean (with 63% probability), and that its more than four times the mean with fairly low probability, (5%). This is a reasonable assumption, because if the overdispersion is much larger than 1, the increased variability implies that the outcome is not rare in some areas. To test the sensitivity of

this prior choice we fit regression model (6) to the Zambia DHS data using the chosen prior and using an alternative prior, Exp(0.25), which puts much less weight near 0. Figure A4 shows that the resulting posterior distributions of the overdispersion parameter are nearly identical.

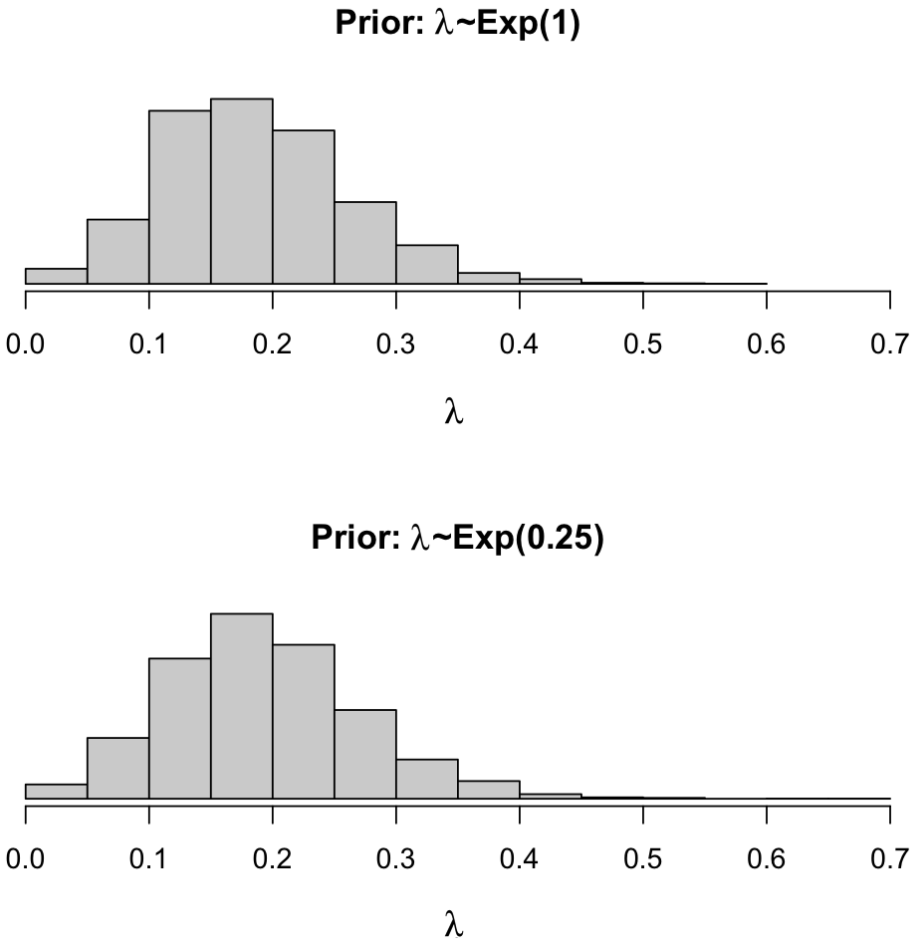


Figure A4. Posterior distribution of the overdispersion parameter in the nested negative binomial model, fit to Zambia neonatal mortality data, under two different choices of prior.

A4 Coefficient of variation for Zambia NMR estimates in motivating example

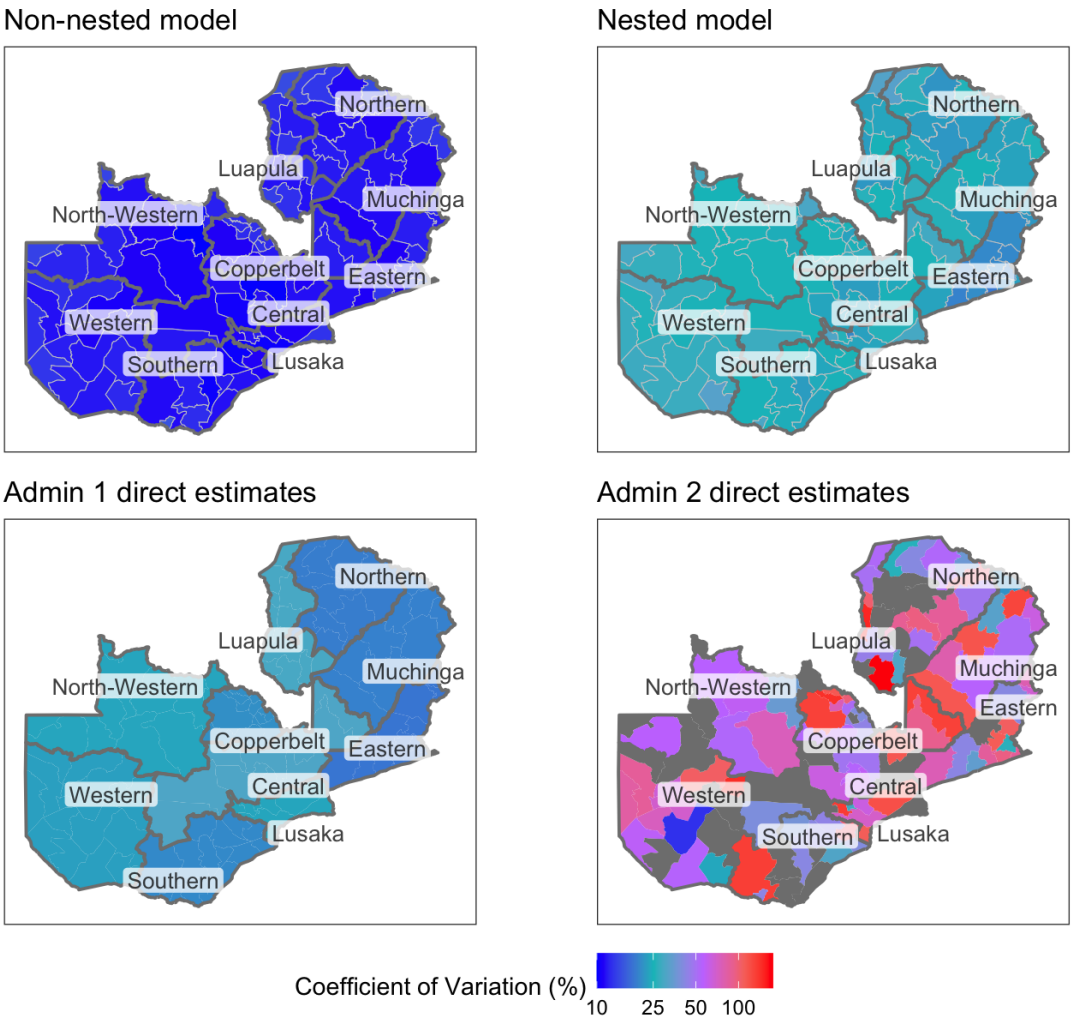


Figure A5. The maps on the top row display coefficients of variation for the NMR estimates under unit-level Bayesian negative binomial models with BYM2 spatial effects at the admin2 level, while the maps on the bottom row display coefficients of variation for the Hájek direct estimates at the admin1 and admin2 level, respectively. The range of coefficients of variation for the non-nested and nested model are $10-16\%$ and $18-32\%$, respectively, while the range of coefficients of variation for the admin1- and admin2-level direct estimates are $18-32\%$ and $13-173\%$, respectively. The latter reflects the extremely high uncertainty of the admin2-level direct estimates, even among those areas which have sufficient data to make valid estimates.

A5 Additional simulation results

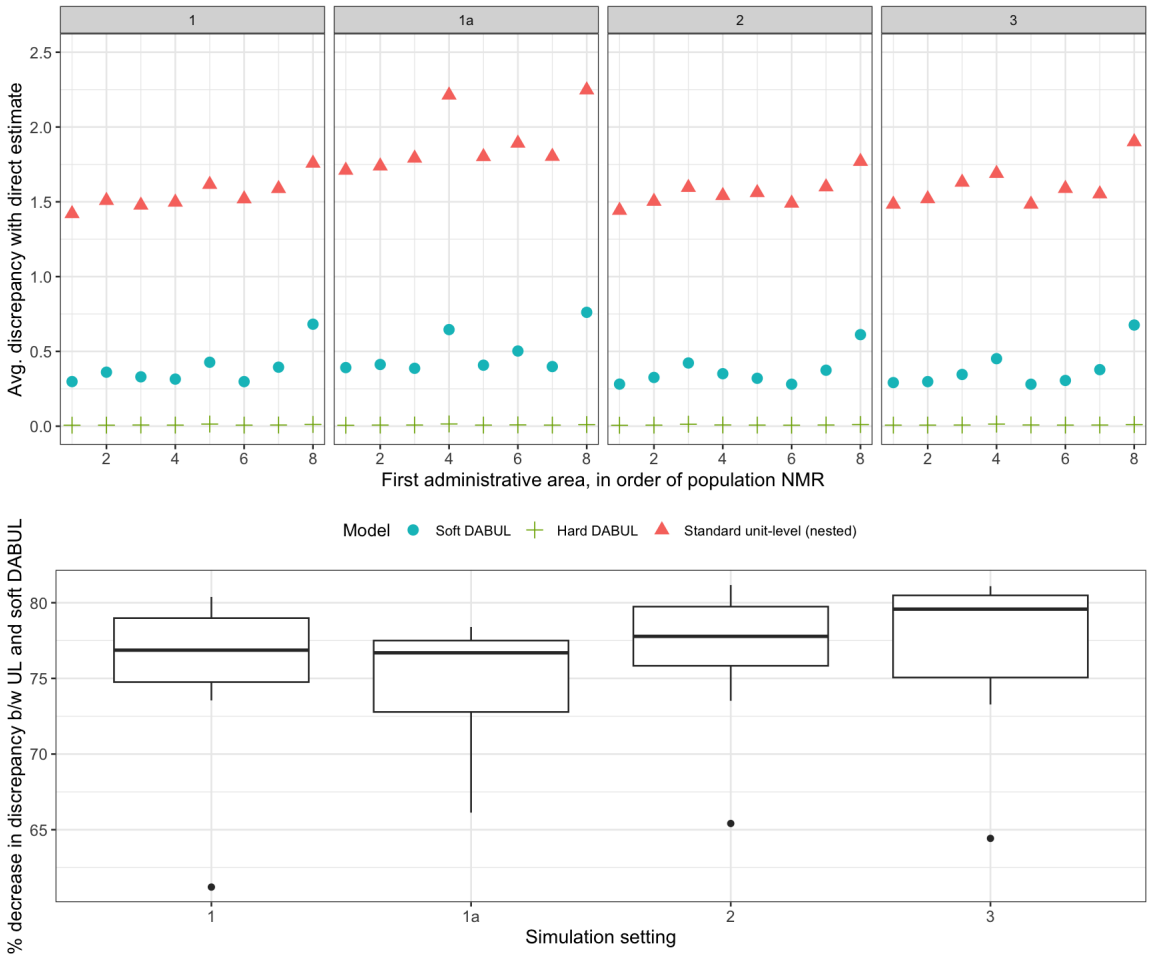


Figure A6. Discrepancy between aggregated model-based admin2-level estimates and design-based admin1-level estimates, among the subset of simulations for which the aggregated standard unit-level model estimate has a discrepancy with the direct estimate that is larger than 0.001.

Figure A6 is similar to Figure 2.6, except we only examine the areas for which the standard unit-level model has significant discrepancy with the direct estimates (greater than 0.001). We observe similar patterns to those in the previous figure, but the decrease in discrepancy, displayed in the bottom panel, is more significant, with aggregated soft DABUL estimates having $74.7 - 76.8\%$ lower discrepancy with the admin1-level direct estimates, on average, compared to aggregated estimates from the standard unit-level model. In other

words, when there is larger discrepancy between the aggregated standard unit-level estimates and direct estimates, the soft DABUL model reduces the discrepancy with direct estimates by a larger magnitude.

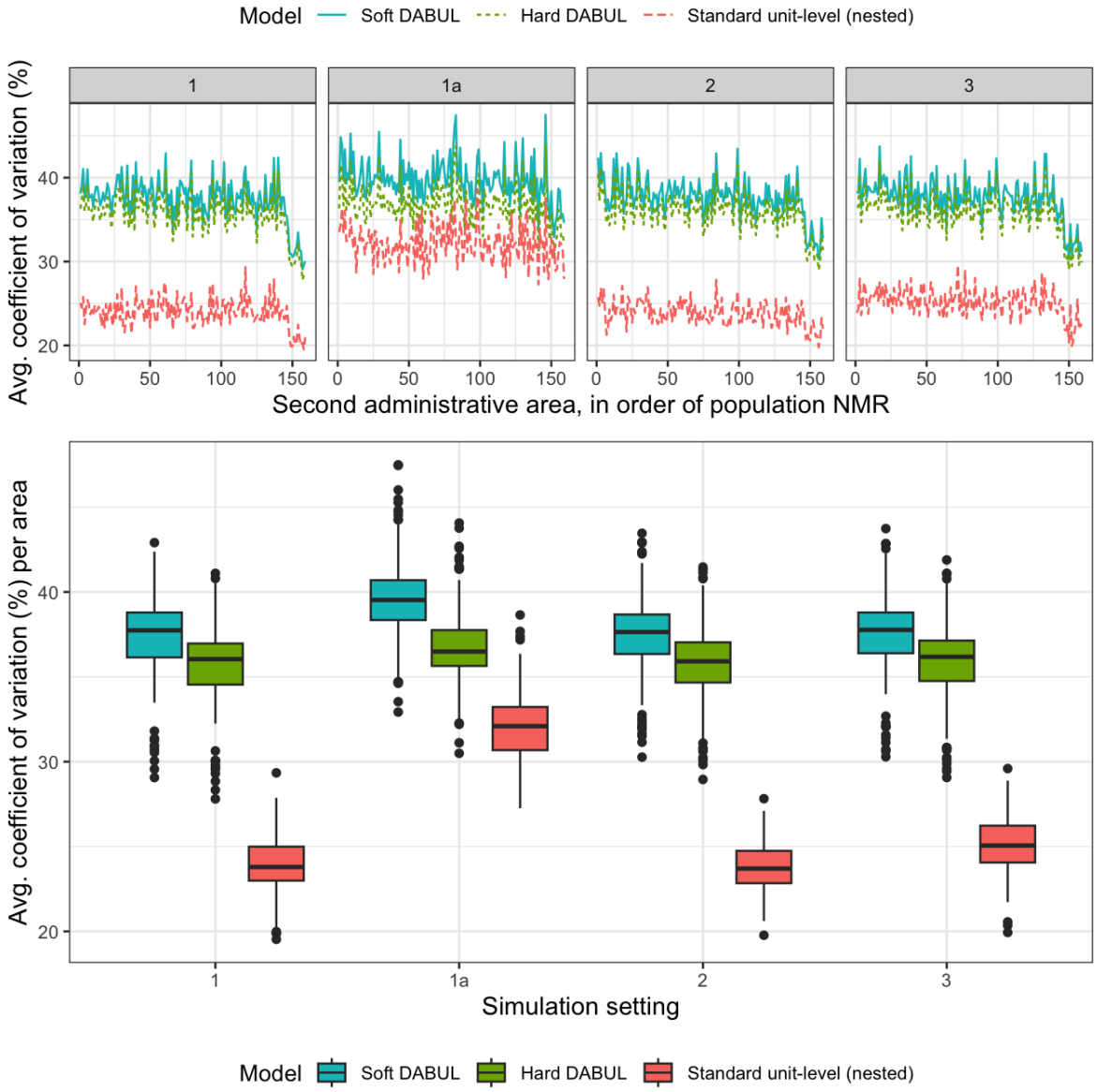


Figure A7. Coefficient of variation for admin2-level estimates, across 500 simulations for each of 4 settings.

A6 Additional plot of Zambia NMR estimates

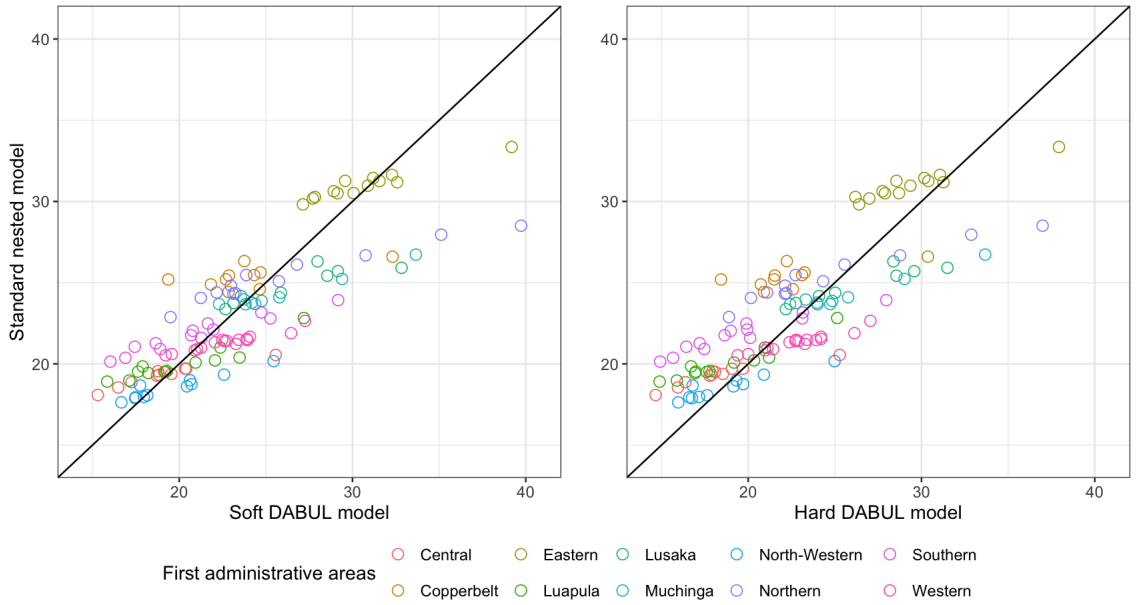


Figure A8. NMR estimates (deaths per 1000 live births) at the admin2-level in Zambia (2009-2013) using the DABUL models, as compared to a standard unit-level nested Bayesian model, both with BYM2 spatial effects at the admin2-level nested based on admin1 level.

Appendix B

APPENDIX TO CHAPTER 3

B1 Definitions and notation

Set	Size	Definition
$\mathcal{H}$	H	sampling strata
$\mathcal{H}_i$	H_i	sampling strata which span area i
U_h	M_h	clusters in strata $h \in \mathcal{H}$
U_{hi}	M_{hi}	subset of U_h, where clusters are contained in area i
S_h	m_h	sampled clusters in strata $h \in \mathcal{H}$
S_{hi}	m_{hi}	subset of S_h, where clusters are contained in area i

Element	Definition	Corresponding term(s)
$m_{\cdot i}$	number of sampled clusters in area i	
N_c	number of individuals in cluster c	
n_c	number of sampled individuals in cluster c	$\mathbf{D_i} = \text{diag}\{n_c : c \in \cup_{h\in\mathcal{H}_i} S_h\}$
w_c	sampling weight for individuals in cluster c	
$w_c^\star$	$n_c w_c$	$\boldsymbol{w_h^\star} = \{w_c^\star : c \in S_h\}$ $\boldsymbol{w_i^\star} = \{w_c^\star : c \in \cup_{h\in\mathcal{H}_i} S_h\}$ $\boldsymbol{W_i} = \text{diag}(\boldsymbol{w_i^\star})$ $\mathbf{T_i} = \mathbf{W_i}\left(\mathbf{I}_{m_{\cdot i}} - \frac{\mathbf{1}\boldsymbol{w_i}^{\star\mathrm{T}}}{\mathbf{1}^\mathrm{T}\boldsymbol{w_i}^\star}\right)$
$\mathbf{B}_h$	$\frac{m_h}{m_h-1}\left(\mathbf{I}_{m_h} - \frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^\mathrm{T}\right)$	$\mathbf{B}_i = \text{blockdiag}\{\mathbf{B}_h : h \in \mathcal{H}_i\}$
$\mathbf{M}_i$	$\mathbf{T}_i^\mathrm{T}\mathbf{B_i}\mathbf{T}_i$	
$\mathbf{q}_i$	nonzero eigenvalues of $\mathbf{D}_i^{1/2}\mathbf{M}_i\mathbf{D}_i^{1/2}$	
$\boldsymbol{v}_{ij}$	eiegenvector corresponding to eigenvalue q_{ij}	
δ_{ij}	noncentrality parameter, $\frac{1}{\sigma_i^2}\left(\boldsymbol{v_{ij}}^\mathrm{T}\boldsymbol{\gamma_i}\right)^2$	
$y_{c\ell}$	outcome for individual ℓ in cluster c	
ζ_{cl}	indicator variable denoting whether individual ℓ in cluster c was sampled	
$\bar{y}_c$	cluster-level sample mean, $\left(\sum_\ell^{N_c}\zeta_{c\ell}y_{c\ell}\right)/n_c$	$\boldsymbol{\bar{y}_h} = \{\bar{y}_c : c \in S_h\}$ $\boldsymbol{\bar{y}_i} = \{\bar{y}_c : c \in S_h, h \in \mathcal{H}_i\}$
θ_i	superpopulation mean for area i	
γ_h	stratum h mean effect	$\boldsymbol{\gamma_i} = \{\gamma_h \mathbf{1}_{m_h} : h \in \mathcal{H}_i\}$
θ_{hi}	stratum-specific mean for area i, $\theta_{hi} = \theta_i + \gamma_h$	
$\hat{\theta}_i$	Hájek estimator for θ_i	
σ_i^2	within-stratum population variance for area i	
V_i	design variance, $\text{Var}(\hat{\theta}_i \mid \theta)$	
$\hat{V}_i$	design-based estimator for V_i	

B2 Form of the design variance estimator

B2.1 Decomposition of summations over clusters inside and outside the domain

Consider an area of interest, i, which is either a planned domain (i.e. $S_{hi} = S_h$ for all $h \in \mathcal{H}_i$) or an unplanned domain (i.e. $S_{hi} \subsetneq S_h$, for some $h \in \mathcal{H}_i$). For all $c \in S_h \setminus S_{hi}$ and $h \in \mathcal{H}_i$, let $w_c^\star = 0$, as these clusters are outside the area of interest. Following Equation 2.3-10 of Korn and Graubard (2011), we apply the Taylor linearization method to expression (19) and obtain the design-consistent estimator for V_i expressed in (20),

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_h} w_c^\star\right)^2} \sum_{h\in\mathcal{H}_i} \frac{m_h}{m_h-1} \sum_{c\in S_h} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_h} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 .$$

To more clearly express the variance contributions of the clusters inside and outside of the area of interest, we can write (20) in terms of only its nonzero elements,

$$\begin{aligned}
\hat{V}_i &= \frac{1}{(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star)^2} \sum_{h\in\mathcal{H}_i} \Bigg\{ \frac{m_h}{m_h-1} \sum_{c\in S_{hi}} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \\
&\qquad + \frac{m_h}{m_h-1} \sum_{c\in S_h\setminus S_{hi}} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \Bigg\} \\
&= \frac{1}{(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star)^2} \sum_{h\in\mathcal{H}_i} \Bigg\{ \frac{m_h}{m_h-1} \sum_{c\in S_{hi}} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \\
&\qquad + \frac{m_h}{m_h-1} \sum_{c\in S_h\setminus S_{hi}} \left[\frac{1}{m_h}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \Bigg\} \\
&= \frac{1}{(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star)^2} \sum_{h\in\mathcal{H}_i} \Bigg\{ \frac{m_h}{m_h-1} \sum_{c\in S_{hi}} \left[w_c^\star(\bar{y}_c - \hat{\theta}_i) - \frac{1}{m_h}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \\
&\qquad + \frac{m_h - m_{hi}}{(m_h-1)m_h} \left[\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'} - \hat{\theta}_i)\right]^2 \Bigg\}
\end{aligned}$$

where the first term inside the larger summation is a sum over the clusters inside the area of interest and the second term is a sum over the clusters outside the area of interest. From this

expression we observe that when the area of interest is an unplanned domain, the number of clusters in the larger planned domain is incorporated in the last term, thus accounting for the additional variability introduced by the fact that the number of sampled clusters in the unplanned domain is random. Moreover, if the area of interest is a planned domain ($m_h = m_{hi}$), the last term becomes zero because the number of sampled clusters is fixed, and the estimator simplifies to

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star\right)^2}\sum_{h\in\mathcal{H}_i}\frac{m_{hi}}{m_{hi}-1}\sum_{c\in S_{hi}}\left[w_c^\star(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'}-\hat{\theta}_i)\right]^2.$$

B2.2 Form of the variance estimator under simple sampling distribution assumptions (Equation 25)

In this section we show that the design variance estimator presented in (20),

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_h} w_c^\star\right)^2}\sum_{h\in\mathcal{H}_i}\frac{m_h}{m_h-1}\sum_{c\in S_h}\left[w_c^\star(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_h}\sum_{c'\in S_h} w_{c'}^\star(\bar{y}_{c'}-\hat{\theta}_i)\right]^2,$$

simplifies to the estimator presented in (25),

$$\hat{V}_i = \frac{1}{m_{\cdot i}(m_{\cdot i}-|\mathcal{H}_i|)}\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}}\left[\bar{y}_c-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}}\bar{y}_{c'}\right]^2,$$

under the design assumptions presented in Section 3.3.1.

Under assumption (i), the area of interest is a planned domain, so (20) can be written as

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star\right)^2}\sum_{h\in\mathcal{H}_i}\frac{m_{hi}}{m_{hi}-1}\sum_{c\in S_{hi}}\left[w_c^\star(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'}-\hat{\theta}_i)\right]^2.$$

Then under assumption (ii), the same number of clusters are sampled from each stratum in

area i (i.e., $m_{hi} = m_{\cdot i}/|\mathcal{H}_i|$) so it follows

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}} w_c^\star\right)^2}\frac{m_{\cdot i}}{m_{\cdot i}-|\mathcal{H}_i|}\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}}\left[w_c^\star(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}} w_{c'}^\star(\bar{y}_{c'}-\hat{\theta}_i)\right]^2.$$

Then under assumptions (iii), (iv), and (v) $w_c^\star = w_0^\star$ for all $c \in \cup_{h\in\mathcal{H}_i} S_{hi}$, so it follows,

$$\begin{aligned}
\hat{V}_i &= \frac{w^{\star 2}}{\left(m_{\cdot i}w^\star\right)^2}\frac{m_{\cdot i}}{m_{\cdot i}-|\mathcal{H}_i|}\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}}\left[\left(\bar{y}_c-\hat{\theta}_i\right)-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}}\left(\bar{y}_{c'}-\hat{\theta}_i\right)\right]^2\\
&= \frac{1}{m_{\cdot i}^2}\frac{m_{\cdot i}}{m_{\cdot i}-|\mathcal{H}_i|}\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}}\left[\left(\bar{y}_c-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}}\bar{y}_{c'}\right)-\left(1-\frac{m_{hi}}{m_{hi}}\right)\hat{\theta}_i\right]^2\\
&= \frac{1}{m_{\cdot i}(m_{\cdot i}-|\mathcal{H}_i|)}\sum_{h\in\mathcal{H}_i}\sum_{c\in S_{hi}}\left[\bar{y}_c-\frac{1}{m_{hi}}\sum_{c'\in S_{hi}}\bar{y}_{c'}\right]^2.
\end{aligned}$$

B2.3 Matrix form of the variance estimator (Equation 29)

In this section we show that the variance estimator (20),

$$\hat{V}_i = \frac{1}{\left(\sum_{h\in\mathcal{H}_i}\sum_{c\in S_h} w_c^\star\right)^2}\sum_{h\in\mathcal{H}_i}\frac{m_h}{m_h-1}\sum_{c\in S_h}\left[w_c^\star(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_h}\sum_{c'\in S_h} w_{c'}^\star(\bar{y}_{c'}-\hat{\theta}_i)\right]^2$$

can be written in the form of expression (29),

$$\hat{V}_i = \frac{1}{\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star\right)^2}\bar{\boldsymbol{y}}_i^{\mathrm{T}}\mathbf{M_i}\bar{\boldsymbol{y}}_i.$$

Recall from Section 3.3.2, $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ is composed of subvectors, $(\bar{\boldsymbol{y}}_{\boldsymbol{i},h})_{h\in\mathcal{H}_i}$, with elements $(\bar{y}_c)_{c\in S_h}$ and $\mathbf{M_i} = \mathbf{T_i}^{\mathrm{T}}\mathbf{B_i}\mathbf{T_i}$. Let $\mathbf{B_i}$ be a block diagonal matrix with blocks $(\mathbf{B}_h)_{h\in\mathcal{H}_i}$ such that $\mathbf{B}_h = \frac{m_h}{m_h-1}\left(\mathbf{I}_{m_h}-\frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)$. Let $\boldsymbol{w_i^\star}$ be a vector composed of subvectors, $(\boldsymbol{w}_h^\star)_{h\in\mathcal{H}_i}$, with elements $\{w_c^\star\}_{c\in S_h}$ and define $\mathbf{T_i} = \mathbf{W_i}\left(\mathbf{I}_{m_{\cdot i}}-\frac{\mathbf{1}\boldsymbol{w_i}^{\star\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star}\right)$, where $\mathbf{W_i} = \mathrm{diag}\left(\boldsymbol{w_i}^\star\right)$. Note that $(\bar{\boldsymbol{y}}_i, \mathbf{B_i}, \boldsymbol{w_i^\star})$ must maintain consistent ordering. By definition, $\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star = \sum_{h\in\mathcal{H}_i}\sum_{c\in S_h} w_c^\star$, so

to prove equality it suffices to show

$$(\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}})^{\mathrm{T}}\,\mathbf{B}_{\boldsymbol{i}}\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{h\in\mathcal{H}_i}\frac{m_h}{m_h-1}\sum_{c\in S_h}\left[w_c^{\star}(\bar{y}_c-\hat{\theta}_i)-\frac{1}{m_h}\sum_{c'\in S_h}w_{c'}^{\star}(\bar{y}_{c'}-\hat{\theta}_i)\right]^2. \qquad (59)$$

By definition, $\mathbf{T}_{\boldsymbol{i}}$ is a square matrix of dimension $\sum_{h\in\mathcal{H}_i}m_h$ with diagonal elements, $\mathbf{T}_{\boldsymbol{i},cc}=w_c^{\star}-\frac{w_c^{\star 2}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w}_{\boldsymbol{i}}{}^{\star}}$, and off-diagonal elements, $\mathbf{T}_{\boldsymbol{i},cc'}=-\frac{w_c^{\star}w_{c'}^{\star}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w}_{\boldsymbol{i}}{}^{\star}}$ for $(c,c')\in S=\cup_{h\in\mathcal{H}_i}S_h$. Then it follows that $\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ is a vector composed of subvectors, $(\boldsymbol{L}_h)_{h\in\mathcal{H}_i}$, where $\boldsymbol{L}_h$ is a vector of length m_h with elements

$$\boldsymbol{L}_{h,c}=w_c^{\star}\bar{y}_c-w_c^{\star}\sum_{c'\in S}\frac{w_{c'}^{\star}\bar{y}_{c'}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w}_{\boldsymbol{i}}{}^{\star}}=w_c^{\star}(\bar{y}_c-\hat{\theta}_i)$$

for $c\in S_h$, and

$$\begin{aligned}(\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}})^{\mathrm{T}}\mathbf{B}_{\boldsymbol{i}}\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}} &= \sum_{h\in\mathcal{H}_i}\boldsymbol{L}_h^{\mathrm{T}}\mathbf{B}_h\boldsymbol{L}_h\\ &= \sum_{h\in\mathcal{H}_i}\frac{m_h}{m_h-1}\boldsymbol{L}_h^{\mathrm{T}}\left(\mathbf{I}_{m_h}-\frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)\boldsymbol{L}_h.\end{aligned}$$

Observe that the matrix $\left(\mathbf{I}_{m_h}-\frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)$ is idempotent, so it is equal to the square of itself, and $\left(\mathbf{I}_{m_h}-\frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)\boldsymbol{L}_h$ is an m_h-length vector with elements

$$\left[\left(\mathbf{I}_{m_h}-\frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)\boldsymbol{L}_h\right]_c=\boldsymbol{L}_{h,c}-\frac{1}{m_h}\sum_{c'\in S_h}\boldsymbol{L}_{h,c'}$$

for $c\in S_h$, so it follows that

$$(\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}})^{\mathrm{T}}\mathbf{B}_{\boldsymbol{i}}\mathbf{T}_{\boldsymbol{i}}\bar{\boldsymbol{y}}_{\boldsymbol{i}}=\sum_{h\in\mathcal{H}_i}\frac{m_h}{m_h-1}\sum_{c\in S_h}\left(\boldsymbol{L}_{h,c}-\frac{1}{m_h}\sum_{c'\in S_h}\boldsymbol{L}_{h,c'}\right)^2$$

which results in the equality in expression (59).

B3 Accuracy of Satterthwhaite approximation

In this section we evaluate how well the Satterthwhaite approximation (31) approximates the survey-weighted sampling distribution (30). We find that the approximation is highly accurate using sampling weights and sample sizes from various admin2 areas in the Kenya 2022 DHS and different sets of parameters.

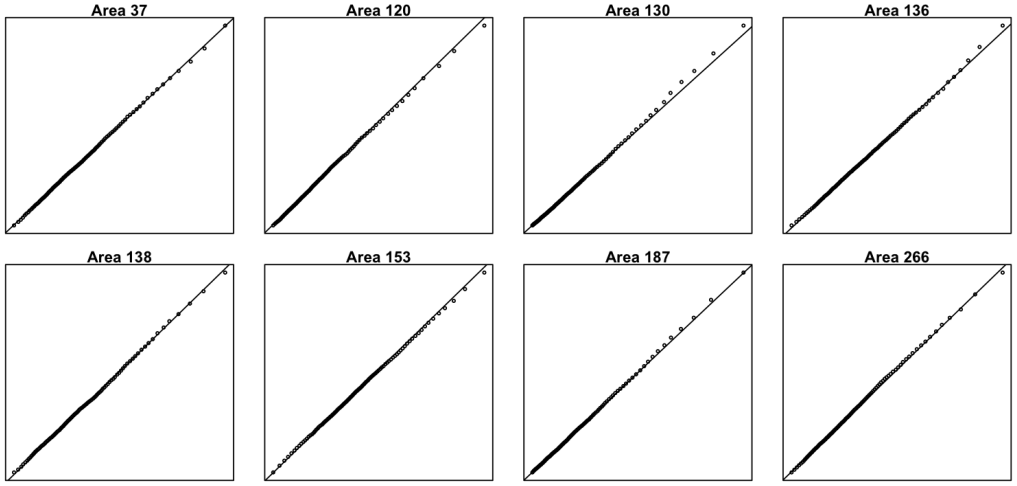


Figure B1. Quantile-quantile plots comparing survey-weighted sampling distribution, on the x-axis, to its Satterthwhaite approximation, on the y-axis, when $\gamma = 1$ and $\sigma^2 = 1$, using the sampling weights and sample sizes from 8 admin2 areas in the 2022 Kenya DHS.

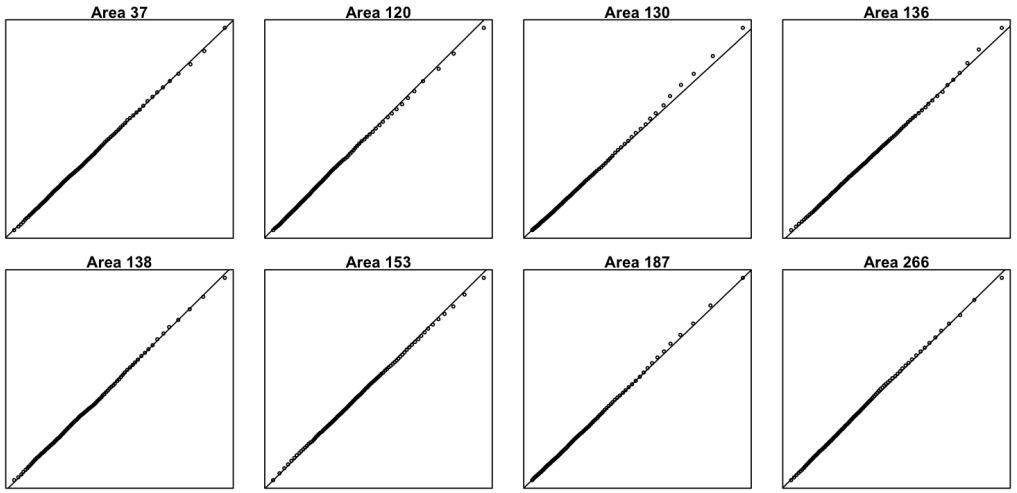


Figure B2. Quantile-quantile plots comparing survey-weighted sampling distribution, on the x-axis, to its Satterthwhaite approximation, on the y-axis, when $\gamma = 2$ and $\sigma^2 = 4$, using the sampling weights and sample sizes from 8 admin2 areas in the 2022 Kenya DHS.

B4 Bias of design variance estimator under SASW sampling distribution (Equation 31)

Recall from Section 3.3.2, $\mathbf{D_i}$ is a a diagonal matrix with elements $(1/n_c)_{c\in\cup_{h\in\mathcal{H}_i}S_h}$, $\boldsymbol{w_i}^\star$ is a vector composed of subvectors, $(\boldsymbol{w}_h^\star)_{h\in\mathcal{H}_i}$, with elements $(w_c^\star)_{c\in S_h}$, and $\mathbf{M_i} = \mathbf{T_i}^{\mathrm{T}}\mathbf{B_i}\mathbf{T_i}$ where $\mathbf{T_i} = \mathrm{diag}\left(\boldsymbol{w_i}^\star\right)\left(\mathbf{I}_{m_{\cdot i}} - \frac{\mathbf{1}\boldsymbol{w_i}^{\star\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{w_i}^\star}\right)$ and $\mathbf{B_i}$ is a block diagonal matrix with blocks $(\mathbf{B}_h)_{h\in\mathcal{H}_i}$ such that $\mathbf{B}_h = \frac{m_h}{m_h-1}\left(\mathbf{I}_{m_h} - \frac{1}{m_h}\mathbf{1}_{m_h}\mathbf{1}_{m_h}^{\mathrm{T}}\right)$. Also recall, $\boldsymbol{q_i}$ is the vector of non-zero eigenvalues of $\mathbf{D_i}^{1/2}\mathbf{M_i}\mathbf{D_i}^{1/2}$ and $\delta_{ij} = \frac{1}{\sigma_i^2}\left(\boldsymbol{v_{ij}}^{\mathrm{T}}\boldsymbol{\gamma_i}\right)^2$, where $\boldsymbol{v_{ij}}$ denotes the $m_{\cdot i}-$length eigenvector corresponding to eigenvalue q_{ij}, scaled such that $\boldsymbol{v_{ij}}^{\mathrm{T}}\mathbf{D_i}\boldsymbol{v_{ij}} = 1$ and $\boldsymbol{\gamma_i}$ is a vector composed of subvectors, $(\gamma_h\mathbf{1}_{m_h})_{h\in\mathcal{H}_i}$. The SASW sampling distribution (31) is defined as

$$\hat{V}_i \mid \sigma_i^2, \boldsymbol{\gamma_i} \sim \frac{V_i^\star(\sigma_i^2)}{\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star}\frac{Q_2}{2Q_1}\chi^2_{\frac{2Q_1^2}{Q_2}},$$

where $Q_1 = \sum_j^{r_i} q_{ij}(1+\delta_{ij})$ and $Q_2 = 2\sum_j^{r_i} q_{ij}^2(1+2\delta_{ij})$.

Then it follows that

$$\mathbb{E}[\hat{V}_i \mid \sigma_i^2, \boldsymbol{\gamma_i}] = \frac{Q_1}{\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star}V_i^\star(\sigma_i^2)$$

so $\hat{V}_i \mid \sigma_i^2, \boldsymbol{\gamma_i}$ is a biased estimator for the theoretical design variance, $V_i^\star(\sigma_i^2)$, by a factor of $Q_1/(\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star)$. We will provide conditions under which $Q_1/(\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star) < 1$.

From the definitions of $\boldsymbol{q_i}$ and $\boldsymbol{\delta_i}$ it follows that

$$Q_1 = \sum_j q_{ij} + \sum_j q_{ij}\delta_{ij} = \mathrm{tr}(\mathbf{M_i}\mathbf{D_i}) + \frac{1}{\sigma_i^2}\boldsymbol{\gamma_i}^{\mathrm{T}}\mathbf{M_i}\mathbf{D_i}\mathbf{M_i}\boldsymbol{\gamma_i}.$$

The first term, $\mathrm{tr}(\mathbf{M_i}\mathbf{D_i})$, can be expressed as $\boldsymbol{w_i}^{\star\mathrm{T}}\mathbf{D_i}\boldsymbol{w_i}^\star - R_i$, where

$$R_i = \sum_{h \in \mathcal{H}_i} \frac{m_h}{m_h - 1} \sum_{c \in S_h} \left\{ \left[\frac{w_c^\star}{\sqrt{n_c}} - \frac{1}{m_h} \sum_{c' \in S_h} \frac{w_{c'}^\star}{\sqrt{n_{c'}}} \right] \right.$$
$$\left. - \frac{\sum_{h \in \mathcal{H}_i} \sum_{c' \in S_h} w_{c'}^\star / n_{c'}}{\sum_{h \in \mathcal{H}_i} \sum_{c' \in S_h} w_{c'}^\star} \left[w_c^\star - \frac{1}{m_h} \sum_{c' \in S_h} w_{c'}^\star \right]^2 \right\},$$

so

$$\frac{Q_1}{\boldsymbol{w_i}^{\star \mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^\star} = \frac{\boldsymbol{w_i}^{\star \mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^\star - R_i + \frac{1}{\sigma_i^2} \boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i}}{\boldsymbol{w_i}^{\star \mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^\star} = 1 + \frac{\frac{1}{\sigma_i^2} \boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i} - R_i}{\boldsymbol{w_i}^{\star \mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^\star}.$$

Then it follows that $Q_1/(\boldsymbol{w_i}^{\star \mathrm{T}} \mathbf{D_i} \boldsymbol{w_i}^\star) < 1$ when $R_i > \boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i} / \sigma_i^2$. Note that both R_i and $\boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i} / \sigma_i^2$ are nonnegative. The second term, $\boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i} / \sigma_i^2$, can be expressed as

$$\left(\frac{m_h}{m_h - 1} \right)^2 \left(\frac{\gamma_h - \bar{\gamma}^W}{\sigma_i} \right)^2 \sum_{c \in S_h} \frac{1}{n_c} \left(w_c^\star - \frac{1}{m_h} \sum_{c' \in S_h} w_{c'}^\star \right)^2$$

where $\bar{\gamma}^W = \frac{\sum_h \gamma_h \sum_{c \in S_h} w_c^\star}{\sum_h \sum_{c \in S_h} w_c^\star}$. We observe that $\boldsymbol{\gamma_i}^{\mathrm{T}} \mathbf{M_i} \mathbf{D_i} \mathbf{M_i} \boldsymbol{\gamma_i} / \sigma_i^2$ tends to zero when $w_c^\star$ takes a similar value for all $c \in S_h$, which is generally the case unless there are extreme differences at the cluster size re-enumeration stage. This term also decreases when the within-stratum variance is larger than the between-stratum variance. On the other hand, R_i only tends to zero when, in addition to $w_c^\star$, n_c also takes similar values for all $c \in S_h$, which is a less reasonable assumption (see Appendix B6).

B5 Additional simulation results

B5.1 Re-enumerated cluster sizes

In the simulation design described in Section 3.5.1 we assume the listed cluster sizes are correct and do not require re-enumeration (i.e. $L_c = N_c$). In this section we evaluate whether cluster re-enumeration, and the resulting difference in sampling weights, significantly modifies the results we observe in the main text. We use the same setup and parameters here as in setting 1, with the modification that each cluster c, is sampled with probability proportional size using L_c (instead of N_c) where L_c is generated from a normal distribution with mean $0.85N_c$ and standard deviation $0.2N_c$, rounded to the nearest whole number. The mean of L_c is chosen to reflect the fact that, due to population growth, we expect the listed cluster size to underestimate the true size, on average. In Figure B3 we observe that there are some very slight differences in model performance when cluster sizes are re-enumerated, namely the RMSE and coverage rates are more variable across areas, but all conclusions made in the main text also hold for this setting.

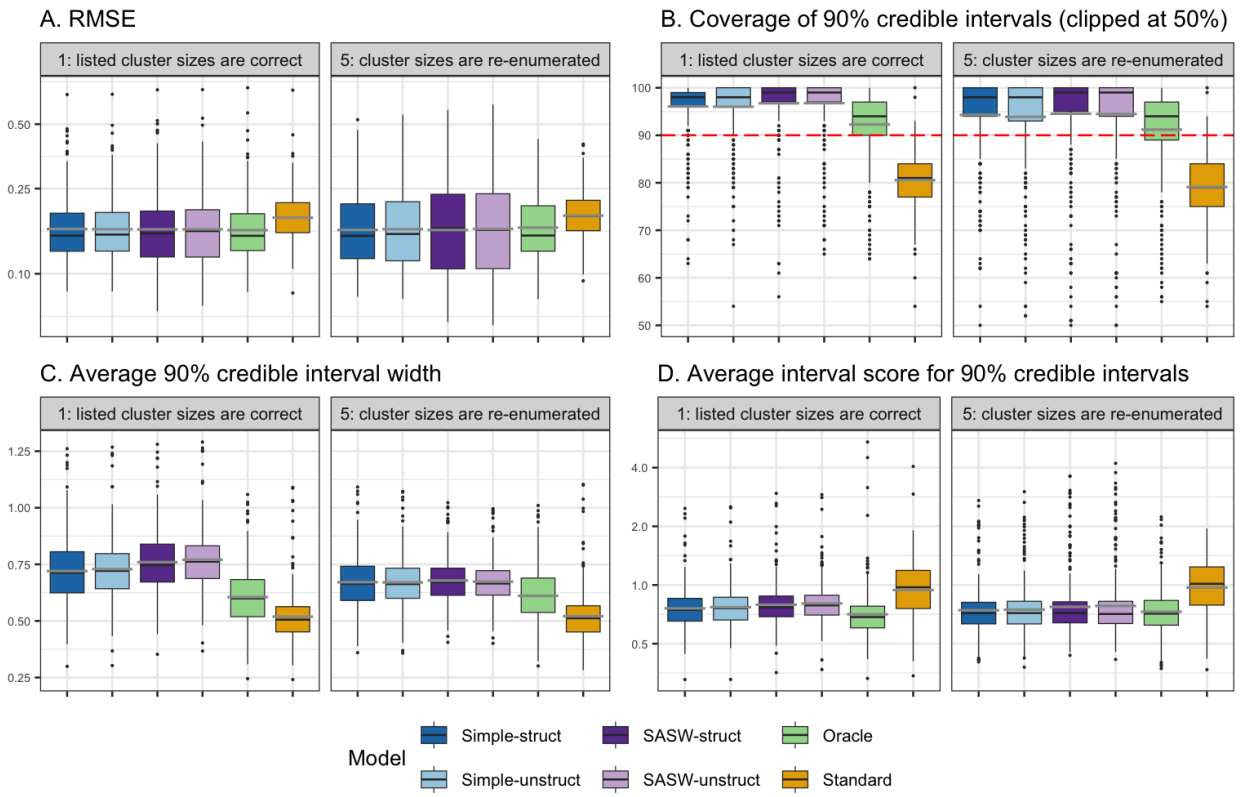


Figure B3. RMSE of area-level point estimates and coverage, average width, and average interval score of 90% credible intervals, for each area and model, across 100 simulations. The gray lines indicate the mean across areas and the red line in panel B indicates the nominal rate of the credible intervals.

B5.2 Comparison of sampling distributions

In this section we assess how closely the simple and SASW sampling distributions adhere to the empirical sampling distribution of the design variance estimator. We do not use model estimates of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\gamma}$, but simply plug in the correct values, which are defined in the simulation design, so that we may assess accuracy of these sampling distributions when parameters are correctly specified.

We use two metrics to evaluate how the simple and SASW sampling distributions compare to the empirical distribution of the design variance estimator, $\hat{V}_i$ for $i = 1, ..., K$, which we have obtained through simulation. To directly compare the sampling and empirical distributions we use the average Wasserstein distance of order 2, an empirical measure for differences between distributions (Sommerfeld and Munk, 2018; Bernton et al., 2019; Panaretos and Zemel, 2019). For each area i, we evaluate the average Wasserstein distance of order 2,

$$\frac{1}{|G_i|} \sum_{g \in G_i} \left\{ \frac{1}{99} \sum_{j=1}^{99} \left(\hat{Q}_i^{(g)}(j) - Q_i(j) \right)^2 \right\}$$

where $\hat{Q}_i^{(g)}(j)$ is the j^{th} percentile of the simple or SASW sampling distribution of $\hat{V}_i$ for dataset g and $Q_i(j)$ is the j^{th} percentile of the empirical distribution of $\hat{V}_i$. Note that the simple and SASW sampling distributions vary by dataset because they depend on the sampling weights and sample sizes of the sampled clusters. To determine the directionality of this difference in distributions, for each area i we evaluate the difference in means,

$$\frac{1}{|G_i|} \sum_{g \in G_i} \left\{ \hat{\mathbb{E}}^{(g)}[\hat{V}_i] - \mathbb{E}[\hat{V}_i] \right\},$$

where $\hat{\mathbb{E}}^{(g)}[\hat{V}_i]$ is the mean of $\hat{V}_i$ under the simple or SASW sampling distribution for dataset g and $\mathbb{E}[\hat{V}_i]$ is the mean of $\hat{V}_i$ under the empirical distribution.

In Figure B4A we observe that, on average, the SASW sampling distribution is a slightly better approximation of the empirical distribution of the design variance estimator. Both

sampling distributions are closer to the empirical distribution when sample size is larger and further when there is within-cluster correlation. In Figure B4B we observe that, on average the simple sampling distribution underestimates the mean of the design variance estimator, while the SASW sampling distribution estimates the mean fairly well, on average, except when there is within-cluster correlation. Poorer performance when there is within-cluster correlation is expected because the theoretical design variance is misspecified, as it is a function of both σ_i^2 and ρ.

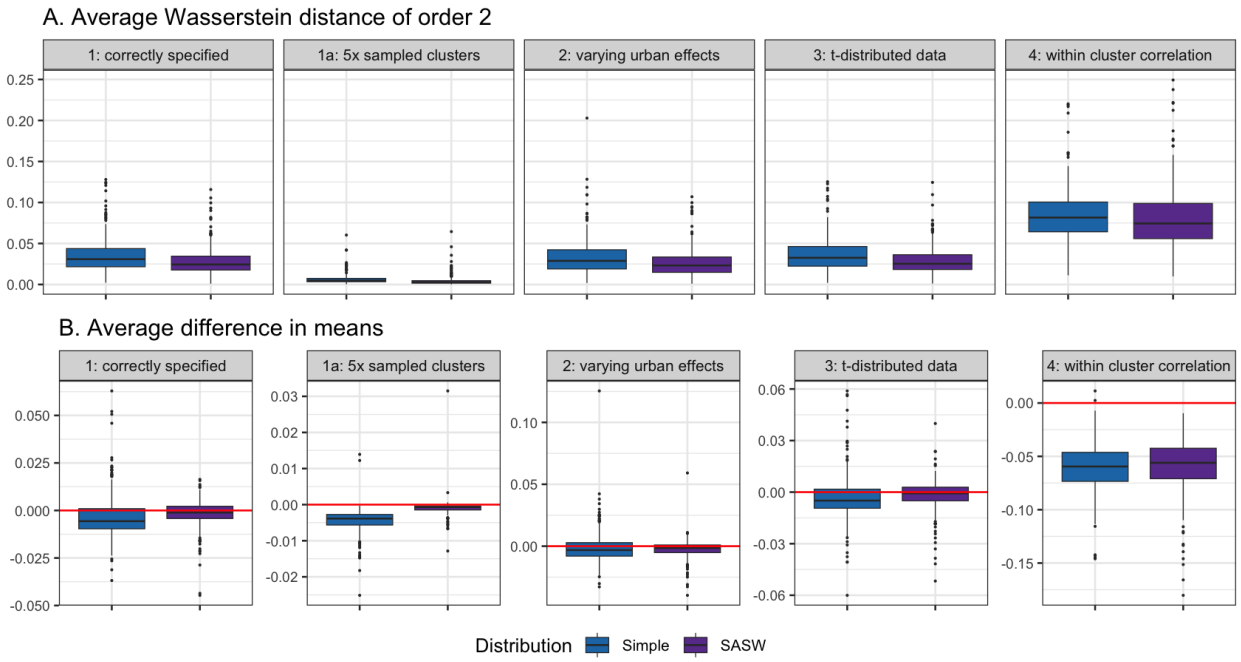


Figure B4. Comparison of simple and SASW sampling distributions to empirical distribution of the design variance estimator. Averages are computed for each area across 100 simulations for each of 5 settings. The red line in panel B indicates the point where the estimate is equal to the truth and values below the red line indicate underestimation.

B6 Distribution of cluster sizes in 2022 Kenya DHS

We present the distribution of cluster sample sizes for urban and rural clusters in the 2022 Kenya DHS, which we use to calibrate the sample sizes in our simulation study. We use a negative binomial distribution with the parameterization,

$$P(n_c|\text{size} = s, \text{mean} = \mu) = \binom{n_c + s - 1}{n_c} \left(\frac{\mu}{\mu + s}\right)^{n_c} \left(\frac{s}{\mu + s}\right)^{s}.$$

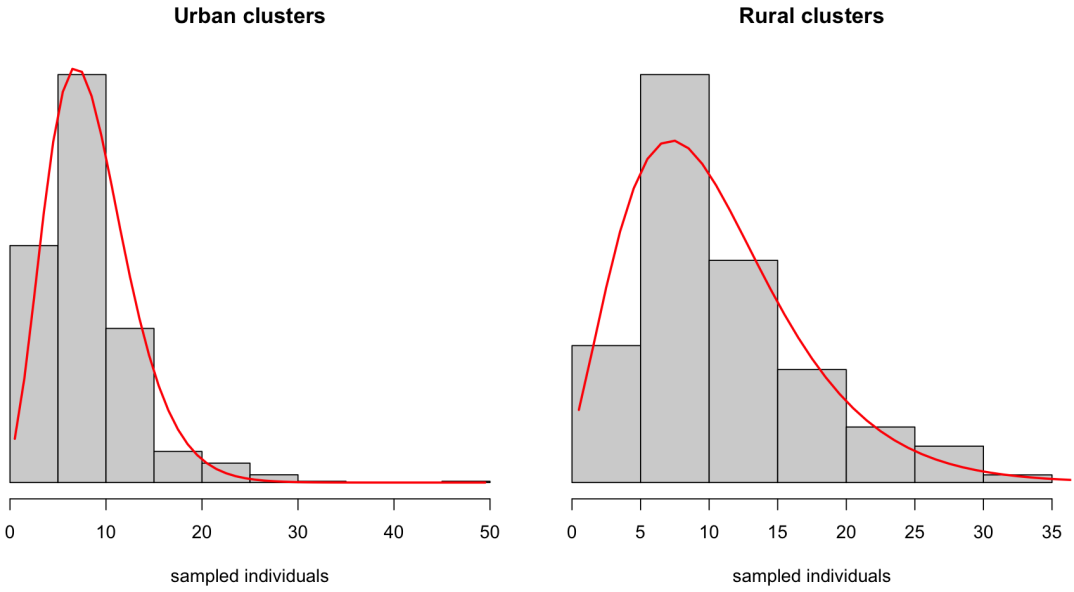


Figure B5. Distribution of urban and rural cluster sample sizes in the 2022 Kenya DHS. The urban cluster size distribution is compared to a negative binomial distribution, denoted by a red curve, with size = 8 and mean = 9. Similarly, the rural cluster size distribution is compared to a negative binomial distribution, with size = 4 and mean = 11.

B7 Auxiliary variables for Kenya example

For the area-specific auxiliary variables in the latent mean model we use the urban under-5 population proportion, population density, yearly average temperature, yearly total precipitation, elevation, motorized travel time to healthcare, and yearly average nighttime lights.

The urban under-5 population proportions are estimated using the method in Wu and Wakefield (2024), a logistic classification model which uses pixel-level population density obtained from WorldPop (Tatem, 2017) and defines a threshold for classifying pixels as urban or rural, which is calibrated by the urban/rural composition at the admin1 level reported by DHS (Corsi et al., 2012). High-resolution surfaces for yearly average temperature, elevation, and yearly total precipitation were obtained from WorldClim (Fick and Hijmans, 2017). High-resolution surfaces for yearly average nighttime lights and motorized travel time to healthcare was obtained from Earth Observation Group, NOAA National Centers for Environmental Information (2022) and Weiss et al. (2018), respectively. For each high-resolution surface, we obtained the area-specific auxiliary variable by taking the population-weighted average of the value over each area. We took the log transformation of variables with skewed distributions (population density, nighttime lights, and time to healthcare) and then standardized each variable (except for the urban proportion).

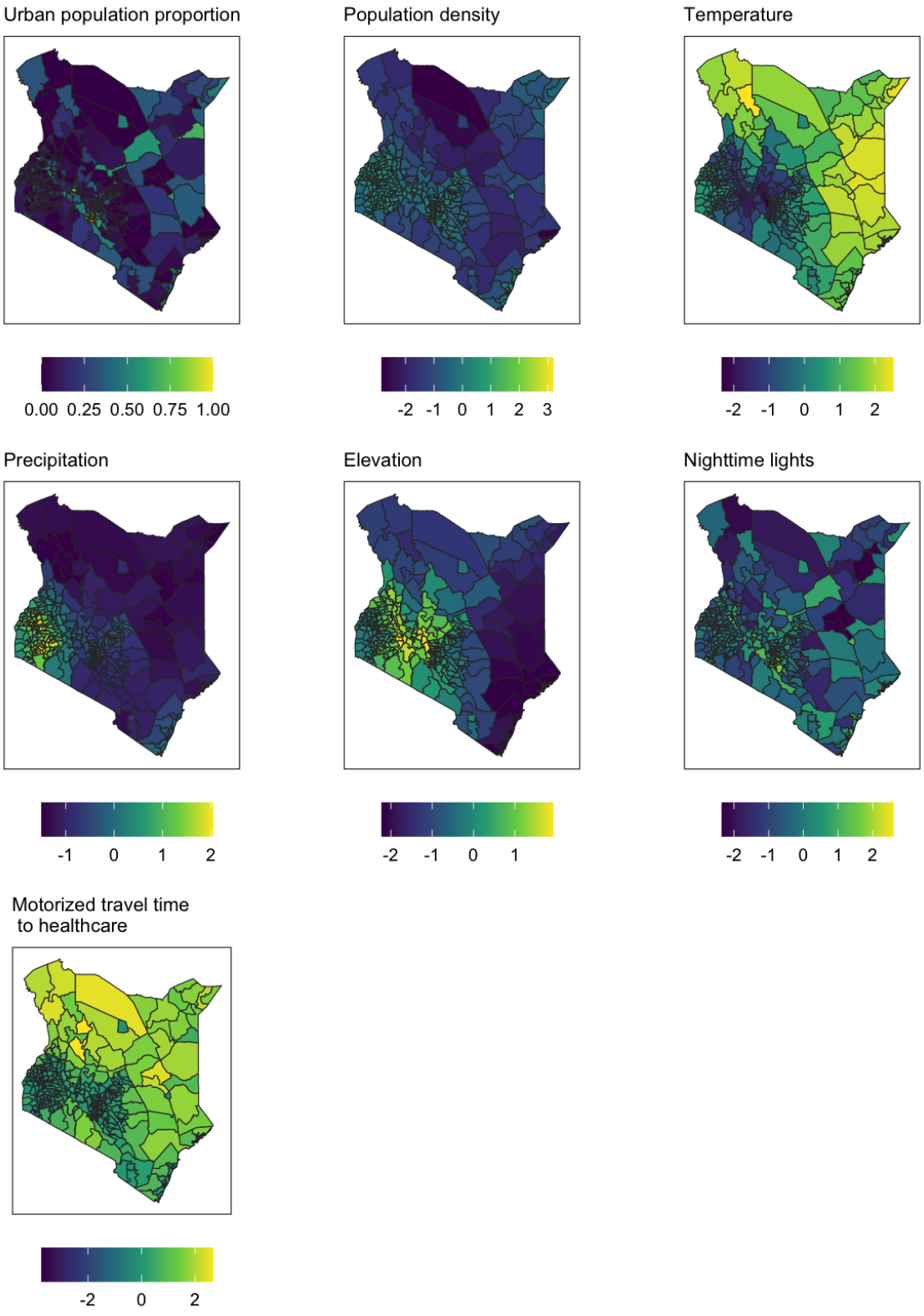


Figure B6. Maps of standardized admin2 area-level auxiliary variables in Kenya

B8 Scatter plots of Kenya HAZ estimates

In this section we compare the mean and standard deviation estimates of HAZ under each model to the design-based estimates. As in Section 3.6 we observe that both variance smoothing models exhibit more shrinkage of the mean and variance compared to the standard model. It is important to note that this shrinkage is not necessarily a negative characteristic, as the design-based estimates are quite noisy due to low sample size. The design-based standard deviation estimates, in particular, are not very reliable, as is noted throughout the main text and exhibited in Section 3.5.3 of the simulation study.

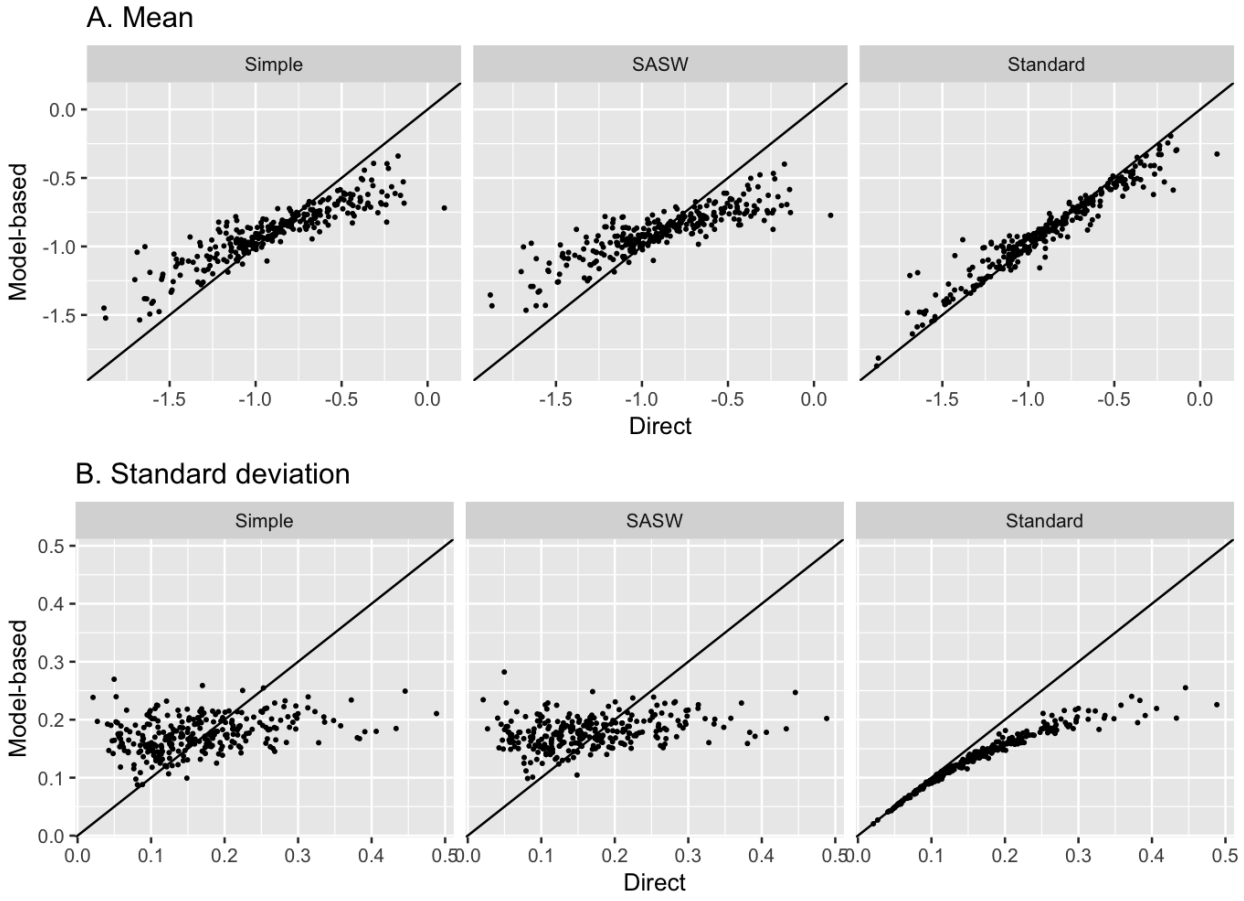


Figure B7. Scatter plots comparing mean and standard deviation estimates of HAZ under each model against the design-based estimates.

Appendix C

APPENDIX TO CHAPTER 4

C1 Derivation of random effect distribution under the SCM models

C1.1 MV-Sym

Let $\boldsymbol{z}^* = \left(\boldsymbol{z}_0^{\mathrm{T}}, \boldsymbol{z}_1^{\mathrm{T}}, ..., \boldsymbol{z}_R^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\boldsymbol{P}^* = \begin{pmatrix} \gamma_1 & 1 & 0 & ... & 0 \\ \gamma_2 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ \gamma_R & 0 & 0 & ... & 1 \end{pmatrix},$$

where $\boldsymbol{z}_k$ and γ_r are as defined as in (50). We will show this implies that the MV-Sym model (50) is a GMBYM2 model (47) with

$$\boldsymbol{\psi}_u = \boldsymbol{P}^* \boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}} \qquad \text{and} \qquad \boldsymbol{\psi}_v = \boldsymbol{P}^* (\mathbf{I}_{R+1} - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}} \tag{60}$$

where $\boldsymbol{\Sigma}_{z^*} = \mathrm{diag}(\sigma^2_{z,0}, \sigma^2_{z,1}, ..., \sigma^2_{z,R})$ and $\boldsymbol{\Phi}_{z^*} = \mathrm{diag}(\phi_{z,0}, \phi_{z,1}, ..., \phi_{z,R})$.

Recall from (50) that $\boldsymbol{z}_k \mid \sigma^2_{z,k}, \phi_{z,k} \sim \mathrm{BYM2}(\sigma^2_{z,k}, \phi_{z,k})$ for $k = 0, 1, ..., R$, so it follows that

$$\boldsymbol{z}^* \mid \boldsymbol{\Sigma}_{z^*}, \boldsymbol{\Phi}_{z^*} \sim \mathcal{N}_{(R+1)m}\left(\mathbf{0}, \boldsymbol{Q}_*^- \otimes (\boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*}) + \mathbf{I}_m \otimes ((\mathbf{I}_R - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*})\right).$$

The random effects under the MV-Sym model (50) can be expressed as $\boldsymbol{b} = (\mathbf{I}_m \otimes \boldsymbol{P}^*)\, \boldsymbol{z}^*$ and thus follow a mean-zero multivariate normal distribution with variance

$$(\mathbf{I}_m \otimes \boldsymbol{P}^*) \left[\boldsymbol{Q}_*^- \otimes \boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*} + \mathbf{I}_m \otimes (\mathbf{I}_R - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*}\right] (\mathbf{I}_m \otimes \boldsymbol{P}^*)^{\mathrm{T}}$$

$$= (\mathbf{I}_m \otimes \boldsymbol{P}^*) \left[\boldsymbol{Q}_*^- \otimes \boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*}\right] \left(\mathbf{I}_m \otimes \boldsymbol{P}^{*\mathrm{T}}\right) + (\mathbf{I}_m \otimes \boldsymbol{P}^*) \left[\mathbf{I}_m \otimes (\mathbf{I}_R - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*}\right] \left(\mathbf{I}_m \otimes \boldsymbol{P}^{*\mathrm{T}}\right)$$

$$= \boldsymbol{Q}_*^- \otimes \boldsymbol{P}^* \boldsymbol{\Phi}_{z^*} \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}} + \mathbf{I}_m \otimes \boldsymbol{P}^* (\mathbf{I}_R - \boldsymbol{\Phi}_{z^*}) \boldsymbol{\Sigma}_{z^*} \boldsymbol{P}^{*\mathrm{T}}$$

$$= \boldsymbol{Q}_*^- \otimes \boldsymbol{\psi}_u + \mathbf{I}_m \otimes \boldsymbol{\psi}_v$$

with $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$ taking the forms in (60).

Next we will derive the pairwise correlation between outcomes $r \neq r'$ implied by $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$. Observe that the diagonal and off-diagonal elements of $\boldsymbol{\psi}_u$ have the forms $\boldsymbol{\psi}_{u,rr} = \phi_{z,0}\gamma_r^2 + \phi_{z,r}\sigma_{z,r}^2$ and $\boldsymbol{\psi}_{u,rr'} = \phi_{z,0}\gamma_r\gamma_{r'}$, respectively. Then it follows that

$$\mathrm{Cor}(\boldsymbol{\psi}_u)_{rr'} = \frac{\boldsymbol{\psi}_{u,rr'}}{\left(\boldsymbol{\psi}_{u,rr}\boldsymbol{\psi}_{u,r'r'}\right)^{1/2}} = \frac{\phi_{z,0}\gamma_r\gamma_{r'}}{\left(\phi_{z,0}\gamma_r^2 + \phi_{z,r}\sigma_{z,r}^2\right)^{1/2}\left(\phi_{z,0}\gamma_{r'}^2 + \phi_{z,r'}\sigma_{z,r'}^2\right)^{1/2}}$$

$$= \frac{\gamma_r\gamma_{r'}}{\left(\gamma_r^2 + \frac{\phi_{z,r}}{\phi_{z,0}}\sigma_{z,r}^2\right)^{1/2}\left(\gamma_{r'}^2 + \frac{\phi_{z,r'}}{\phi_{z,0}}\sigma_{z,r'}^2\right)^{1/2}}.$$

Similarly, $\mathrm{Cor}(\boldsymbol{\psi}_v)_{rk}$ can be derived by replacing $\phi_{z,\cdot}$ with $(1 - \phi_{z,\cdot})$.

C1.2 MV-Asym

Let $\boldsymbol{z} = \left(\boldsymbol{z}_1^{\mathrm{T}}, ..., \boldsymbol{z}_R^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\boldsymbol{P} = \begin{pmatrix} 1 & \gamma_2 & ... & \gamma_R \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix},$$

where $\boldsymbol{z}_r$ and γ_r are as defined in (52). We will show this implies that the MV-Asym model (52) is a GMBYM2 model (47) with

$$\boldsymbol{\psi}_u = \boldsymbol{P}\boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}} \qquad \text{and} \qquad \boldsymbol{\psi}_v = \boldsymbol{P}(\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}} \tag{61}$$

where $\boldsymbol{\Sigma}_z = \mathrm{diag}(\sigma_{z,1}^2, ..., \sigma_{z,R}^2)$ and $\boldsymbol{\Phi}_z = \mathrm{diag}(\phi_{z,1}, ..., \phi_{z,R})$.

Recall from (52) that $\boldsymbol{z}_k \mid \sigma^2_{z,k}, \phi_{z,k} \sim \text{BYM2}(\sigma^2_{z,k}, \phi_{z,k})$ for $k = 1, ..., R$, so it follows that

$$\boldsymbol{z} \mid \boldsymbol{\Sigma}_z, \boldsymbol{\Phi}_z \sim \mathcal{N}_{mR}\left(\mathbf{0}, \boldsymbol{Q}_*^- \otimes (\boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z) + \mathbf{I}_m \otimes ((\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z)\right).$$

The random effects under the MV-Asym model (52) can be expressed as $\boldsymbol{b} = (\mathbf{I}_m \otimes \boldsymbol{P})\,\boldsymbol{z}$ and thus follow a mean-zero multivariate normal distribution with variance

$$(\mathbf{I}_m \otimes \boldsymbol{P})\left[\boldsymbol{Q}_*^- \otimes \boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z + \mathbf{I}_m \otimes (\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\right](\mathbf{I}_m \otimes \boldsymbol{P})^{\mathrm{T}}$$

$$= (\mathbf{I}_m \otimes \boldsymbol{P})\left[\boldsymbol{Q}_*^- \otimes \boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z\right]\left(\mathbf{I}_m \otimes \boldsymbol{P}^{\mathrm{T}}\right) + (\mathbf{I}_m \otimes \boldsymbol{P})\left[\mathbf{I}_m \otimes (\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\right]\left(\mathbf{I}_m \otimes \boldsymbol{P}^{\mathrm{T}}\right)$$

$$= \boldsymbol{Q}_*^- \otimes \boldsymbol{P}\boldsymbol{\Phi}_z\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}} + \mathbf{I}_m \otimes \boldsymbol{P}(\mathbf{I}_R - \boldsymbol{\Phi}_z)\boldsymbol{\Sigma}_z\boldsymbol{P}^{\mathrm{T}}$$

$$= \boldsymbol{Q}_*^- \otimes \boldsymbol{\psi}_u + \mathbf{I}_m \otimes \boldsymbol{\psi}_v$$

with $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$ taking the forms in (61).

Next we will derive the pairwise correlation between primary outcome, $r = 1$, and auxiliary outcome, $r' > 1$, implied by $\boldsymbol{\psi}_u$ and $\boldsymbol{\psi}_v$. Observe that $\boldsymbol{\psi}_{u,11} = \phi_{z,1}\sigma^2_{z,1} + \sum_{\ell=2}^R \gamma^2_\ell \phi_{z,\ell}\sigma^2_{z,\ell}$, $\boldsymbol{\psi}_{u,r'r'} = \phi_{z,r'}\sigma^2_{z,r'}$ and $\boldsymbol{\psi}_{u,1r'} = \gamma^2_{r'}\phi_{z,r'}\sigma^2_{z,r'}$, respectively. Then it follows that

$$\text{Cor}(\boldsymbol{\psi}_u)_{1r'} = \frac{\boldsymbol{\psi}_{u,1r'}}{\left(\boldsymbol{\psi}_{u,11}\boldsymbol{\psi}_{u,r'r'}\right)^{1/2}}$$

$$= \frac{\gamma_{r'}\phi_{z,r'}\sigma^2_{z,r'}}{(\phi_{z,1}\sigma^2_{z,1} + \sum_{\ell=2}^R \gamma^2_\ell \phi_{z,\ell}\sigma^2_{z,\ell})^{1/2}(\phi_{z,r'}\sigma^2_{z,r'})^{1/2}} = \frac{\gamma_{r'}\sqrt{\phi_{z,r'}}\sigma_{z,r'}}{(\phi_{z,1}\sigma^2_{z,1} + \sum_{\ell=2}^R \gamma^2_\ell \phi_{z,\ell}\sigma^2_{z,\ell})^{1/2}}.$$

Similarly, $\text{Cor}(\boldsymbol{\psi}_v)_{1k}$ can be derived by replacing $\phi_{z,\cdot}$ with $(1 - \phi_{z,\cdot})$.

C2 Additional plots for Section 4.4.2

In Figure C1 we show that changes in σ_r^2, $\hat{v}_{ir}$, and ϕ_r modify overall levels, but the same patterns shown in in Figure 4.2 persist. For this illustration, we choose $\omega = 0.5$ and vary ϕ_r, σ_r, and $\hat{v}_{ir}$. As in the main text, $\hat{v}_{i1} = \sigma_1^2$, and $\hat{v}_{ir} = d^2\sigma_r^2$ for $r = 2, 3$ where $d = 1$ for the equal signal-to-noise ratio setting and $d = 0.25$ for the unequal signal-to-noise ratio setting. Therefore, as we vary σ_r^2, $\hat{v}_{ir}$ varies along with it. The first column of Figure C1 uses the same parameter values for ϕ_r, σ_r, and $\hat{v}_{ir}$ as in Section 4.4.2, so it is identical to the third column in Figure 4.2. In the second column, we choose $\phi_r = 0$ for all outcomes r, which effectively reduces these GMBYM2 models to their IID analogs because we are now assuming all spatial variability is unstructured. We observe that the same patterns persist, indicating that our theoretical findings are not specific to the GMBYM2 model class, but generalizable to the broader class of multivariate FH models. In the third column, we choose $\sigma_r^2 = 0.5^2$ for all outcomes r, and, again, observe that the same patterns persist.

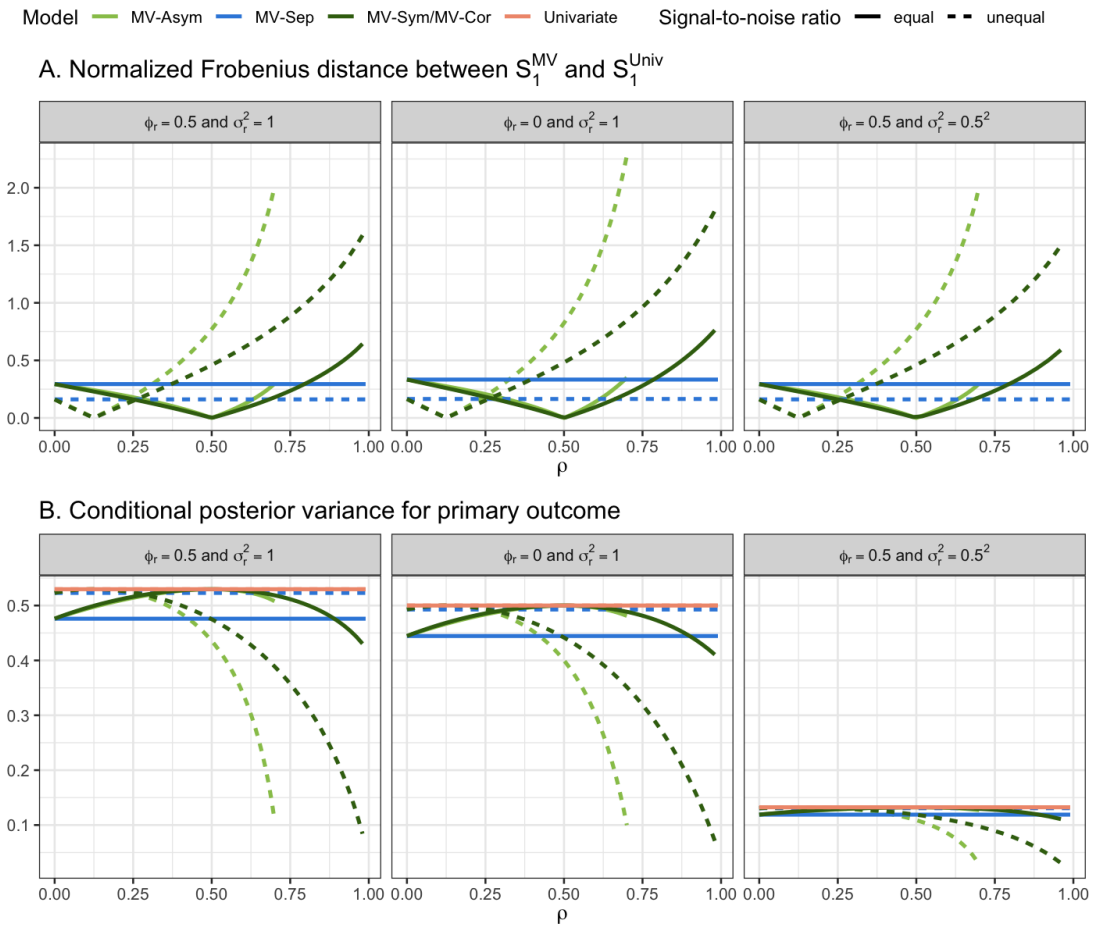


Figure C1. Properties of the conditional posterior distribution under GMBYM2-FH models to that under the univariate FH model as a function of pairwise latent correlation, for different values of ϕ_r, σ_r^2, and $\hat{v}_{ir}$.

C3 Bound on ρ in MV-Asym model for Section 4.4.2

We show that under the assumptions that $Cor(\boldsymbol{\psi}_u)_{1r'} = Cor(\boldsymbol{\psi}_v)_{1r'} = \rho \in [0, 1)$, for $r' = 2, ..., R$, and $\phi_r = \phi^*$, for $r = 1, ..., R$, it follows that ρ is bounded above by $\frac{1}{\sqrt{R-1}}$ when $(\sigma_1^2, ..., \sigma_R^2)$ are fixed. Recall from Section 4.3.1, that under the MV-Asym model,

$$\text{Cor}(\boldsymbol{\psi}_u)_{1r'} = \frac{\gamma_{r'}\sqrt{\phi_{z,r'}}\sigma_{z,r'}}{\left(\phi_{z,1}\sigma_{z,1}^2 + \sum_{\ell=2}^{R}\gamma_\ell^2\phi_{z,\ell}\sigma_{z,\ell}^2\right)^{1/2}}$$

and

$$\text{Cor}(\boldsymbol{\psi}_v)_{1r'} = \frac{\gamma_{r'}\sqrt{(1-\phi_{z,r'})}\sigma_{z,r'}}{\left((1-\phi_{z,1})\sigma_{z,1}^2 + \sum_{\ell=2}^{R}\gamma_\ell^2(1-\phi_{z,\ell})\sigma_{z,\ell}^2\right)^{1/2}}$$

for $r' > 1$. By applying the above assumptions we obtain

$$\rho = \frac{\gamma_k\sigma_{z,r'}}{\left(\sigma_{z,1}^2 + \sum_{\ell=2}^{R}\gamma_\ell^2\sigma_{z,\ell}^2\right)^{1/2}}.$$

For this hold for all $r' > 1$, it must be true that

$$\rho = \frac{d}{(\sigma_{z,1}^2 + (R-1)d^2)^{1/2}} \tag{62}$$

for ≥ 0.

Also, it follows directly from (52) that the variance of $(b_{11}, ..., b_{m1})$ under the MV-Asym model,

$$\sigma_1^2 = \sigma_{z,1}^2 + \sum_{r=2}^{R}\gamma_r^2\sigma_r^2 = \sigma_{z,1}^2 + (R-1)d^2.$$

All variances must be nonnegative, so it follows that

$$\sigma_{z,1}^2 = \sigma_1^2 - (R-1)d^2 \geq 0$$

which implies that

$$d \leq \frac{\sigma_1}{\sqrt{R-1}}.$$

Then, plugging this back into (62), we obtain

$$\rho = \frac{d}{\sigma_1} \leq \frac{1}{\sqrt{R-1}}.$$

C4 Sufficient conditions for equivalence between multivariate and univariate FH conditional posterior means

Suppose the sampling correlation matrix, $\text{Cor}(\hat{\boldsymbol{V}}_i) = \boldsymbol{\Omega}$ for all areas $i = 1, ..., m$, and the latent dependence, $\text{Cor}(\boldsymbol{\psi}_u) = \text{Cor}(\boldsymbol{\psi}_v) = \boldsymbol{\Lambda}$, for positive semi-definite matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$. Further, suppose the latent spatial dependence parameter is equal across all outcomes (i.e., $\phi_r = \phi^*$), all areas have equal sampling variance (i.e., $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{v}}^*$), and the sampling and latent variances are proportional across outcomes (i.e., $\boldsymbol{\sigma}^2 = c\hat{\boldsymbol{v}}_i = c\hat{\boldsymbol{v}}^*$). We will show that under these conditions, $\boldsymbol{\Lambda} = \boldsymbol{\Omega}$ implies $\mathcal{S}^{\text{MV}} = \mathcal{S}^{\text{Univ}}$, and, equivalently, the conditional posterior mean under the multivariate FH model is equal to that under the univariate FH model.

Under the above conditions we can write,

$$\hat{\boldsymbol{V}}_i = \hat{\boldsymbol{V}}^* = \boldsymbol{\Upsilon}^{1/2}\boldsymbol{\Omega}\boldsymbol{\Upsilon}^{1/2}, \qquad \boldsymbol{\psi}_u = c\phi\boldsymbol{\Upsilon}^{1/2}\boldsymbol{\Lambda}\boldsymbol{\Upsilon}^{1/2}, \qquad \text{and} \qquad \boldsymbol{\psi}_u = c(1-\phi)\boldsymbol{\Upsilon}^{1/2}\boldsymbol{\Lambda}\boldsymbol{\Upsilon}^{1/2},$$

where $\boldsymbol{\Upsilon} = \text{diag}(\hat{\boldsymbol{v}}^*)$. It follows that if $\boldsymbol{\Lambda} = \boldsymbol{\Omega}$, then $\hat{\boldsymbol{V}}^* = \frac{1}{c\phi}\boldsymbol{\psi}_u = \frac{1}{c(1-\phi)}\boldsymbol{\psi}_v$, so

$$\mathcal{S}^{\text{MV}} = \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}\boldsymbol{V}^{-1}\right)^{-1} = \left(\mathbf{I}_{mR} + (\boldsymbol{Q}_*^- \otimes \boldsymbol{\psi}_u + \mathbf{I}_m \otimes \boldsymbol{\psi}_v)(\mathbf{I}_m \otimes \boldsymbol{V}^{*-1})\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + (\boldsymbol{Q}_*^- \otimes \boldsymbol{\psi}_u\boldsymbol{V}^{*-1} + \mathbf{I}_m \otimes \boldsymbol{\psi}_v\boldsymbol{V}^{*-1})\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + (\boldsymbol{Q}_*^- \otimes c\phi^*\mathbf{I}_R + \mathbf{I}_m \otimes c(1-\phi^*)\mathbf{I}_R)\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + (\boldsymbol{Q}_*^- \otimes c\phi^*\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} + \mathbf{I}_m \otimes c(1-\phi^*)\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + \left(\boldsymbol{Q}_*^- \otimes \phi^*\boldsymbol{\Sigma} + \mathbf{I}_m \otimes (1-\phi^*)\boldsymbol{\Sigma}\right)\left(\mathbf{I}_m \otimes c\boldsymbol{\Sigma}^{-1}\right)\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}^{\text{Sep}}\left(\mathbf{I}_m \otimes \text{diag}(\hat{\boldsymbol{v}}^{*-1})\right)\right)^{-1}$$

$$= \left(\mathbf{I}_{mR} + \boldsymbol{\Psi}^{Sep}\text{diag}(\mathbf{V})^{-1}\right)^{-1} = \mathcal{S}^{\text{Univ}}.$$

C5 Estimates of auxiliary outcomes in Kenya example

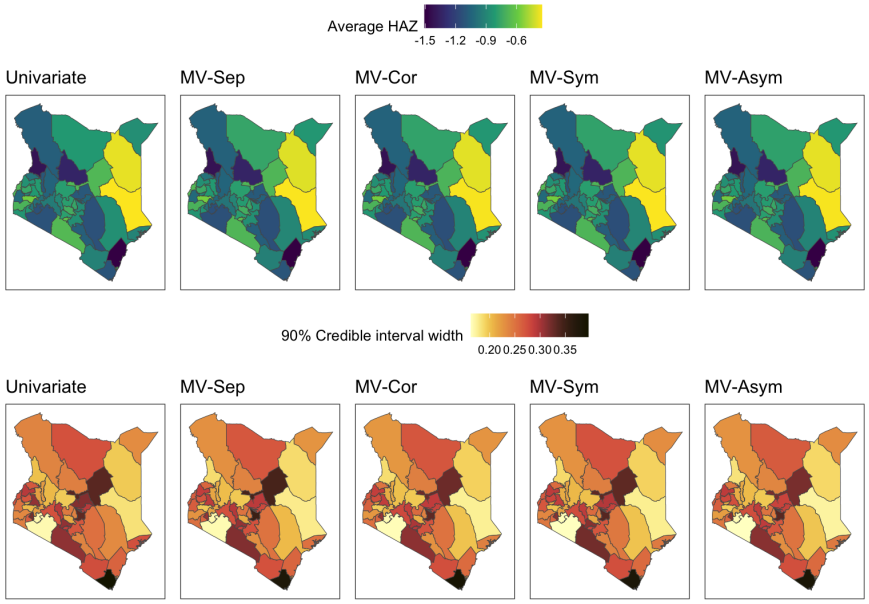


Figure C2. Estimated average height-for-age z-scores and widths of 90% credible intervals for admin1 areas in Kenya under the univariate and GMBYM2-FH models, using 2022 Kenya DHS data.

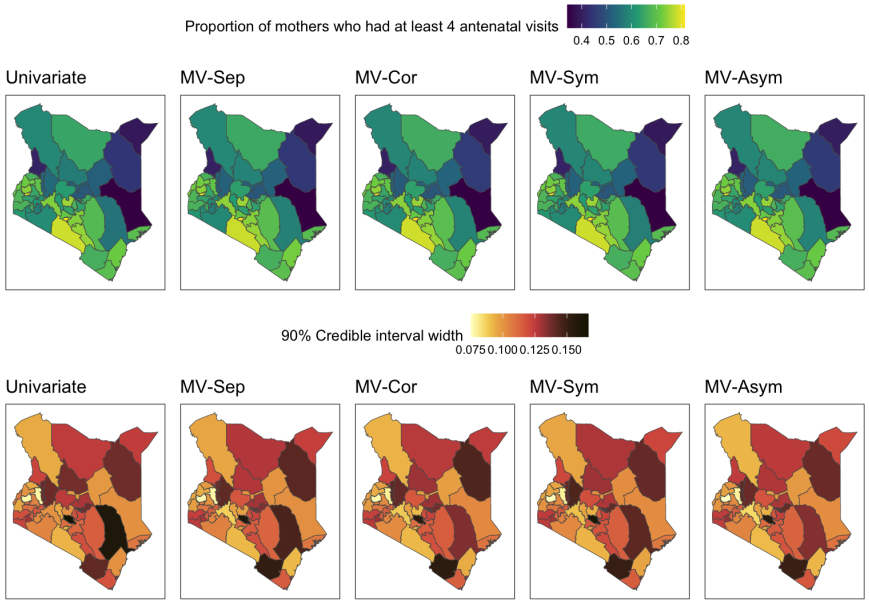


Figure C3. Estimated proportion of mothers who had at least 4 antenatal care visits and widths of 90% credible intervals for admin1 areas in Kenya under the univariate and GMBYM2 FH models, using 2022 Kenya DHS data.