Direct-Assisted Bayesian Unit-level Modeling for Small Area Estimation of Rare Event Prevalence

Alana McGovern[1*], Katherine Wilson[2], and Jon Wakefield[1,2]

[1] Department of Statistics, University of Washington, Seattle WA, USA
[2] Department of Biostatistics, University of Washington, Seattle WA, USA

*Corresponding author, amcgov@uw.edu
This work was supported by the National Institutes of Health [R01 HD112421-02].

Abstract

Small area estimation using survey data can be achieved by using either a design-based or a model-based inferential approach. Design-based direct estimators are generally preferable because of their consistency, asymptotic normality, and reliance on fewer assumptions. However, when data are sparse at the desired area level, as is often the case when measuring rare events, these direct estimators can have extremely large uncertainty, making a model-based approach preferable. A model-based approach with a random spatial effect borrows information from surrounding areas at the cost of inducing shrinkage towards the local average. As a result, estimates may be over-smoothed and inconsistent with design-based estimates at higher area levels when aggregated. We propose two unit-level Bayesian models for small area estimation of rare event prevalence which use design-based direct estimates at a higher area level to increase consistency in aggregation. This model framework is designed to accommodate sparse data obtained from two-stage stratified cluster sampling, which is particularly relevant to applications in low and middle income countries. After introducing the model framework and its implementation, we conduct a simulation study to evaluate its properties and apply it to the estimation of the neonatal mortality rate in Zambia, using 2014 Demographic Health Surveys data.

Statement of significance: This work contributes to the literature regarding small area estimation of rare event prevalence using sparse complex survey data - such applications are extremely important in low and middle income countries in particular. We introduce a Bayesian model which integrates design-based estimators into a unit-level hierarchical model, which encourages agreement between aggregated small area estimates and design-based estimates at a higher area level, which will be advantageous in global and public health settings.

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, 1964). 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 small area estimation (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 from surrounding 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 paper is a pair of unit-level Bayesian models with random spatial effects 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 do not. We propose two variations of our model because while an exact benchmarking approach is often desirable, there may be cases in which a practitioner wishes to take the uncertainty of the benchmark into account. 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 we will introduce the context of SAE in low- and middle-income countries (LMICs), which motivates the new model. In section 3 we will present a standard Bayesian unit-level model for rare events and in section 4 we will extend this model and introduce the soft and hard direct-assisted Bayesian unit-level models and discuss its implementation. In section 5 we conduct a simulation study comparing the DABUL models to a standard unit-level Bayesian model before concluding in section 6 with an application of the DABUL models to estimation of the neonatal mortality rate (NMR) in Zambia.

2. SMALL AREA ESTIMATION IN LOW- AND MIDDLE-INCOME COUNTRIES

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 et al., 2015) to obtain estimates in LMICs at the level of the first or second administrative area. The majority of LMICs carry out DHS and/or MICS, 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 in order to get closer to the Sustainable Development Goal (SDG) of no more than 12 deaths per 1000 live births by 2030 (`https://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 surveys are powered at the first administrative 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 second administrative level are desired because this is often the level at which health interventions are administered. As a result of this sampling design, it is generally necessary to use unit-level models to obtain small area estimates, 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 paper 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 first administrative 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'.

3. BAYESIAN UNIT-LEVEL GLMMS FOR RARE EVENT PREVALENCE

3.1 Sampling Model

Suppose we seek to estimate the prevalence of a rare event, neonatal mortality, at the second administrative level using sparse survey data. It is common to fit a generalized linear mixed model (GLMM) 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 sampled births and neonatal deaths in a fixed time period in sampled cluster c, respectively, where each cluster is contained in one of m_2 second administrative areas. Then, for neonatal mortality rate, r_c, and overdispersion parameter λ, we assume that for each cluster, c,

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

with the parameterization

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

Under this parameterization, $\mathbb{E}[Z_c|n_c, r_c, \lambda] = n_c r_c$ and $\text{Var}(Z_c|n_c, r_c, \lambda) = (1+\lambda) n_c r_c$. 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 a second administrative level random spatial effect and $\delta_k(j)$ is the set of clusters in k^{th} administrative area j. When this model is fit with sparse data there is often a large amount of shrinkage towards the local mean because information in each second administrative 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 first administrative level and a set of random spatial effects at the second administrative level, i.e., replacing (3) with

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

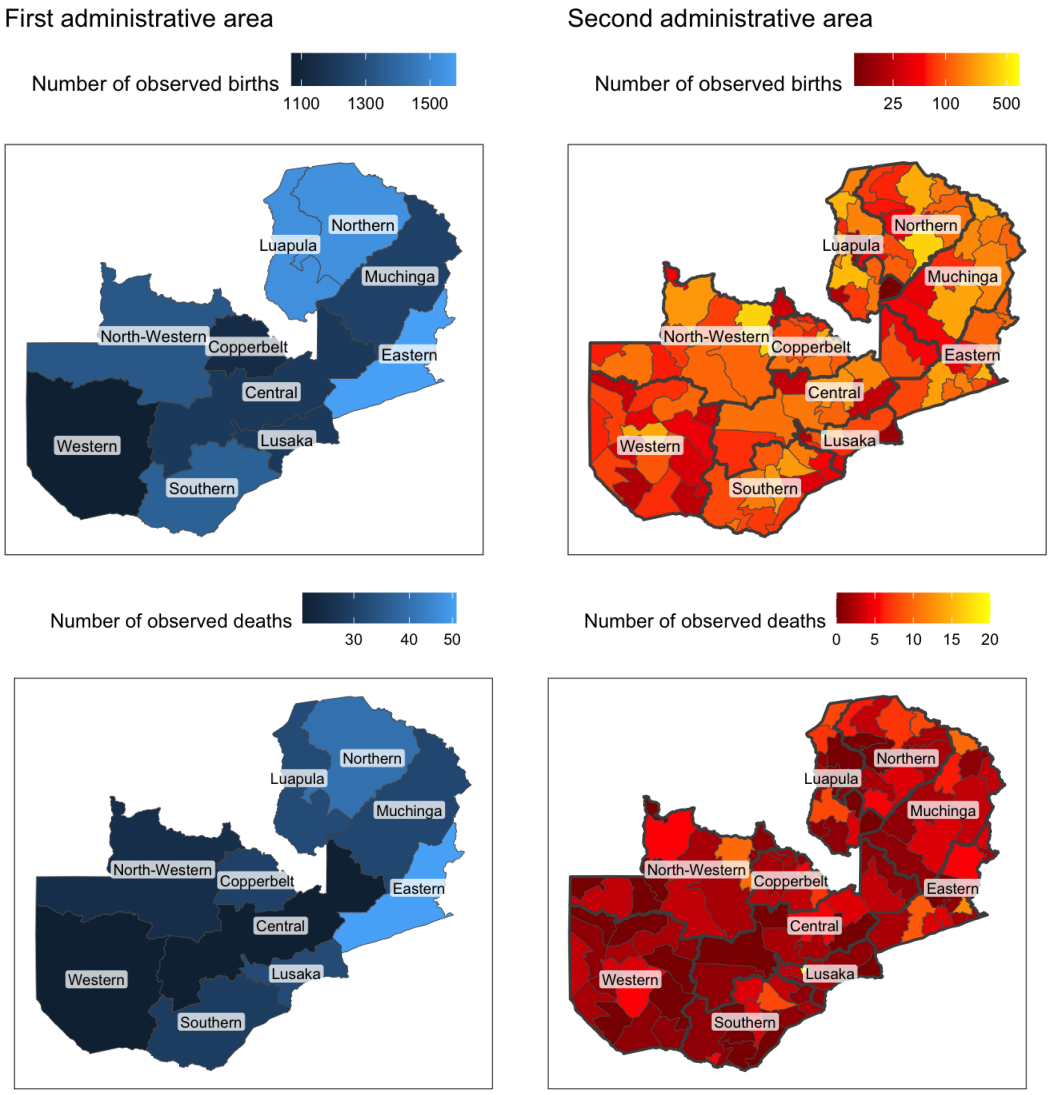
where β_i is a first administrative level fixed effect and β_1 is fixed at 0 to preserve identifiability.

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 second administrative level using all births between 2009 and 2013 recorded in the 2014 Zambia DHS (Zambia Central Statistical Office, Zambia Ministry of Health, and ICF International, 2014). Zambia has 10 first administrative areas and 115 second administrative areas. A summary of this data is displayed in Figure 1. The numbers of observed births and neonatal deaths in each first administrative 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 second administrative 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 required for estimation at the second administrative level.

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

Figure 1. Summary of births and neonatal deaths between 2009 and 2013 recorded in the 2014 Zambia DHS. The top row provides maps of the total number of observed births in each first and second administrative area. The bottom row provides maps of the total number of neonatal births in each first and second administrative area. Each map has a different color scale to make differences between areas more clear.



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

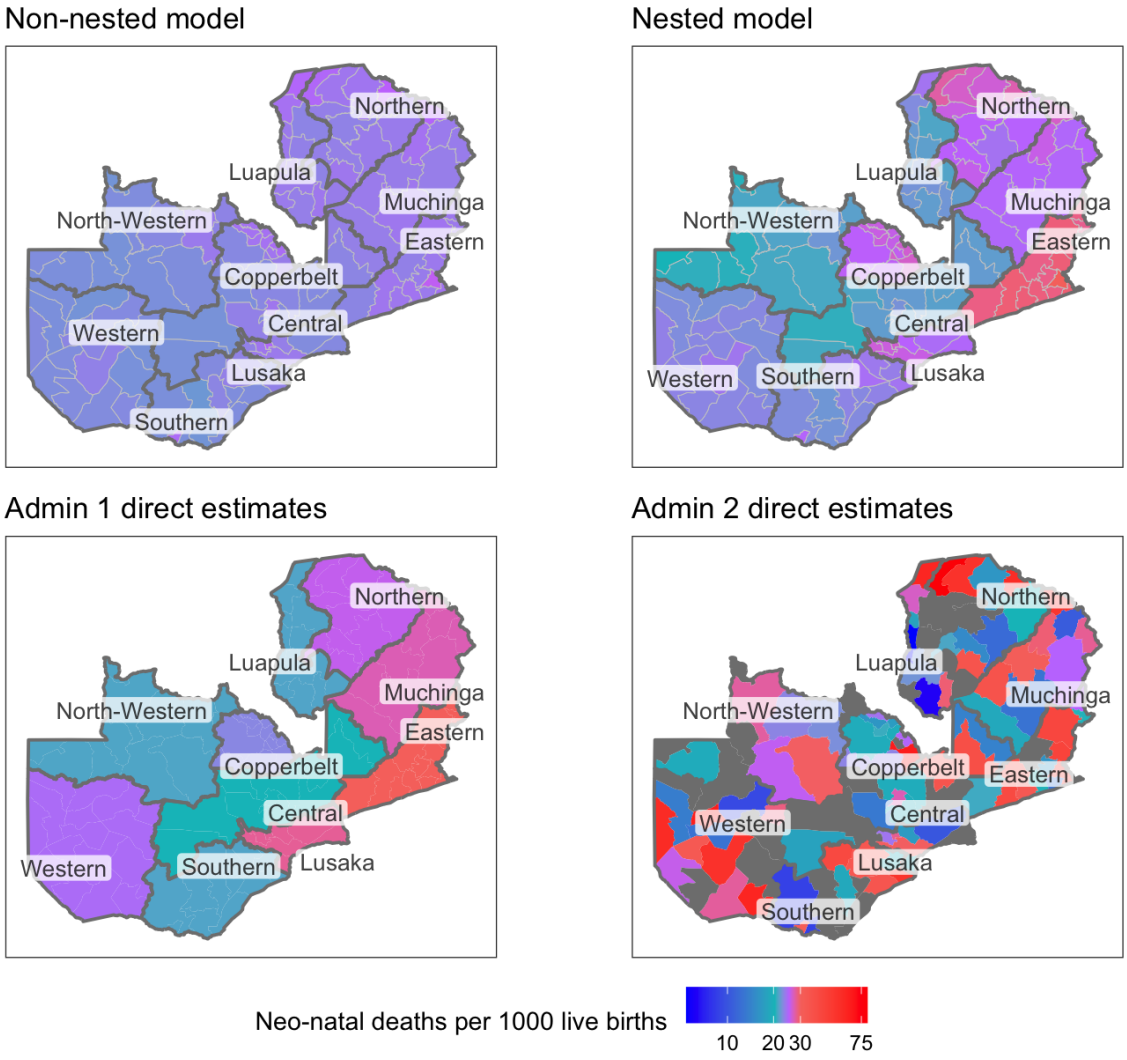
$$\log(r_c) = \alpha_U \mathbb{1}_{c \in \delta_U} + \alpha_R \mathbb{1}_{c \in \delta_R} + b_j \mathbb{1}_{c \in \delta_2(j)} \tag{5}$$

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

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 second administrative level under regression models (5) and (6) using the `Stan` software (Stan Development Team, 2024). While a variety of spatial models at the second administrative 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-Mollie (BYM) model (Besag et al., 1991), which includes both unstructured IID spatial effects and structured ICAR spatial effects (Besag, 1974). The BYM2 model has two parameters: σ^2, which indicates the total variance of the spatial 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. 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. For hyperpriors, we set $\phi \sim Beta(0.5, 0.5)$ 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: $\lambda \sim \text{Exp}(1)$, $(\alpha_U, \alpha_R) \overset{\text{iid}}{\sim} \mathcal{N}(-3.5, 3^2)$ 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. More information regarding these prior choices is presented in Appendix A. We aggregate the urban and rural estimates within each second administrative area using urban/rural population fractions estimated using the method in Wu and Wakefield (2024).

Figure 2. Map of NMR estimates in Zambia (2009-2013) under various models. The maps on the top row display results from unit-level Bayesian negative binomial models with BYM2 spatial effects at the second administrative level, while the maps on the bottom row display Hájek direct estimates at the first and second administrative level, respectively. The map on the top left displays the estimates obtained from the model with spatial effects only at the second administrative level (5), while the map on the top right displays the estimates obtained from the model with nested spatial effects (6).Note that data at the second administrative level is sparse, so the bottom right map is only included to visualize the magnitude of shrinkage.



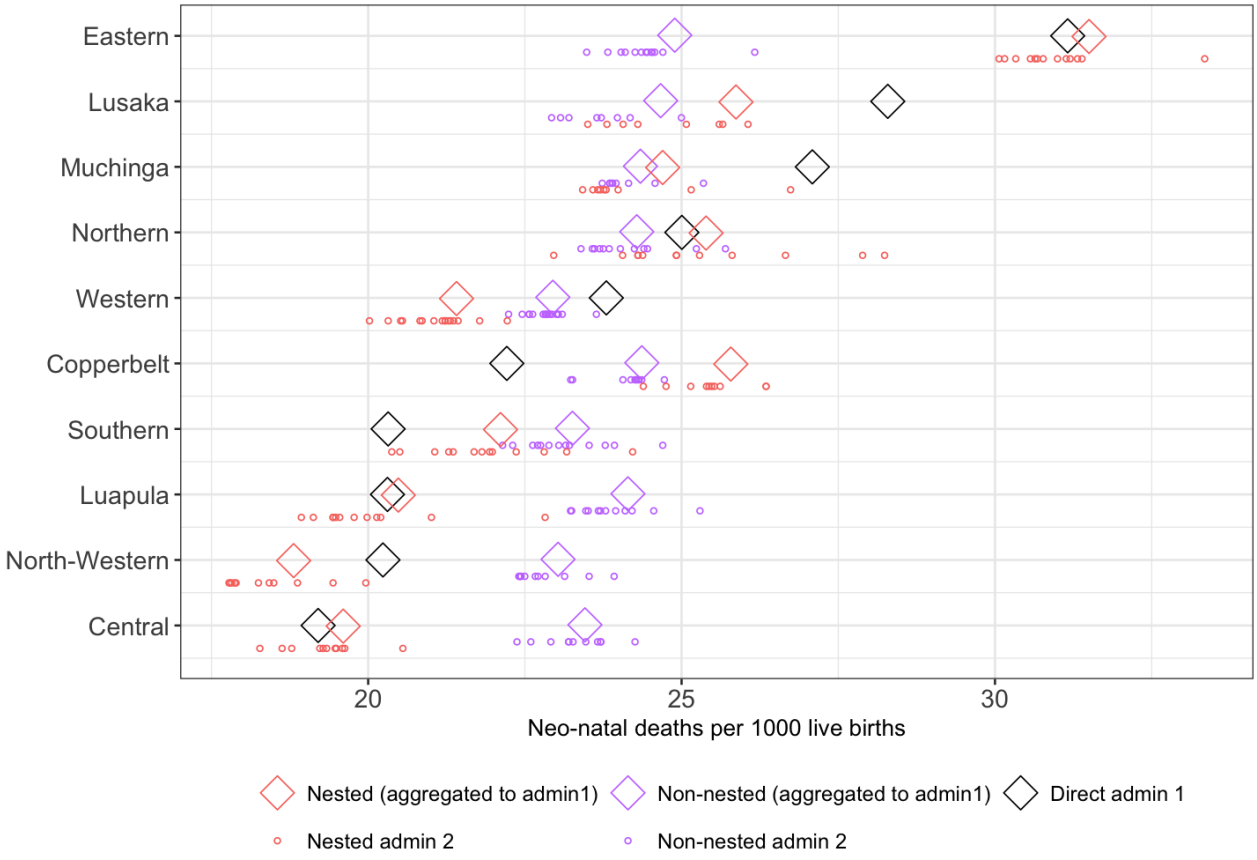
In figure 2 we map the NMR (deaths per 1000 live births) estimates under non-nested, nested, and direct methods. We observe that including a set of fixed effects at the first administrative level in the nested model significantly reduces shrinkage. The nested model

estimates, mapped in the middle, have a range of $18 - 33$ deaths per 1000 live births, while the estimates from the non-nested model, mapped on the left, are nearly homogeneous with a range of $22 - 26$ deaths per 1000 live births. Although we still observe shrinkage within each first administrative level, there is still vast improvement. In figure 2, we also include a map of the second administrative 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. To further compare the model-based and direct estimates, we aggregate the second administrative level model-based estimates to the first administrative level with population weights calculated using WorldPop (`worldpop.org`). In figure 3, observe that the aggregation of the estimates to the first administrative level can be quite far from the consistent design-based estimate. In some first administrative areas, such as the Central, Luapala, and Eastern regions, aggregated estimates from the nested model are in close agreement with their corresponding design-based estimates. However, for other areas, such as the Southern, Western, and Muchinga 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.

4. DIRECT-ASSISTED BAYESIAN UNIT-LEVEL MODEL

In the following section we introduce the Direct-assisted Bayesian unit-level (DABUL) models: an extension of the standard unit-level Bayesian model with nested spatial effects which incorporates direct design-based estimates to encourage consistency with design-based estimates in aggregation.

Figure 3. NMR estimates at the second administrative level in Zambia (2009-2013) under various models. The colored circles denote the second administrative level estimates using unit-level Bayesian negative binomial models with BYM2 spatial effects at the second administrative level and the colored diamonds denote their aggregations to the first administrative level. The black diamonds denote Hájek direct estimates at the first administrative level.



Let $\mathbf{Y}$ be the cluster-length vector containing the total number of neonatal deaths, Y_c, in each cluster c. Let $\mathbf{Y_i}$ be the sub-vector containing the elements of $\mathbf{Y}$ corresponding to all clusters in the first administrative area i, and Y_{i+} be the sum of the elements in this sub-vector. Similarly, 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 first administrative area, i. Also, let $\mathbf{Nr}$ be the vector containing elements $N_c r_c$, and $\mathbf{Nr}_i$ be the sub-vector containing the elements of $\mathbf{Nr}$ corresponding to the clusters in the first administrative area, i. Then an equivalent way to define the negative binomial distribution expressed in (2) is

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

for first administrative area, i, where

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

and

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

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,

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

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 θ, then $x_+ = \sum_{i=1}^n x_i$ follows a negative binomial distribution with mean $\sum_{i=1}^n \theta_i$, and overdispersion parameter θ. It also holds that $\mathbf{x}|x_+,\boldsymbol{\theta}$ follows a DCM distribution with parameter vector $\boldsymbol{\theta}/\theta$.

When all births and neonatal deaths in each cluster are observed, this alternative form of the negative binomial distribution is equivalent to (2), but consider a survey sampling context in which 1) only some of the births and corresponding neonatal deaths are observed, i.e., $\mathbf{Y}$ is not known, and 2) each Y_{i+} can be estimated through a design-based estimate.

4.1 Accounting for survey sampling

Let us set point 2) for a moment and assume Y_{i+} is known for all i. 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 sampled births and neonatal deaths across all of the 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,

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

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 this model pertains to the situation in which there is sufficient data to use direct design-based estimates at the first administrative level, though not at the second administrative level. We will denote the design-based prevalence estimate for a first administrative area i as $\hat{r}_{Di}$. 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 $Y_{i+} = \hat{r}_{D_i} \times N_i$, where $N_i = \sum_{c \in \delta_1(i)} N_c$, to apply the 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 on the logit scale (Breidt and Opsomer, 2017, Section 3), so we can specify

$$\text{logit}(\hat{r}_{Di})|Y_{i+} \sim \mathcal{N}\left(\text{logit}\left(Y_{i+}/N_i\right), \hat{V}_i\right) \tag{11}$$

where $N_i = \sum_{c \in \delta_1(i)} N_c$ and $\hat{V}_i$ is the design-based variance of logit($\hat{r}_{Di}$), 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.

Then, the joint distribution of $(\mathbf{Z}, \hat{\mathbf{r}}_\mathbf{D})$ conditional on $(\mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma})$ under the hard DABUL model can be expressed as,

$$\begin{aligned} P(\mathbf{Z}, \hat{\mathbf{r}}_\mathbf{D} | \mathbf{Y}, \mathbf{Y}_+, \boldsymbol{\gamma}) &= P(\mathbf{Z} | \hat{\mathbf{r}}_\mathbf{D}, \mathbf{Y}, \boldsymbol{\gamma}) P(\hat{\mathbf{r}}_\mathbf{D} | \mathbf{Y}_+) \\ &\propto \prod_{i=1}^{m_1} P(\mathbf{Z_i} | \hat{r}_{Di}, \mathbf{Y_i}, \boldsymbol{\gamma}_i) I(\hat{r}_{Di} = Y_{i+}/N_i) \end{aligned} \tag{12}$$

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

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

By definition, $\hat{r}_{Di}$ is a weighted sum of $\mathbf{Z_i}$, so it follows that $P(\mathbf{Z_i} | \hat{r}_{Di}, \mathbf{Y_i}, \boldsymbol{\gamma}_i)$ is a product of Hypergeometric distributions conditional on the survey-weighted sum of $\mathbf{Z_i}$, $g_{Di}(\mathbf{Z_i})$, being equal to $\hat{r}_{Di}$. More explicitly, for first administrative area i,

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

Note that $g_{Di}(\mathbf{Z_i}) = \hat{r}_{Di}|\mathbf{Y_i}, \boldsymbol{\gamma}_i$ follows the distribution of a weighted sum of Hypergeometric random variables.

4.3 Combining likelihood components

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

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

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 first administrative 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 be uninformative because there is no data for these clusters, so the total neonatal death counts from all of the unobserved clusters in a first administrative area can be collapsed into one group. Hence, we replace (8) with

$$(\mathbf{Y_i^{(s)}}, Y_{i+} - Y_{i+}^{(s)})|Y_{i+}, \mathbf{r_i}, \lambda \sim \text{DCM}\left(Y_{i+}, \left((\mathbf{Nr_{i:\gamma_c=1}})/\lambda, \left(\sum_{\substack{c\in\delta_1(i)\\ \gamma_c=0}} N_c r_c\right)/\lambda\right)\right) \quad (16)$$

where $Y_{i+}^{(s)}$ is the sum of the elements in $\mathbf{Y_i^{(s)}}$ and $\mathbf{Nr_{i:\gamma_c=1}}$ is the sub-vector of $\mathbf{Nr}$ containing elements corresponding to the observed clusters in first administrative 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

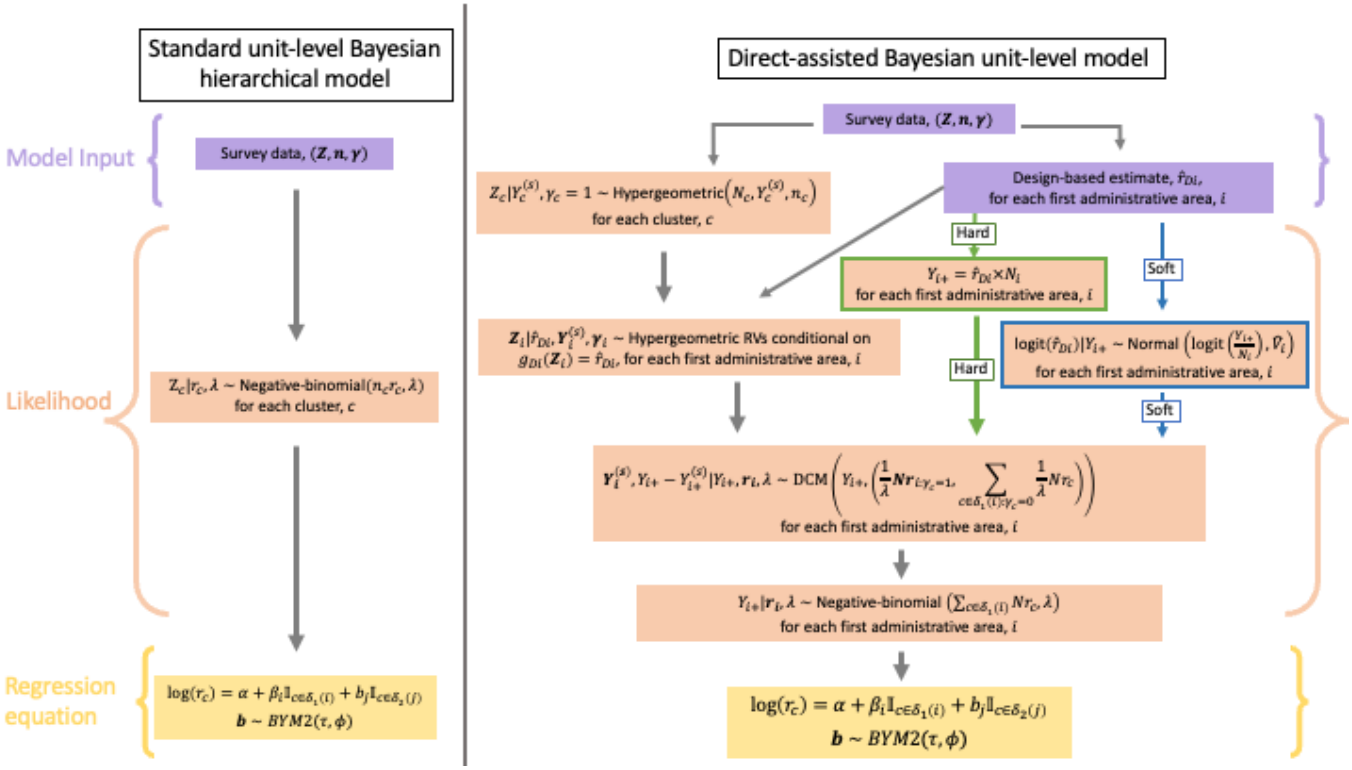
smaller magnitude. This reduction increases numerical stability and computational efficiency.

This completes the derivation of the posterior distribution under the DABUL models, which can be expressed as follows,

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

where $\boldsymbol{\kappa}$ is the set of regression parameters and hyperparameters. Figure 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 $\hat{\mathbf{r}}_\mathbf{D}$ is used to estimate $\mathbf{Y}_+$.

Figure 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.



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 foundational 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.

All components of the posterior distribution (17) have a closed form except for $P(g_{Di}(\mathbf{Z_i}) = \hat{r}_{Di}|\mathbf{Y_i}, \boldsymbol{\gamma}_i)$, the probability mass function of a weighted sum of independent Hypergeometric random variables. Although $g_{Di}(\mathbf{Z_i})|\mathbf{Y_i}, \boldsymbol{\gamma}_i$ technically follows a discrete distribution, 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 $P(g_{Di}(\mathbf{Z_i}) = \hat{r}_{Di}|\mathbf{Y_i}, \boldsymbol{\gamma}_i)$ 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 $\mathbf{Y}^{(s)}$. 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 $g_{Di}(\mathbf{Z_i})|\mathbf{Y_i}, \boldsymbol{\gamma}_i$. Gaussian kernel density estimation is then applied

to this empirical distribution, using the `density` function in `R`, to approximate $P(g_{Di}(\mathbf{Z_i}) = \hat{r}_{Di}|\mathbf{Y_i}, \boldsymbol{\gamma}_i)$. The bandwidth is chosen using Silverman's rule-of-thumb method Silverman (2018).

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

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

for m in 1:M **do**

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

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 first administrative area, i **do**

Draw $Y_{i+}^{(m)} \sim P(Y_{i+}|\mathbf{Y^{(s)}}^{(m-1)}, \boldsymbol{\gamma}, \mathbf{r}^{(m)}, \lambda^{(m)}, \hat{r}_{Di}, \sigma^2_{\hat{r}_{Di}})$ using inverse transform sampling

end for

for each cluster, c **do**

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

end for

end for

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

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

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

$Y_{i+} = \hat{r}_{Di} N_i$ for each first administrative area i

for m in 1:M **do**

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

for each cluster, c **do**

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

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

end for

end for

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

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 first administrative areas and 159 second administrative areas. There is a minimum number of 13 second administrative areas in each first administrative area and a maximum of 27. A map of these areas is provided in figure 5. This neighborhood structure and map is modified from that of Angola. The total number of urban and rural clusters in each first administrative 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 first administrative area, the clusters are distributed evenly across second administrative areas so that the number of births and clusters in a second administrative area are similar to others within its first administrative area.

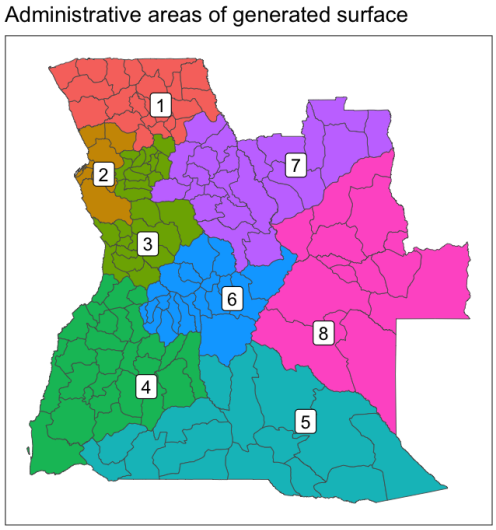
The total number of neonatal deaths in each cluster, Y_c, is drawn from a negative binomial distribution with mean $N_c r_c$ and overdispersion parameter $\lambda = 0.25$, where r_c is defined as

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

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 may be different between urban and rural areas. 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 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 $\phi = 0.7$. Through these settings we can observe whether the relative performance of the DABUL

model is affected by spatial precision or dependence.

Figure 5. Second administrative area neighborhood structure for simulation study



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 1. A proportion of urban and rural clusters are sampled from each first administrative 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 is this is often the case for DHS and MICS. Similarly, note that clusters are sampled at the first administrative level to mimic DHS and MICS, which are powered at the first administrative level. Due to this sampling design, there may be second administrative 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 displayed in Table 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 first administrative level using the

Table 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

`SUMMER` (Li et al. (2024)) and `survey` (Lumley (2024)) packages in R.

2. Second administrative level prevalence estimates from a standard unit-level (UL) 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. Second administrative 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. Second administrative 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.

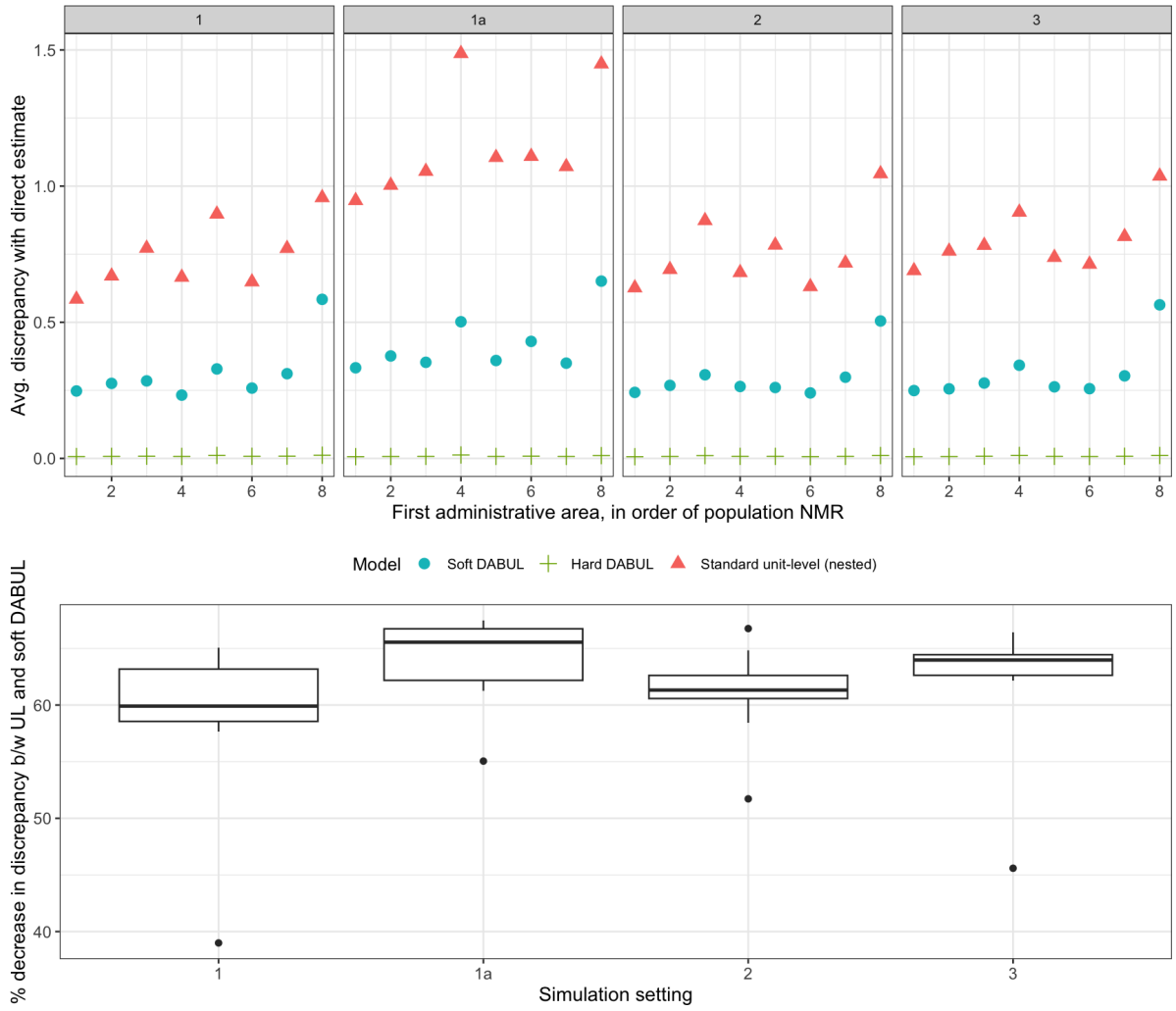
As in the motivating example, we aggregate the urban and rural estimates within each second administrative area using urban/rural population fractions. The run time of each DABUL model was approximately 130 minutes, while the run time of each standard unit-level model in `Stan` was approximately 2 minutes.

We compare the performance of the second administrative level prevalence estimates using four metrics. First, we quantify the discrepancy, or absolute difference, between the

first administrative level aggregated estimates and the direct estimates, averaged across all simulations. For first administrative area i, the average discrepancy is calculated by $\frac{1}{500}\sum_{k=1}^{500}|\sum_{j\in\zeta(i)} w_j\hat{r}_j^{(k)} - \hat{r}_{Di}|$ where $\hat{r}_j^{(k)}$ is the second administrative level prevalence estimate for the k^{th} sample dataset, w_j is the population weight for second administrative area j, and $\zeta(i)$ is the set of second administrative areas in first administrative area i. Second, we evaluate the absolute error of the second administrative level estimates, averaged across all simulations, calculated by $\frac{1}{500}\sum_{k=1}^{500}|\hat{r}_j^{(k)} - r_j|$, where r_j is the population prevalence for second administrative area j. Then, we compare the coverage and average width of the second administrative level 90% credible intervals, calculated by $\frac{1}{500}\sum_{k=1}^{500} I(r_j \in [Q_j^{(k)}(5), Q_j^{(k)}(95)])$ and $\frac{1}{500}\sum_{k=1}^{500}(Q_j^{(k)}(95) - Q_j^{(k)}(5))$, respectively, where $Q_j^{(k)}(q)$ is the q^{th} quantile of the posterior distribution for second administrative area j and sample dataset k. Additionally, coefficients of variation of the second administrative level estimates, calculated by $\frac{1}{500}\sum_{k=1}^{500}\frac{SD_j^{(k)}}{\hat{r}_j^{(k)}}$, are presented in Appendix B.

Figure 6 displays the discrepancies between aggregated second administrative level model-based estimates and first administrative level direct estimates. The top panel depicts the average discrepancies across simulations for each first administrative 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 first administrative 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 first administrative level direct estimates, on average, compared to aggregated estimates from the standard unit-level model. This decrease in discrepancy is slightly larger when the

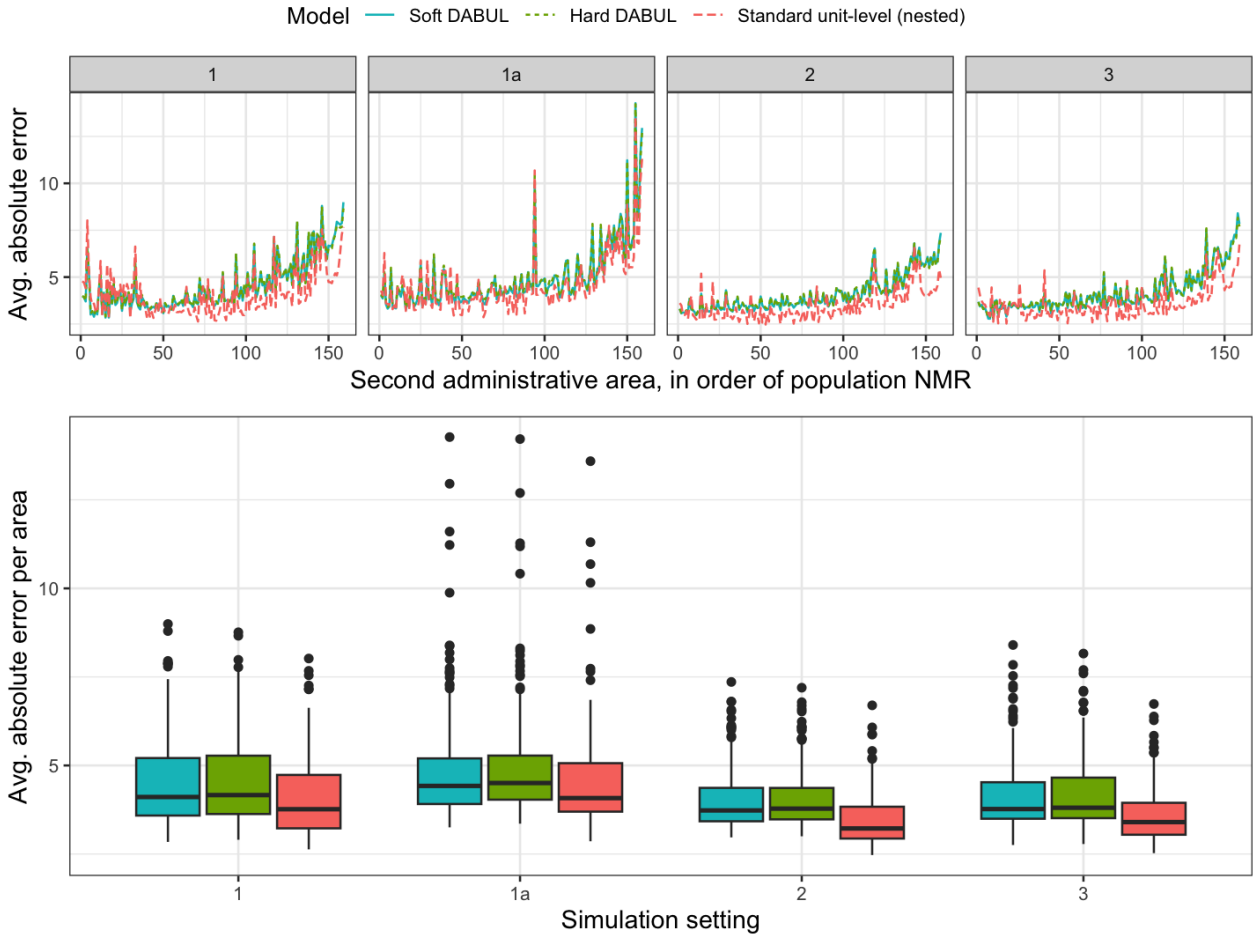
Figure 6. Discrepancy between aggregated model-based second administrative level estimates and design-based first administrative level estimates, across 500 simulations for each of 4 settings.



standard unit-level model has larger discrepancy with the direct estimates (scenario 1a). Figure 15 presented in Appendix B further shows this pattern.

Figure 7 displays the absolute error of the second administrative 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 first administrative area, model, and simulation setting. The bottom panel depicts the distribution of the average absolute error of each second administrative

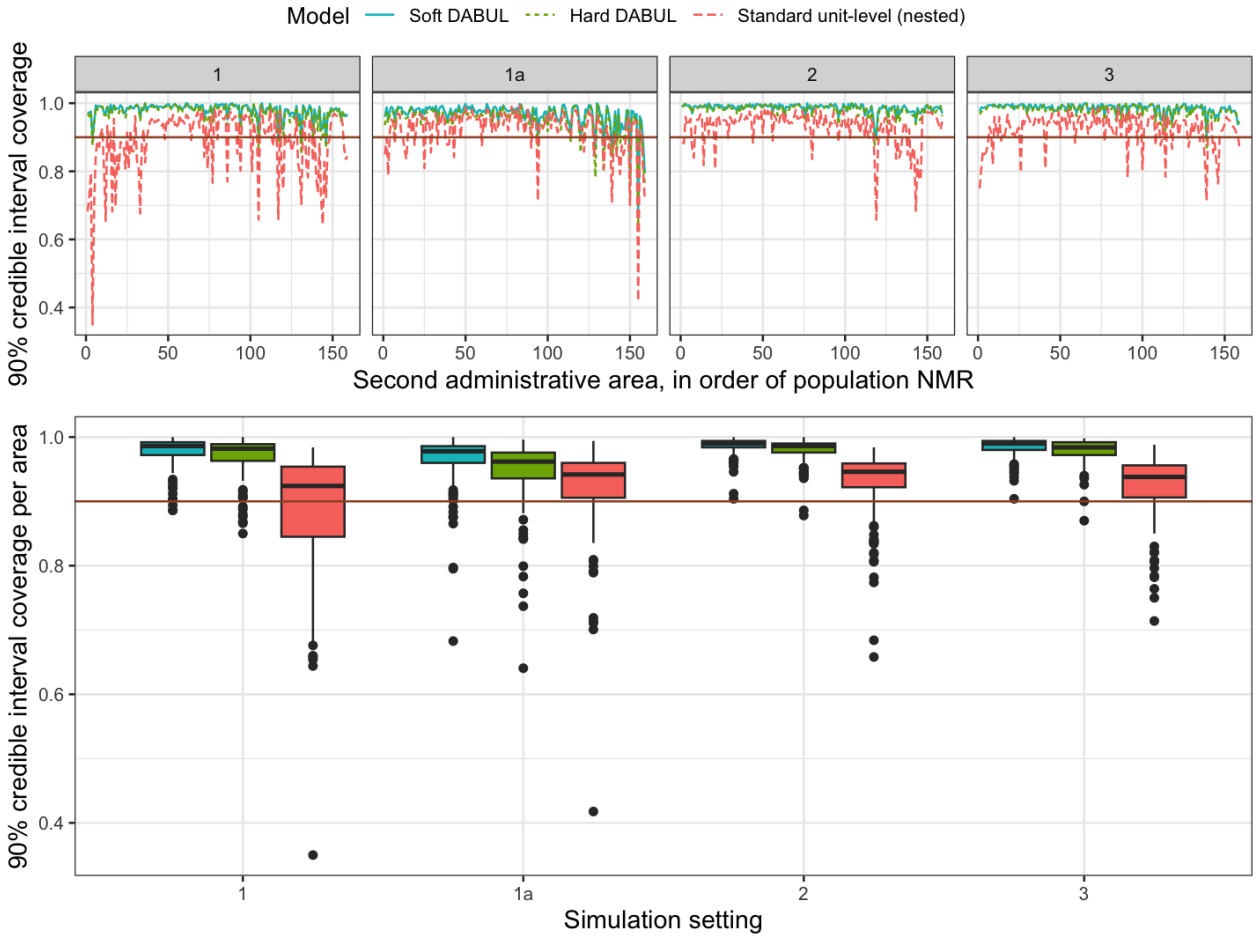
Figure 7. Absolute error of second administrative level estimates, across 500 simulations for each of 4 settings.



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 8 and 9 compare the coverage and width of the 90% credible intervals for the second administrative level estimates. Note that coverage of area-specific Bayesian credible intervals in unit-level models are expected to attain the desired coverage *on average*, but the credible 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 unit-level models was not accurate in a range of

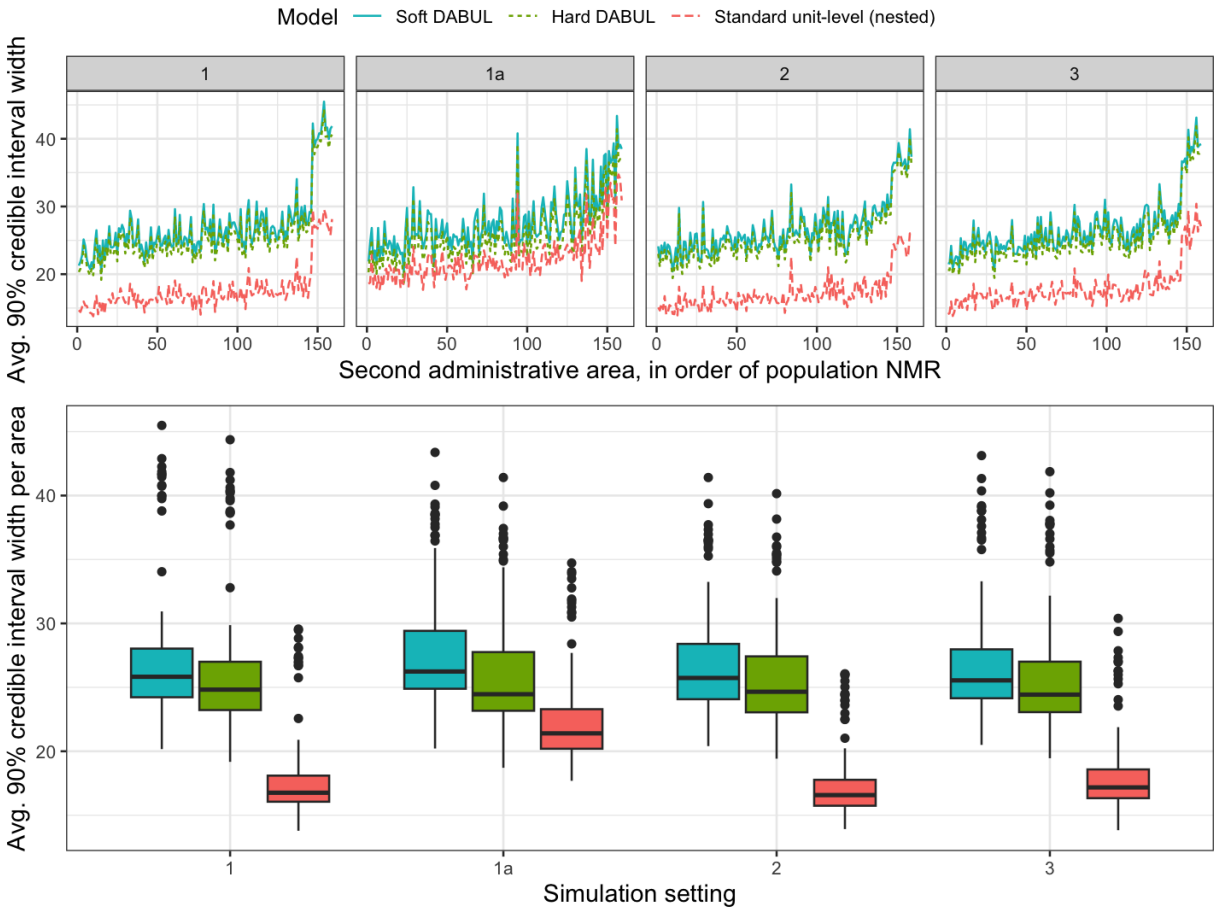
Figure 8. Coverage of 90% credible intervals for second administrative level estimates, across 500 simulations for each of 4 settings.



scenarios. From figures 8 and 9, we observe that the DABUL estimates have wider credible 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 credible 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.

Figure 9. Width of 90% credible intervals (in deaths per 1000 births) for second administrative level estimates, across 500 simulations for each of 4 settings.

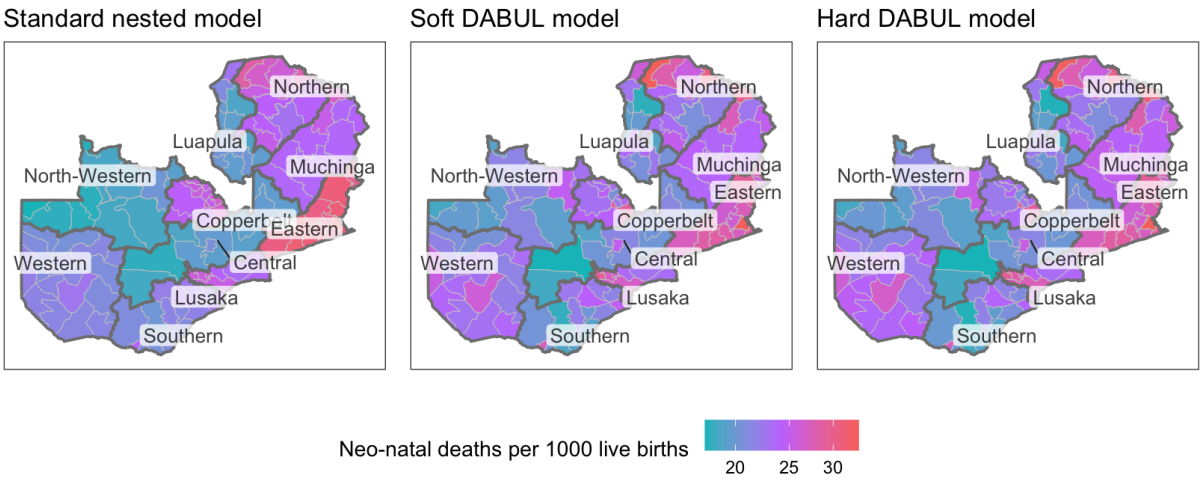


6. APPLICATION TO ZAMBIA DHS DATA

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 same priors and hyperpriors as in the motivating example. 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.

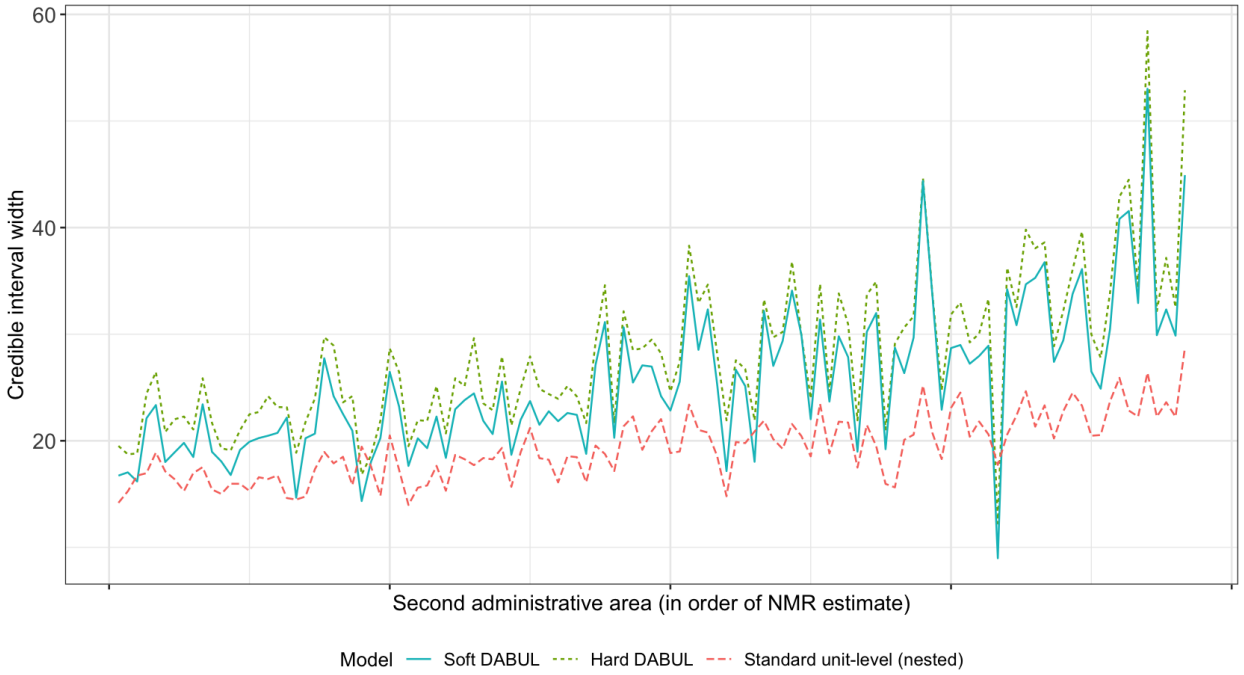
Figure 10 provides a map of the NMR estimates resulting from each of the two models

Figure 10. Map of NMR estimates at the second administrative 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 second administrative level nested based on first administrative level.



and figure 11 compares the credible interval width for each second administrative area. A scatter plot comparing the NMR estimates from each of the two models is presented in Appendix C. From these figures, we observe that the NMR point estimates are fairly similar between models, with the DABUL estimates having less shrinkage towards the first administrative 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. In figure 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 larger design-based variance estimates, 0.086 and 0.091, respectively, compared to the other first administrative area level direct estimates, which have design-based variance estimates between 0.034 and 0.056. This is an expected consequence of taking the uncertainty of the benchmarks into account. Additionally, the

Figure 11. Width of 90% credible intervals of NMR estimates (in deaths per 1000 live births) at the second administrative level in Zambia (2009-2013) using the DABUL models, as compared to a standard unit-level Bayesian model.

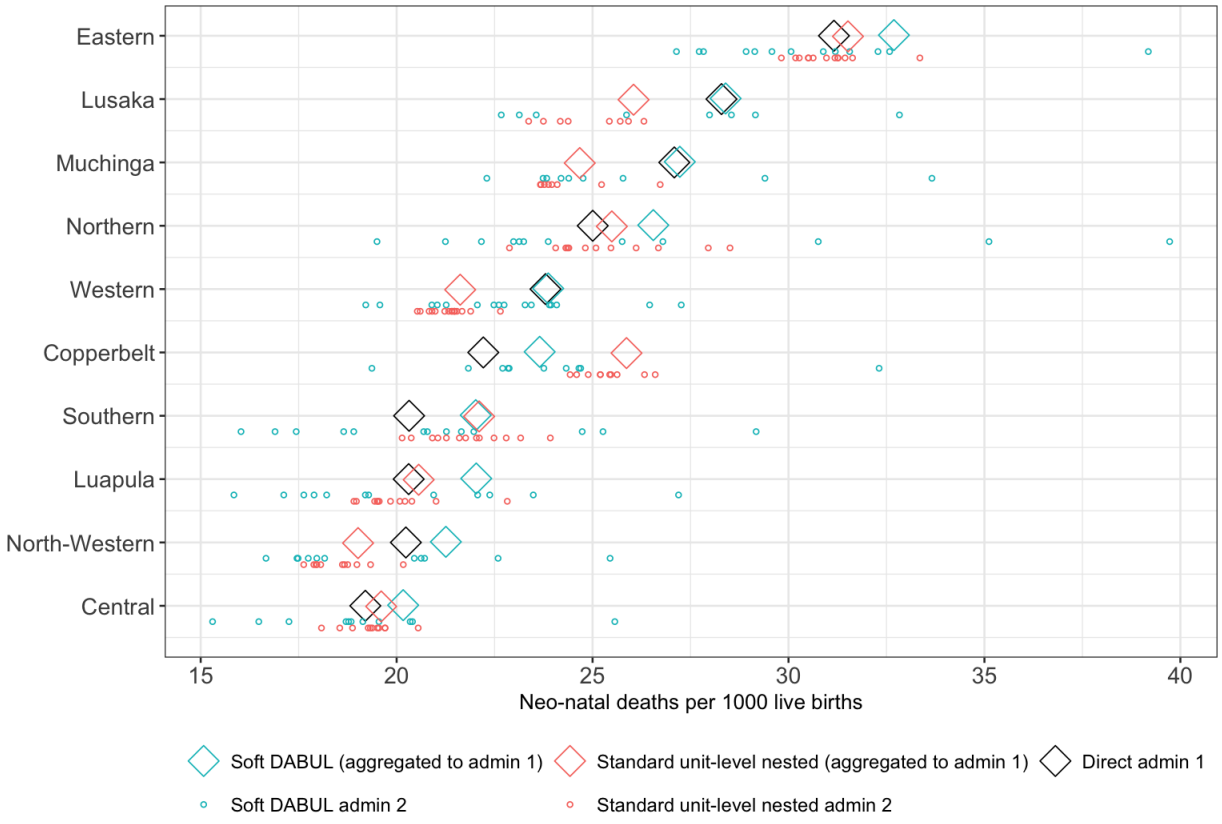


DABUL models produce results with reduced shrinkage, so if the more populous second administrative areas happen to be more extreme areas, the aggregated estimate will be pulled in that direction. If estimates under the soft DABUL model do not have sufficient agreement in aggregation for a particular 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.

7. DISCUSSION

We first observe that when there are multiple nested areas, using nested models is to be recommended to reduce overshrinkage, with the nested unit-level model being a very standard choice. Here we have developed two alternative DABUL models. The DABUL models could also be extended to include covariates, which is encouraged whenever

Figure 12. NMR estimates at the second administrative 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 second administrative level estimates and the colored diamonds denote their aggregations to the first administrative level. The black diamonds denote Hájek direct estimates at the first administrative level.



covariates with predictive power are available. We do not include covariates here because previous work has shown that available covariates in low and middle income countries have little effect on predictive power in spatial models for NMR (Golding et al., 2017,Wakefield et al., 2019). It is straightforward to include covariates in the DABUL models, though the aggregation step can introduce extra error, as is true for any loglinear model.

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 $\tilde{Y}_i|Y_{i+}$ 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 normal distributions implied by the Gaussian likelihood.

We have demonstrated the value of a 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 which imposes an 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. Further study is necessary to determine if this conclusion holds over a wider array of scenarios, however, the soft DABUL model remains the more conservative, principled approach because of its proper accounting for the large uncertainty of the benchmarks.

Data Availability Statement

The processed 2014 Zambia DHS data used in this article are available in this article's online supplementary material.

References

Astfalck, L., Sen, D., Patra, S., Cripps, E., and Dunson, D. (2018). Posterior projection for inference in constrained spaces. *arXiv e-prints*, pages arXiv–1812.

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.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36(2):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.

Breidt, F. J. and Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science*, 32(2):190 – 205.

Corsi, D. J., Neuman, M., Finlay, J. E., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology*, 41(6):1602–1613.

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

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*, page qnae023.

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

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, 59(2):286–295.

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(366a):269–277.

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(10108):2171–2182.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523.

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(1):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(260):663–685.

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*.

Lumley, T. (2024). Survey: Analysis of complex survey samples. R package version 4.4.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):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(2):137–152.

Okonek, T. and Wakefield, J. (2024). A computationally efficient approach to fully Bayesian benchmarking. *Journal of Official Statistics*, 40(2):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(2):318–342.

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(1):73–84.

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(4):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(2):319–392.

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.

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(1):123–150.

UNICEF et al. (2015). Multiple Indicator Cluster Survey. Statistics and Monitoring. *Resource document. UNICEF. http://www. unicef. org/statistics/index_24302. html. Accessed*, 24.

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(9):2614–2634.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418.

Wang, J., Fuller, W. A., and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34(1):29.

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*.

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(3):431–439.

Yu, C. and Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319–335.

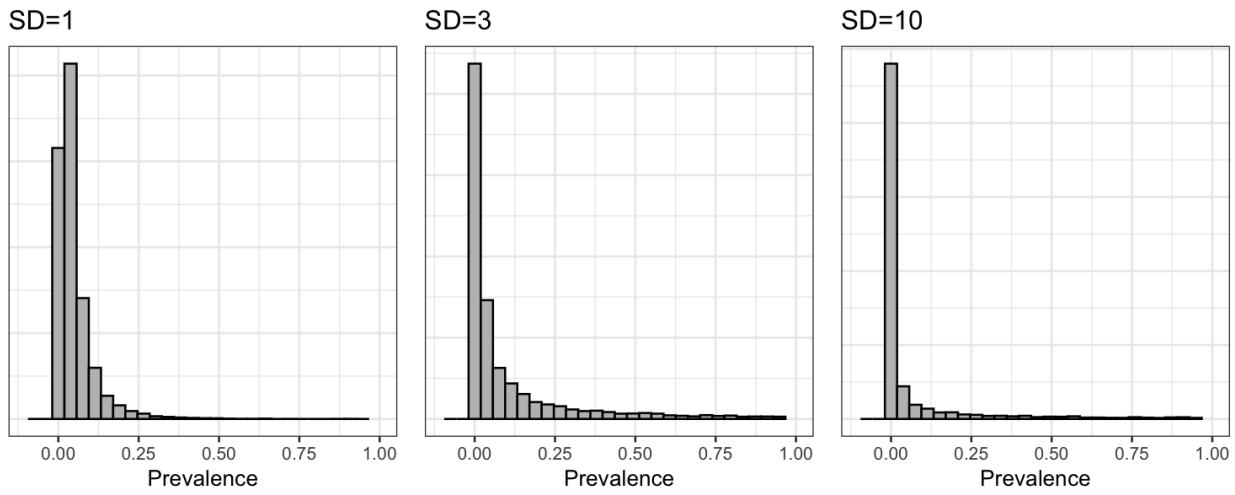
Zambia Central Statistical Office, Zambia Ministry of Health, and ICF International (2014). Zambia Demographic and Health Survey 2013-14. Rockville, Maryland, USA.

Zhang, J. L. and Bryant, J. (2020). Fully Bayesian benchmarking of small area estimation models. *Journal of Official Statistics*, 36(1):197–223.

Appendix A 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 13 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.

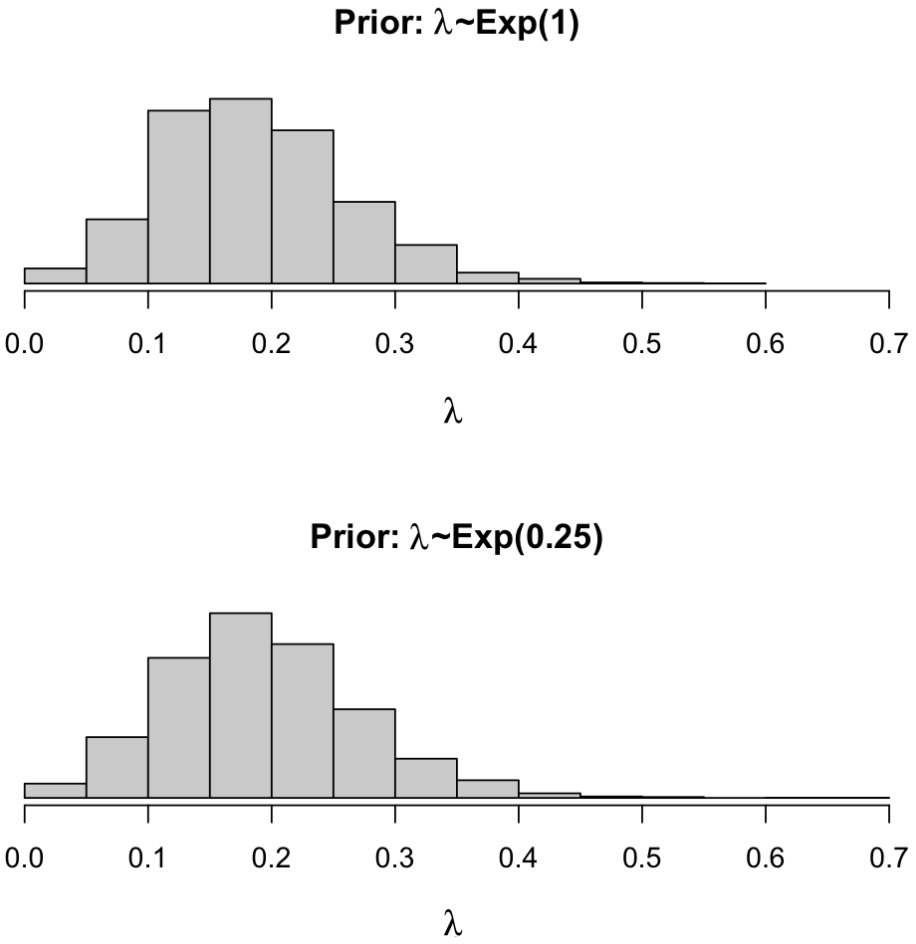
Figure 13. Exponential transformations of Normal distributions with mean -3.5 and variances 1^2, 3^2, and 10^2, respectively.



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 14 shows that the resulting posterior distributions of the overdispersion parameter are nearly identical.

Figure 14. Posterior distribution of the overdispersion parameter in the nested negative binomial model, fit to Zambia neonatal mortality data, under two different choices of prior.



Appendix B Additional simulation results

Figure 15. Discrepancy between aggregated model-based second administrative level estimates and design-based first administrative 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.

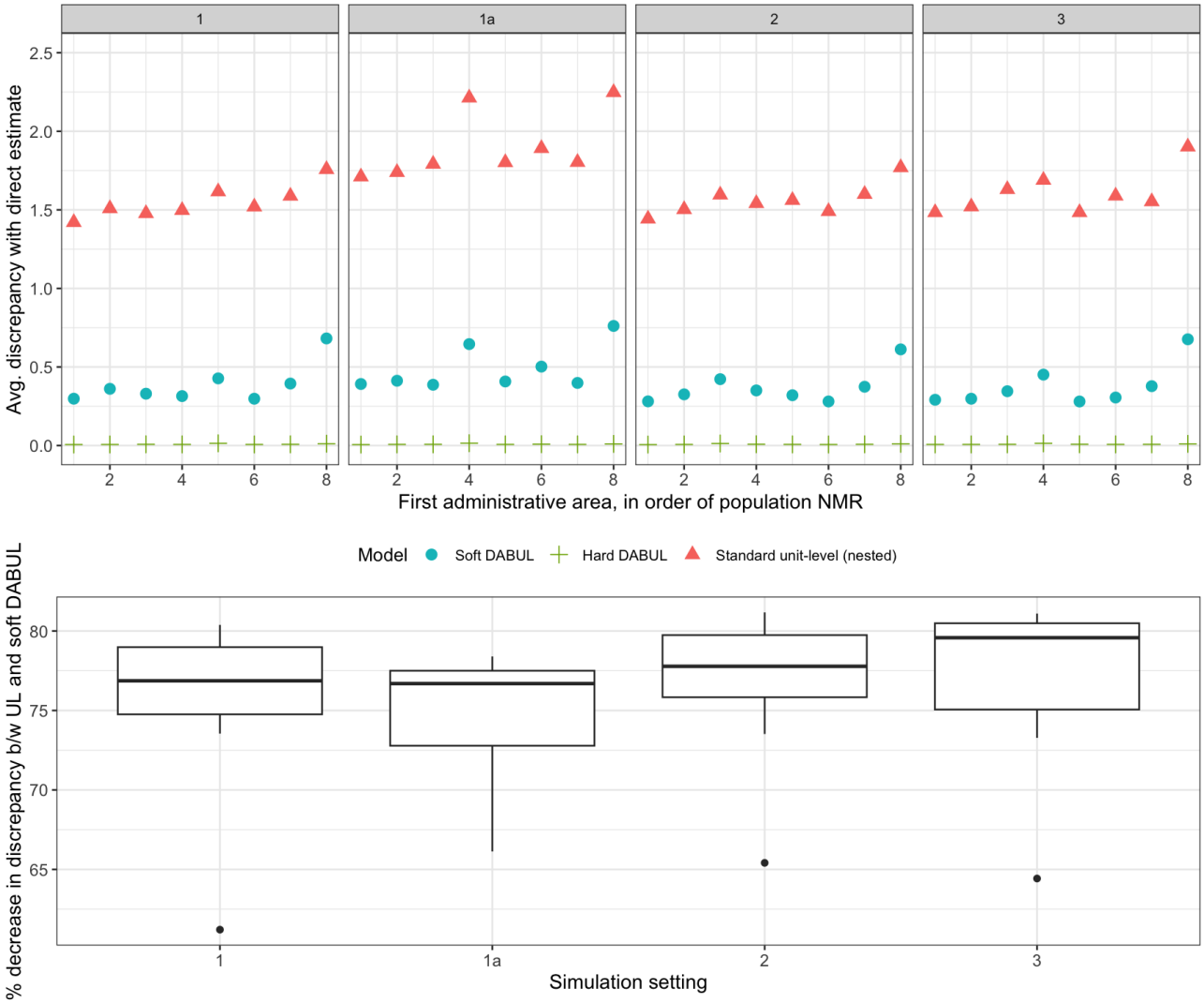
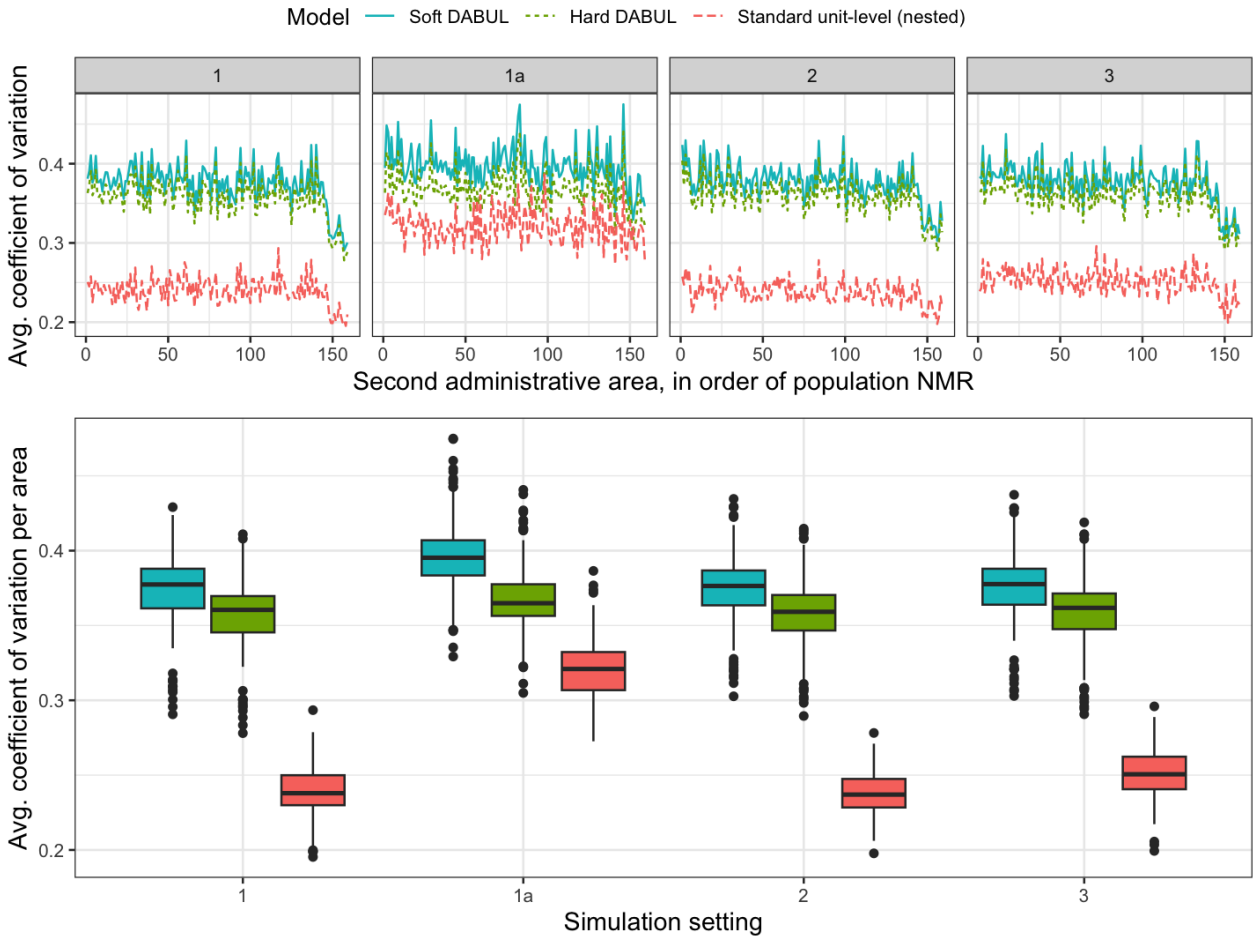


Figure 15 is similar to figure 6, except we only examine the areas for which the standard unit-level model has significant discrepancy with the direct estimates (greater than 0.001). We observe similar patterns to those in the previous figure, but the decrease in discrepancy, displayed in the bottom panel, is more significant, with aggregated soft

DABUL estimates having 74.7-76.8% lower discrepancy with the first administrative 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 16. Coefficient of variation for second administrative level estimates, across 500 simulations for each of 4 settings.



Appendix C Additional plot of Zambia NMR estimates

Figure 17. NMR estimates (deaths per 1000 live births) at the second administrative 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 second administrative level nested based on first administrative level.