Submitted to Statistical Science

The Two Cultures of Prevalence Mapping: Small Area Estimation and Model-Based Geostatistics

Jon Wakefield, Peter A. Gao, Geir-Arne Fuglstad and Zehang Richard Li

Abstract. In low- and middle-income countries (LMICs), accurate estimates of subnational health and demographic indicators are critical for guiding policy and identifying disparities. Many indicators of interest are proportions of binary outcomes and the task of estimating these fractions is often called prevalence mapping. In LMICs, health and vital records data are limited, so prevalence mapping relies on data from household surveys with complex sampling designs. However, estimates are often desired at spatial resolutions at which data are insufficient for reliable weighted estimation. We review two families of approaches to prevalence mapping: small area estimation (SAE) methods (from the survey statistics literature) and model-based geostatistics (MBG) methods (from the spatial statistics literature). SAE models can be "area-level" or "unit-level" and commonly use area-specific random effects and rely upon high-quality covariate data, often obtained from administrative sources. Unit-level models for binary responses are relatively underdeveloped. MBG approaches explicitly specify binary response models, incorporate continuous spatial random effects, and leverage alternative sources of data such as those arising from satellite imagery. These models are usually studied under a Bayesian framework. SAE methods often address the design by incorporating sampling weights or modeling the sampling mechanism. Two delicate issues arise when using MBG methods for prevalence mapping. First, aggregating unit level predictions to create area-level summaries requires population-level information that is rarely directly available. Second, MBG approaches typically assume the sampling design is ignorable. We review both SAE and MBG approaches to prevalence mapping, and argue that binary response models can be improved using insights from both the survey sampling and the spatial statistics literature. We highlight these issues using household survey data from the Zambia 2018 Demographic Health Survey to estimate subnational HIV prevalence for woman aged 15–49.

Key words and phrases: complex survey designs, demographic and health indicators, design-based inference, Gaussian random field models, geostatistical models.

Jon Wakefield is Professor, Departments of Statistics and Biostatistics, University of Washington, USA (e-mail: jonno@uw.edu). Peter A. Gao is Assistant Professor, Department of Mathematics and Statistics, San José State University, USA (e-mail: peter.gao@sjsu.edu). Geir-Arne Fuglstad is Professor, Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway (e-mail: geir-arne.fuglstad@ntnu.no). Zehang Richard Li is Assistant Professor, Department of Statistics, University of California Santa Cruz, USA (e-mail: lizehang@ucsc.edu).

1. INTRODUCTION

In low- and middle-income countries (LMICs), subnational estimates of key health and demographic variables are often used for determining progress toward targets, designing effective policy interventions and assessing disparities. For example, the United Nations General Assembly's 2030 Agenda established a set of Sustainable Development Goals for global development (General Assemby of the United Nations, 2015), each associated with specific aims such as eliminating preventable deaths under five years of age and reducing maternal mortality (UN

System Chief Executives Board for Coordination, 2017). Accurate tracking of the goals requires subnational estimates of indicators such as the neonatal mortality rate (proportion of children who die within the first month of life), vaccination coverage, poverty measures, contraceptive use and female attainment of secondary education. We focus on indicators that can be expressed as a *prevalence*, meaning the proportion of individuals in a group that satisfy a specific condition. Producing subnational estimates of the prevalence of an outcome is often called *prevalence mapping*.

When the data are sparse, this problem is an example of small area estimation (SAE), itself defined as the task of "producing reliable estimates of parameters of interest...for subpopulations (areas or domains) of a finite population for which samples of inadequate sizes or no samples are available" (Rao and Molina, 2015, p. xxiii). Much of the work on SAE is based on finite population estimation from the survey sampling literature. Historically, this research has focused on applications in high-income countries with readily available auxiliary data, often derived from administrative sources. However, in LMICs, health and vital records data are limited and although the most reliable sources of data are from household surveys, traditional SAE methods are not universally used. As an alternative, a body of research has emerged that uses so-called model-based geostatistics (MBG) approaches, drawn from the spatial statistics literature, for prevalence mapping (Diggle and Giorgi, 2016).

We draw a distinction between two distinct families of approaches to prevalence mapping in LMICs. The first route, which we call the SAE approach, comprises methods drawn primarily from the survey sampling literature. All major surveys in LMICs collect data via a complex survey design, which involves stratification and cluster sampling, which we detail shortly. SAE methods often account for the complex survey design explicitly and, when sparsity of data prevents the use of design-based weighted estimators, frequently rely upon linear mixed effects models (LMEMs), leveraging auxiliary administrative data. The second route, which we call the MBG approach, encompasses a broader collection of spatial statistical modeling methods that have been adapted specifically for estimation in low-data settings. These strategies generally model the risk of experiencing the outcome of interest using a spatially continuous surface and use alternative sources of data such as satellite imagery. However, MBG approaches often pay less attention to the sampling design of the surveys, or to the nuances of aggregating point level predictions, which is required to produce area-level estimates.

Inferentially, MBG approaches often take a Bayesian approach and use modern computational techniques such as the integrated nested Laplace approximation (INLA) of Rue et al. (2009) which provides a computationally efficient alternative to Markov chain Monte Carlo (MCMC) approaches. Subjective priors may also be specified. The SAE literature has seen an adoption of Bayesian methods, but computation is generally carried out with MCMC and "uninformative priors" are often favored (Rao and Molina, 2015, Chapter 10).

In this article, we review both SAE and MBG approaches and examine their differences. In some respects, these differences result from a basic divergence with respect to the target estimand of interest. SAE approaches estimate *prevalence*, and focus on a finite population of inference. The approach begins with a fixed set of individuals in a sampling frame. By contrast, MBG approaches assume the existence of a continuous spatial *risk* surface, representing the probability of experiencing the outcome of interest for a hypothetical individual located at any particular location, and the average risk over an area is typically used as a proxy for the prevalence of that area. This review follows in the footsteps of previous reviews of SAE (Ghosh and Rao, 1994; Pfeffermann, 2013) but focuses on prevalence mapping in LMICs and gives an overview of MBG methods, which are less familiar to those in the design-based SAE community. Similarly, those in the spatial community may be less familiar with design-based finite population approaches.

To motivate the discussion, we briefly introduce an application in which we examine spatial variation in female HIV prevalence. We begin with key definitions: all countries are divided into principal administrative divisions, called Admin-1 regions, which are further subdivided into secondary administrative regions, called Admin-2 regions. In our example, we wish to characterize variation in HIV prevalence in women across 10 provinces (Admin-1 areas) and 115 districts (Admin-2 areas) of Zambia, based on data from the 2018 Demographic Health Survey (DHS), which used stratified two-stage unequal probability cluster sampling. The clusters correspond to census enumeration areas (EAs) and the strata are the urban/rural status of the cluster crossed with Admin-1 areas. The two stages of sampling are clusters within strata and households within clusters. For DHS (and more recent MICS) surveys, responses are reported with their cluster location, so that all individuals in the cluster are reported to be located at the geographical location of the cluster. Figure 1 maps the approximate locations of the 545 sampled clusters, indicating which were urban/rural in the original sampling frame based on the 2010 census. HIV infection and geographic data were recorded for 12,893 women aged 15–49, from the 13,625 sampled households. The sampling units are households, while the observation units are women. The Admin-1 and Admin-2 boundaries are also shown in Figure 1, and we see a relatively large number of both urban and rural clusters in each Admin-1 area (as expected since these are planned domains),

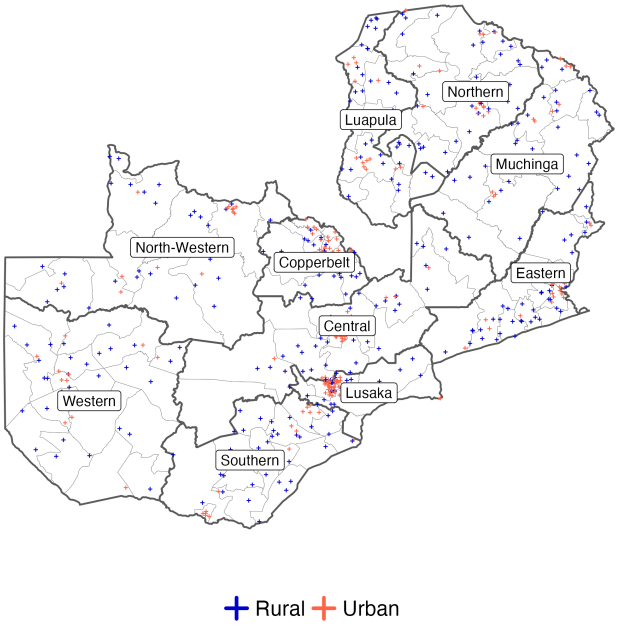


FIG 1. *Locations of sampled urban and rural clusters in the 2018 Zambia DHS (jittered to preserve privacy) with Admin-1 and Admin-2 boundaries indicated and Admin-1 labels.*

but sparser sampling in Admin-2 areas (unplanned domains). Government agencies often require high precision when reporting domain estimates. For example, Statistics Canada, has guidelines (Cloutier and Langlet, 2014, Table 5) for area-level estimates: for an area with a coefficient of variation (CV) of less than 16.7% the estimate can be used without restriction, when the CV is above this but less than 33.3%, it should be used with caution, and an estimate with a CV greater than 33.3% is deemed too unreliable to be published. For this example, all Admin-1 (weighted) estimates have CVs less than 16.7%. However, at the Admin-2 level, out of 115 areas, three areas have no clusters, and fourteen have insufficient data to yield usable variance estimates for weighted estimates; only 27 of the areas have a CV smaller than 16.7% while 32 of the areas have a CV larger than 33.3%. Hence, at the Admin-2 level, there is a need for modeling. In Section 6, we fit a variety of SAE and MBG models that include spatial smoothing terms and auxiliary data.

The structure of the paper is as follows. The journey begins in Section 2, with a description of household surveys and an outline of the particular characteristics of prevalence mapping in LMICs that complicate the use of traditional SAE methods and have led to the emergence of MBG approaches. In Section 3, we review traditional SAE methods, and discuss the finite population estimation perspective and design-based and model-based estimators. Section 4 considers MBG approaches, focusing on a number of issues including the aggregation of a continuous surface based on limited population information. In Section 5, we discuss a number of miscellaneous topics related to prevalence mapping. There is a huge literature on the SAE and MBG approaches and what we review, in the context of LMICs, is by necessity selective and not comprehensive. Section 6 compares SAE and MBG methods in the context of mapping HIV prevalence among women aged 15–49 in Zambia, and provides suggestions for modeling, taking elements from both of the "cultures". The paper ends with a discussion in Section 7. Model details, additional results and code can be found in the Supplementary Materials (Wakefield et al., 2025). The code is also available at `https://github.com/peteragao/two-cultures-prev-map`.

2. PREVALENCE MAPPING USING HOUSEHOLD SURVEY DATA

2.1 Issues with Household Survey Data in LMICs

We begin by introducing some of the challenges that complicate prevalence mapping in LMICs, before contrasting, at a high level, traditional SAE and MBG approaches.

Complex sampling design

In many LMICs, household surveys including the DHS[1] and the Multiple Indicator Cluster Surveys (MICS)[2] collect data on health outcomes. Both survey programs implement a stratified multistage design. We will describe the DHS sampling design in detail, since it is relevant to our HIV prevalence mapping example. It is representative of the designs used by other large household surveys, including MICS, that are carried out in LMICs.

In a given country, DHS aims to conduct household surveys approximately every 5 years, with the *sampling units* being households. The sampling frame is typically constructed from the most recent national census and aims to enumerate the households in the country at the time of the census and groups households into EAs (clusters). In the survey sampling literature the clusters are referred to as primary sampling units (PSUs). Clusters are classified as urban or rural by national authorities at the time of the census, based on a country-specific definition (Alkema et al., 2013), but in some cases the classifications are updated between the time of the census and the survey. For example, the last census in Nigeria was 2006, but a new definition was taken to label each cluster as urban or rural, in 2017.

For each survey, households are generally sampled using a stratified two-stage cluster sampling design, as detailed by ICF International (2012). The strata are usually

[1] `https://dhsprogram.com/`
[2] `https://mics.unicef.org/`

defined by splitting each Admin-1 area into urban and rural parts. In the first stage of sampling, a specific number of urban clusters and rural clusters are sampled from each Admin-1 area using probability-proportional-to-size (PPS) sampling, where a cluster's size is often defined as the number of households in the cluster. Urban clusters are often over-sampled. Since the list of households based on the census sampling frame will be outdated, the surveyor re-enumerates the households in each selected cluster. In the second stage of sampling, a fixed number of households (25 in the 2018 Zambia DHS) are sampled with equal probability from the updated list of households in each cluster. In the language of survey sampling, the households are termed secondary sampling units (SSUs). Household members are interviewed to provide information on health and demographic variables. As a result of the sampling procedure, each household has an inclusion probability that describes the *a priori* probability that it will be included in the sample. The response rate of DHS surveys are high, but the reported weights do contain a non-response adjustment. Only the final weights are reported, and not the constituent parts, and the weight is scaled, so the weights alone cannot be used to estimate totals, only ratios, such as the prevalence.

Sparse response data at the resolution of interest

Often, estimates are desired not only at the Admin-1 level, but also at the Admin-2 level, as many health policy decisions are made at this level (Hosseinpoor et al., 2016). At this resolution, many areas may have severely limited data on an outcome of interest (and some may have no response data at all), since the surveys are typically powered to produce reliable estimates for Admin-1 regions. As a result, weighted estimation may not be feasible and model-based approaches are needed.

Acknowledging the Design and Aggregation

Failing to account for stratification, clustering, and unequal probability sampling in a model-based approach can produce biased small area estimates (because of informative sampling in which the selection probabilities depend on the outcome values) and poorly calibrated estimates of uncertainty. Urban households are often overrepresented in in DHS and if the outcome is associated with urban/rural location, then estimates of prevalence will be biased if the oversampling is not acknowledged. Unfortunately, relevant design variables, such as the number of households in a cluster, which is used in the PPS sampling, are unavailable. In addition, model-based methods that treat observations as independently and identically distributed (iid) are inappropriate for cluster sampling.

When moving from points to areas, aggregation of a risk surface is required and, formally, the locations of all of the target population, along with any associated auxiliary information that is used in a non-linear (binary data) model, is required, but this is never available. Hence, by necessity, the aggregation will be an approximate operation.

Limited availability of high-quality auxiliary data

Traditional SAE model-based approaches often rely on covariate information from administrative sources such as national censuses. If recent and reliable census data is unavailable, as is often the case in LMICs, then alternative data is needed to explain differences between areas.

2.2 An Overview of SAE and MBG Approaches

We first establish some terminology (Rao and Molina, 2015). *Direct estimators* use response data from the area in question only when estimating the prevalence in an area. In areas where data are limited, *indirect estimators* use statistical models that share information between areas. SAE approaches, which we review in depth in Section 3, typically account for features of survey design to minimize bias. The most basic SAE methods are based on Hájek estimators that incorporate the sampling weights. In the SAE literature, models are usually categorized as either *area-level*, meaning that area-level weighted estimators are treated as response data in a smoothing model, or *unit-level*, meaning that individual survey responses are modeled. The well-known Fay-Herriot model (Fay and Herriot, 1979) is the most popular area-level model. In the first step of the Fay-Herriot approach, weighted estimates and associated estimates of the sampling variances are computed for each area. In the second step, these direct estimates are used as response data in a smoothing model that incorporates area-level covariates and Gaussian random effects, that are assumed (in the original formulation) to be iid across areas. The resulting estimators account for the sampling design because they take as input the direct estimates and their associated variances.

Unfortunately, if the target areas contain few samples, direct estimators (and the corresponding variance estimators) may be unreliable. For example, when using DHS data to map health outcomes at the Admin-2 level, it is common for many areas to have few or even zero sampled clusters, making direct estimation infeasible. Area-level modeling may still be possible, particularly if variance smoothing is incorporated, but for sparse data, unit-level models are the only option. These models specify a sampling model for individual responses, and are more flexible, incorporating unit-level covariates and smoothing via random effects. However, unit-level approaches often assume the sampling design is *ignorable*, meaning that the same response model is appropriate for the sampled and non-sampled units. But, as already noted, the data required to fit an appropriate model may be unavailable for the sampled responses, and so we may not be able to specify a model for which the design is ignorable.

The variance model should also account for the effects of clustering. Moreover, unit-level models produce individual level predictions which must be aggregated to produce area-level estimates. For non-linear models and unit-level covariates, aggregation requires covariate information for each unobserved unit, but in the LMICs context, sampling frame information, such as the locations of all clusters and the population sizes within those clusters is not available. In Sections 3.4 and 4.3 we describe approximate aggregation strategies.

Like the unit-level models used in SAE, MBG approaches to prevalence mapping in LMICs treat individual responses as data and leverage recent advances in continuous spatial modeling and alternative sources of auxiliary information, but less attention is paid to the survey design and in particular little reference is made to the sampling frame and finite populations.

As already briefly discussed, SAE approaches estimate finite population *prevalences* while geostatistical approaches estimate a continuously indexed spatial surface representing *risk*. For example, when estimating an outcome such as the proportion of children who die within the first month of life, geostatistical models assume the existence of an unobservable spatial surface representing the hypothetical risk of death at *any* location. Each point-referenced observation is viewed as the result of an individual being exposed to the risk at that given location. Continuous spatial models are widely used in the environmental sciences to represent quantities such as pollutant levels or climate-related variables over a continuously-indexed study area and this is certainly reasonable in many contexts where the outcome variable (e.g., airpollution or temperature) does truly exist at every location. SAE approaches thus perform *prevalence mapping*, estimating the empirical proportions for areas, while MBG approaches approximate prevalences through *risk mapping*, extracting areal estimates by averaging over a spatially continuous risk surface.

Practically, geostatistical approaches borrow strength across space using Gaussian random field (GRF) models and spatial covariates, which may help to address some of the shortcomings of SAE methods in low-data settings. Using continuous spatial random effects enables smoothing even when response data is sparse or unavailable. MBG approaches often use covariates derived from data sources such as satellite imagery. Such data, which may include variables such as intensity of night time lights or vegetation indices, are usually provided as raster data with measurements on a high-resolution grid across the domain of interest. Risk predictions can be made for any location covered by the covariate rasters. Of course, traditional SAE unit-level models could utilize satellite imagery data, but following the rationale of SAE approaches, for non-linear models, one would require the set of locations of all units in each area, which is unavailable. In practice, geostatistical unit-level models overcome this issue by generating predictions on high-resolution (often 1×1 km or 5×5 km) pixel maps across the spatial domain of interest, and weighting by population density. Rasters of population density are commonly used for weighting, but the population density values are themselves modeled quantities with (often, unspecified) uncertainty and it is not clear how to propagate this uncertainty into the areal estimates. In addition, predictions that are optimal at the pixel level are not guaranteed to result in optimal predictions at coarser levels such as Admin-2 when aggregated using imprecise and inaccurate knowledge about the population. Benchmarking, to ensure consistency across nested geographical levels, is not carried out as part of MBG analyses, though there is a large SAE literature on this issue, see Rao and Molina (2015, Section 6.4.6).

Interpreting and quantifying the uncertainty of small area estimates produced by MBG approaches is difficult. As noted, geostatistical models often assume implicitly that the sampling procedure is ignorable under the specified model and may not adjust for relevant design variables.

Traditional SAE approaches have been applied to a number of indicators in LMICs by international organizations. These include the World Bank, in the context of poverty mapping (Molina et al., 2019; Corral et al., 2021; Edochie et al., 2025) and the United Nations Inter-Agency Group on Mortality Estimation (UN-IGME), who produced subnational estimates of child mortality indicators using SAE methods (Wakefield et al., 2019; Wu et al., 2021), leaning heavily on discrete spatial models. HIV prevalence mapping is also common using SAE methods, see for example, Gutreuter et al. (2019).

Similarly, a number of prominent research organizations produce subnational risk maps using MBG approaches, including the Institute for Health Metrics and Evaluation (IHME) (Golding et al., 2017; Burstein et al., 2018; Osgood-Zimmerman et al., 2018; Graetz et al., 2018; Mosser et al., 2019; Local Burden of Disease Vaccine Coverage Collaborators and others, 2021), WorldPop (Utazi et al., 2018, 2020; Ferreira et al., 2022) and the Incorporated City Fund (ICF, who historically implemented DHS) (Mayala et al., 2019; Janocha et al., 2021). Other examples include Giorgi et al. (2015); Diggle and Giorgi (2016) and Giorgi et al. (2018).

As evidence of the cultural divide between the SAE and MBG approaches, prominent SAE texts (Rao and Molina, 2015; Pratesi, 2016; Morales et al., 2021) and the SAE guidelines report of Corral et al. (2022) cover spatial modeling only briefly and do not mention the MBG approach. From the other side of the divide, SAE is not mentioned in the index of the book-length treatment on MBG approaches of Diggle and Giorgi (2019), and the classic text of Rao and Molina (2015) is not referenced anywhere in the book.

3. SMALL AREA ESTIMATION

In this section, we review area-level and unit-level SAE approaches more formally. We also outline direct estimation since it is an important ingredient for area-level modeling, though it is not strictly an SAE approach, since no between-area modeling is carried out. We begin by establishing notation and briefly reviewing the principles of finite population inference.

3.1 Finite Population Estimation and Inference

We consider a finite target population in a study area that consists of N *observation units*. The target population can be represented as the set of indices $U = \{1, 2, \ldots, N\}$ where observation unit k has an associated response value $y_k \in \{0, 1\}$ for $k \in U$. When a country is divided into m administrative areas, these areas partition the population U into m disjoint subpopulations, $U = U_1 \cup U_2 \cup \cdots \cup U_m$, where U_i is the set of units that belong to area i.

This notation is standard for direct estimates and area-level models. For unit-level models and MBG approaches, it is easier to use nested indexing. Let N_i denote the number of units in U_i for $i = 1, \ldots, m$. Then, in anticipation of our SAE unit-level development, we can index population responses as y_{ij} for unit $j = 1, \ldots, N_i$ in area $i = 1, \ldots, m$. There is a one-to-one correspondence between unit j in area i and a unit $k[i, j] \in U$, and we switch between notations depending on the type of model considered.

The target parameters are the area-specific prevalences. With the former notation,

$$p_i = \frac{1}{N_i} \sum_{k \in U_i} y_k, \quad i = 1, \ldots, m, \tag{1}$$

and in the latter notation

$$p_i = \bar{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}, \quad i = 1, \ldots, m.$$

We let $S = \{k_1, \ldots, k_n\} \subset U$ denote the set of n sampled indices, where S is partitioned into m areas, $S = S_1 \cup \cdots \cup S_m$. We assume a probability sampling design, and for all $k \in U$, let π_k denote the probability that unit k is sampled and $w_k = 1/\pi_k$ denote the design weight for unit k. In general, weights may contain non-response and post-stratification adjustments. In the DHS the weights include the former but not the latter. This is common for household surveys in LMICs, where observed population data (from, for example, a census) is generally neither recent nor reliable enough for the purpose of post-stratification.

When conducting inference based on survey data, we can distinguish between design-based inference and model-based inference. The design-based approach assumes a fixed finite population of responses, and inference is based on the randomization distribution over the space of possible samples defined by the sampling design. Asymptotic analysis for sample survey estimators usually considers a sequence of finite populations of growing size and an associated sequence of sampling designs. The sample size is typically assumed to grow at the same rate as the finite population and various other assumptions can be made about the sampling design as well as the finite population responses. Under this design-based framework, estimators that are unbiased, or at least design consistent, are preferred. For an overview of sample survey asymptotics, see Breidt and Opsomer (2017). The model-based approach assumes that the finite population responses are drawn from a data-generating model and inference is based on this model, though design consistency of model-based estimates is desirable. The sampling design is assumed to be ignorable with respect to the model. SAE methods adopt both design-based and model-based perspectives, but generally design-based inference is used when there is sufficient data for using direct weighted estimators while model-based inference is more commonly used when data is more limited.

3.2 Direct Estimates

Direct weighted estimators acknowledge the sampling design by incorporating sampling weights. For example, the Hájek estimator (Hájek, 1971) replaces the numerator and denominator in Equation (1) with survey weighted estimates,

$$\hat{p}_i^{\mathrm{w}} = \frac{\sum_{k \in S_i} w_k y_k}{\sum_{k \in S_i} w_k}, \quad i = 1, 2, \ldots, m. \tag{2}$$

Under the design-based perspective, the random quantity in (2) is S_i. The estimators of the numerator and the denominator are unbiased, and $\hat{p}_i^{\mathrm{w}}$ is consistent for p_i. Design-based estimates of the variance can be constructed with linearization-based approximations or resampling methods such as jackknife or bootstrap estimators (Lohr, 2021; Pedersen and Liu, 2012).

Achieving reliable weighted estimates with low variances requires sufficient sample sizes for each target area, making direct estimators inadequate for SAE with small domains (Rao and Molina, 2015, Chapter 2). For example, the m administrative areas may be unplanned areas (Lehtonen and Veijanen, 2009) which do not match the strata used for sampling, so that some administrative areas have sparse, if any, data.

3.3 Area-Level Models

Simultaneous modeling of the data from all areas can increase the precision of estimates, relative to the direct alternatives, effectively increasing the sample size.

The most commonly used approach is the two-stage Fay-Herriot model (Fay and Herriot, 1979) that uses auxiliary information in the form of area-level covariates and introduces area-specific random effects. In the original paper, the random effects were assumed to be iid normal and a frequentist approach to inference was taken. Let $\hat{\theta}_i^{\mathrm{w}}$ represent a direct estimate of an area-level parameter and V_i be the sampling variance of $\hat{\theta}_i^{\mathrm{w}}$, which is estimated from data but often treated as known. The Fay-Herriot model is:

$$\hat{\theta}_i \mid \theta_i \sim_{\mathrm{iid}} \mathrm{N}(\theta_i, V_i), \tag{3}$$

$$\theta_i = \alpha + \boldsymbol{x}_i^{\top}\boldsymbol{\beta} + u_i, \tag{4}$$

$$u_i \mid \sigma_u^2 \sim_{\mathrm{iid}} \mathrm{N}(0, \sigma_u^2), \quad i = 1, \ldots, m, \tag{5}$$

where α is the intercept, $\boldsymbol{x}_i$ are area-level covariates with associated coefficients contained in the vector $\boldsymbol{\beta}$ and u_i represent between-area differences not explained by covariates. Equation (3) is referred to as the *sampling model* and equation (4) the *linking model*. The Fay-Herriot model implicitly acknowledges the sampling design through the use of sampling weights when computing the direct estimate and its standard error, $\sqrt{V_i}$. The direct estimates may be transformed to make the normal approximation to the sampling distribution more accurate and to ensure that the resultant prevalence estimates are on $[0, 1]$. In particular, for each area i, we can define $\hat{\theta}_i^{\mathrm{w}} = h(\hat{p}_i^{\mathrm{w}})$ where $h(\cdot)$ represents a transformation. The Fay-Herriot model can then be applied to model the transformed $\hat{\theta}_i^{\mathrm{w}}$ parameters with the resulting modeled estimates being transformed back to the original range. The sampling variance of the transformed $\hat{\theta}_i^{\mathrm{w}}$ parameters must be estimated or approximated. Transformations used with the Fay-Herriot model include the log (Fay and Herriot, 1979; Slud and Maiti, 2006), logit (Jiang and Tang, 2011; Mercer et al., 2015), dual power (Sugasawa and Kubokawa, 2015), and arcsine (Hirose et al., 2023). We emphasize that θ_i represents a finite population parameter, such as the logit prevalence, $\mathrm{logit}(p_i)$, for area i and not a super-population characteristic, such as the logit risk.

This model can be used to generate small area estimates using either frequentist or Bayesian methods. The Fay-Herriot model is an example of a LMEM and so the empirical best linear unbiased predictor (EBLUP), a standard summary for such models, can be used (Rao and Molina, 2015, Chapter 6). Bayesian approaches are also available, but in the SAE literature, "diffuse" (rather than subjective/informative) priors are often sought (Rao and Molina, 2015, Chapter 10). For SAE, under frequentist inference, a large emphasis is placed on estimating the mean squared error (MSE) associated with an areal estimate, with the bootstrap being a common tool. A good discussion of design-based and model-based approaches to MSE estimation is provided by Datta et al. (2011). Under a fully Bayesian approach, as described in Section 3.4, the natural analogue is the posterior variance, which is a model-based uncertainty measure. Another common uncertainty summary is the coefficient of variation, which is the square root MSE (or posterior standard deviation), divided by a point estimate.

Area-level models are popular as they are generally computationally easy to fit and they can improve upon the precision of direct estimators by explaining between-area differences using auxiliary data and random effects. The resulting estimates exhibit *shrinkage*, so that they are (conditional on the true value) biased, but their variance is reduced, and in general this results in a lower MSE for the complete collection of estimates, when compared to the MSE of direct estimates. Another major advantage of area-level models is that they account for the sampling design and achieve design consistency, in the sense that the sampling bias and variance of the estimated prevalence can be shown to converge to zero asymptotically under reasonable assumptions on the finite population and sampling design. Intuitively, since the Hájek estimator is design consistent, Fay-Herriot estimators, though based on a model, are design consistent also, as the within-area sample size increases.

In a high-income setting, the Fay-Herriot model has seen a vast number of applications, as summarized in Rao and Molina (2015) and Ghosh et al. (2020). A well-known example is the production of county-level estimates of poverty for the Small Area Income and Poverty Estimation (SAIPE) program in the United States (Bell et al., 2016).

We continue by discussing possible extensions to the basic Fay-Herriot model that can be used to address particular challenges of prevalence mapping in LMICs. The basic Fay-Herriot model is called a matched model in the sense that both the sampling and linking models are on the same scale (for example, logit), each with normal models. For a prevalence p_i and a nonlinear transformation $h(\cdot)$,

$$\mathrm{E}[\, h(\hat{p}_i^{\mathrm{w}}) \mid h(p_i) \,] \neq h(p_i),$$

even if $\hat{p}_i^{\mathrm{w}}$ is design unbiased (Rao and Molina, 2015, Section 4.2). This motivates an unmatched model in which it may be assumed that,

$$\hat{p}_i^{\mathrm{w}} \mid p_i \sim_{\mathrm{iid}} \mathrm{N}(p_i, V_i), \quad i = 1, 2, \ldots, m, \tag{6}$$

but with a linking model that assumes that $\theta_i = h(p_i)$, and not p_i, is normally distributed. You and Rao (2002b) assume a normal sampling model but a log-normal linking model. For estimating small area proportions, sampling model (6) along with a logistic linking model has been used (Mohadjer et al., 2012). Other unmatched models for proportions are explored by Liu et al. (2014) and Franco and Bell (2013). A general empirical Bayes (EB) approach for SAE with unmatched sampling and linking area-level models is outlined by Sugasawa et al. (2018).

The basic Fay-Herriot model assumes normally distributed iid area-level random effects, but the model may be easily extended to allow for random effects with other correlation structures. In particular, spatial and spatiotemporal covariance matrices may be used to smooth estimates across space and space-time. Chung and Datta (2020) describe a range of spatial models including a conditionally autoregressive (CAR) model; they provide a comparison of the traditional Fay-Herriot model with spatial alternatives, finding that a spatial area-level model can improve estimation when good covariates are not available. Ghosh et al. (1998) applied an intrinsic CAR (ICAR) prior (Besag and Kooperberg, 1995) to the random effects, while other methods have focused on the use of simultaneous autoregressive (SAR) spatial models (Saei and Chambers, 2003; Petrucci and Salvati, 2006; Pratesi and Salvati, 2008; Marhuenda et al., 2013). One approach that we have used extensively is a model that decomposes the random effect into the sum of an unstructured iid normal random effect and a spatial ICAR random effect, which is known as the BYM model (Besag et al., 1991). In our analyses in Section 6, we adopt the reparameterization known as the BYM2 model (Riebler et al., 2016), in which the vector of random area effects has structure,

$$\boldsymbol{u} = \sigma_u \left(\sqrt{1-\phi}\boldsymbol{e} + \sqrt{\phi}\boldsymbol{S} \right),$$

where σ_u is the total standard deviation, ϕ is the proportion of the variance that is spatial, $\boldsymbol{e}$ is a vector of iid standard normal random variables and $\boldsymbol{S}$ follows a scaled ICAR prior so that the geometric mean of the marginal variances of S_i is equal to 1, under a sum-to-zero constraint that is imposed to ensure identifiability when there is an intercept in the model (Rue and Held, 2005). More details on this model are contained in the Supplementary Materials.

Spatial models have been extensively used in disease mapping (Wakefield et al., 2000). Disease mapping is distinct from SAE in that it is based on a full enumeration of events (up to accurate case enumeration). Hence, the prevalence is observed, but one is interested in risk. Consequently, model-based, as opposed to design-based, inference is carried out. Generalized linear mixed models (GLMMs) are a common approach, and the BYM model has been extensively used in disease mapping applications.

While the vast majority of applications of the Fay-Herriot model have assumed normal iid random effects or spatial random effects with CAR, ICAR or SAR forms, a number of authors have considered different shrinkage forms. For example, Datta and Mandal (2015) suggested spike and slab mixture priors and Tang et al. (2018) consider horseshoe priors and Tang and Ghosh (2023) extend such priors to include spatial effects.

The inputs to the Fay-Herriot model are the weighted estimates and their variances. When the data are sparse in an area, the usual variance estimation methods may produce estimates that are zero, undefined, or unstable. For example, when the logit transform is used, problems arise for areas in which $\hat{p}_i^{\mathrm{w}} = 0/1$. To alleviate difficulties with the sampling variances, they may be modeled using generalized variance functions (Wolter, 2007, Chapter 7), which leverage the mean-variance relationship between θ_i and V_i, and/or incorporate covariates (Otto and Bell, 1995; Mohadjer et al., 2012; Franco and Bell, 2013; Liu et al., 2014). Uncertainty in the sampling variances may also be incorporated into the model by considering a joint model for the direct estimates and the associated sampling variance estimates (You and Chapman, 2006; Maiti et al., 2014; Sugasawa et al., 2017; Gao and Wakefield, 2023b). If there are just a small number of areas with no data then one may still fit Fay-Herriot models, treating these areas as having missing data. Spatial random effects models are particularly appealing in this regard, and the situation brightens considerably if there are strong associations with covariates. It is, however, very difficult to give guidelines on when the proportion of missing areas becomes too large to follow such a strategy.

Synthetic estimators are developed under the assumption that small areas have similar characteristics to larger areas (Rao and Molina, 2015, Section 3.2). In simple models, synthetic estimators correspond to Fay-Herriot models with no random effects, which therefore lean on the regression part of the model. In general, the bias from such approaches can be considerable when between-area residual variation is significant. However, if one requires estimates for a set of geographic areas for which few were sampled, a spatial model is not useful, and strongly predictive covariates are essential. In this paper, we focus on the situation in which the majority of the target areas provide samples.

When data are indexed by both areas and sub-areas (such as Admin-1 and Admin-2 geographies), one may introduce two levels of random effects, to provide more efficient estimation by borrowing information at each level. Torabi and Rao (2014) propose a Fay-Herriot model of this type, see also Erciulescu et al. (2019). Such a model is very appealing, but in LMICs the data will often be too sparse to produce reliable direct estimates and variances at Admin-2, and these are required as inputs.

There is a strand of research that models area-level responses on [0,1] using a variety of beta sampling models, see for example, Janicki (2020) and De Nicolò et al. (2024).

Surveys with temporal information are quite commonplace, and various models have been proposed; see for example, Rao and Molina (2015, Section 4.4.3). A more recent review of combining surveys over time, is provided

by Pfeffermann (2022). In the LMICs setting, Li et al. (2019) use a spatiotemporal Fay-Herriot model to estimate Admin-1 under-5 mortality in 35 African countries, using BYM2 spatial priors, random walk temporal priors (Rue and Held, 2005) and space-time interactions, as described in Knorr-Held (2000).

3.4 Unit-Level Models

When it is not possible to obtain reliable area-level direct estimates and standard errors, one must consider alternative approaches. Unlike area-level models, unit-level models treat individual responses as random variables drawn from some superpopulation model (Royall, 1970; Valliant et al., 2000), making the finite population small area means and totals random as well.

For continuous responses without a clustering component, Battese et al. (1988) proposed the nested error regression model, also called the basic unit-level model. Again define N_i as the population size in area i, so that $N_i = |U_i|$ with $j = 1, \ldots, N_i$, indexing units in the area i population. The unit-level nested error model is defined for all members of the population as,

$$Y_{ij} = \alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i + \varepsilon_{ij}, \tag{7}$$

for $j = 1, \ldots, N_i$, $i = 1, \ldots, m$, and where α is the intercept, $\boldsymbol{x}_{ij}$ are unit-level covariates, and $\boldsymbol{\beta}$ are regression coefficients. We have area-level random effects $u_i \mid \sigma_u^2 \sim_{\text{iid}} \text{N}(0, \sigma_u^2)$ and $\varepsilon_{ij} \mid \sigma_\varepsilon^2 \sim_{\text{iid}} \text{N}(0, \sigma_\varepsilon^2)$ represent iid unit-level effects, which are a combination of true unit-level variation, and measurement error. We use Y_{ij} to avoid ambiguity between the stochastic variable (in the model) and the realized value y_{ij} in the fixed finite population. A key requirement is that (7) holds for all members of the population, sampled and unsampled. We define S_i^{u} as the set of sampled units in area i, with $n_i = |S_i^{\text{u}}|$ so that the observed data are $\{y_{ij}, j \in S_i^{\text{u}}, i = 1, \ldots, m\}$.

For any area i, the finite population mean of the continuous response according to the model is, under the assumption that ε_{ij} represents true signal,

$$\overline{Y}_i = \alpha + \overline{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + u_i + \overline{\varepsilon}_i,$$

where

$$\overline{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}, \quad \overline{\boldsymbol{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \boldsymbol{x}_{ij}, \quad \overline{\varepsilon}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \epsilon_{ij},$$

denote the area means, for $i = 1, \ldots, m$. This illustrates that for a linear model, only area-level population covariate averages are needed. Since $\overline{\varepsilon}_i \mid \sigma_\varepsilon^2 \sim \text{N}(0, \sigma_\varepsilon^2/N_i)$, we can set $\overline{\varepsilon}_i = 0$, for large finite population sizes. Hence, inference for $\overline{Y}_i$, is close to that for,

$$\begin{aligned} \mu_i &= \text{E}\left[\, \overline{Y}_i \mid \alpha, \boldsymbol{\beta}, u_i \,\right] \\ &= \alpha + \overline{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + u_i, \qquad i = 1, \ldots, m. \end{aligned}$$

Further, $\overline{Y}_i$ is a description of our beliefs about the unknown population average $\overline{y}_i$. Hence, model-based estimation of μ_i is asymptotically equivalent to estimation of the finite population mean. The area-level summary is the product of an *aggregation step* which is more difficult when attempted with nonlinear prevalence models, as we see shortly.

There have been many applications of unit-level models Rao and Molina (2015) and extensions to the original model. For example, Molina and Rao (2010) proposed a linear nested error model to estimate nonlinear poverty measures.

As with area-level models, unit-level models can be analyzed using frequentist or Bayesian approaches. Model-based approaches are more common than design-based, since inference is based on a super-population model that is responsible for the randomness. However, design consistency for the unit-level linear model, via a pseudo-EBLUP approach that explicitly uses the weights, was considered by You and Rao (2002a), see also Jiang and Lahiri (2006); Pfeffermann and Sverchkov (2007) and Guadarrama et al. (2018).

Our interest is primarily in binary outcomes such as the prevalence of health and demographic indicators. For such a response, a logistic unit-level model may be specified. For binary variables, multilevel logistic regression models have been used previously (Malec et al., 1997; Jiang and Lahiri, 2001; Congdon and Lloyd, 2010; Zhang et al., 2014; Hobza and Morales, 2016).

A unit-level binary response model for all members of the population is,

$$Y_{ij} \mid r_{ij} \sim \text{Bernoulli}(r_{ij}) \tag{8}$$

$$\text{logit}(r_{ij}) = \alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i \tag{9}$$

for $j = 1, \ldots, N_i$, $i = 1, \ldots, m$, and with individual-level covariates $\boldsymbol{x}_{ij}$ with associated log odds ratios $\boldsymbol{\beta}$. This is an example of a GLMM and inference can again proceed from either frequentist or Bayes perspectives. The EB approach extends the EBLUP method beyond the LMEM to the GLMM (Rao and Molina, 2015, Chapter 9). Sometimes this approach is referred to as empirical best prediction (EBP).

Aggregation under nonlinear models is not straightforward. Under the population model (8) and (9), the area-level risk is

$$r_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \text{expit}(\alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + u_i), \tag{10}$$

for $i = 1, \ldots, m$. Notice that we require covariates, $\boldsymbol{x}_{ij}$, on all the individuals, which is difficult to achieve in LMICs. Specifically, censuses are often infrequently carried out and may be inaccurate, and if we wish to use satellite imagery or meteorological variables, we need the

locations of all units, which is never achievable. These challenges are further discussed in Sections 4.3 and 5.2.

The above description adopts notation that is typical of that used in the SAE literature, with units j in areas i. We now define notation and provide a development for data available from household surveys in LMICs, in which units j in the same cluster are recorded with the same location, and so we combine together the data within each cluster. Note that, for a given indicator, all units in the same cluster have the same design weight and the same non-response adjustment (these adjustments are constant within sampling strata, since strata are used as the response groups), which provides further justification for combining data within the same cluster.

Let S_i^{c} index the set of sampled clusters c in area i, with $n_i = |S_i^{\text{c}}|$. An added complication is that not all households (and therefore individuals) are sampled in the selected clusters. We let S_{ic}^{c} index the set of sampled individuals (units) in cluster c of area i, with $n_{ic} = |S_{ic}^{\text{c}}|$. In the population, the total number of clusters in area i is C_i and the total number of individuals in cluster c is N_{ic}, $c = 1, \ldots, C_i$, $i = 1, \ldots, m$. In the sampled clusters, the observed totals for the sampled units are,

$$y_{ic}^{\text{CL}} = \sum_{j \in S_{ic}^{\text{c}}} y_{ij}, \qquad c \in S_{ic}^{\text{c}}, \qquad i = 1, \ldots, m.$$

These totals should be distinguished from the unobserved population totals, for all clusters,

$$y_{ic}^{\text{POP}} = \sum_{j=1}^{N_{ic}} y_{ij}, \qquad c = 1, \ldots, C_i, \qquad i = 1, \ldots, m.$$

We denote the cluster-specific *risk parameter* by r_{ic}. This will not cause ambiguity with individual-level risks r_{ij} since the models will only involve cluster-level risks. We assume that all individuals in the cluster experience the same risk so that $\text{E}[Y_{ic}^{\text{POP}}/N_{ic} \mid r_{ic}] = r_{ic}$, and assume, for all clusters in the population, the observation model is,

$$Y_{ic}^{\text{POP}} \mid r_{ic} \sim \text{Binomial}(N_{ic}, r_{ic}), \tag{11}$$

for $c = 1, \ldots, C_i, i = 1, \ldots, m$, and where Y_{ic}^{POP} and y_{ic}^{POP} again differentiate between stochastic variables in the model and realized values in the fixed finite population. This model is assumed to also hold for selected individuals, so that the sampling model is,

$$Y_{ic}^{\text{CL}} \mid r_{ic} \sim \text{Binomial}(n_{ic}, r_{ic}), \tag{12}$$

for $c \in S_i^{\text{c}}$, $i = 1, \ldots, m$.

We model the risk parameters using a logistic model, with cluster-level covariates $\boldsymbol{z}_{ic}$ (where we assume that the covariates are common to all units in the cluster) and area-level random effects u_i. As we have described, DHS and MICS data are collected via stratified two-stage cluster sampling, with clusters selected within strata, and households sampled within clusters. To account for between-area variability and within-cluster variation, two levels of random effects can be used, for areas and for clusters (Stukel and Rao, 1997; Marhuenda et al., 2017). This gives the logistic regression model, assumed to hold for all clusters in the population, as

$$\text{logit}(r_{ic}) = \alpha + \boldsymbol{z}_{ic}^{\top}\boldsymbol{\beta} + u_i + \delta_{ic} \tag{13}$$

where $\delta_{ic} \mid \sigma_\delta^2 \sim_{\text{iid}} \text{N}(0, \sigma_\delta^2)$ denote cluster level random effects for $c = 1, \ldots, C_i$, $i = 1, \ldots, m$. The area-level random effects may be iid or assigned a spatial model, such as the BYM2. All units (households) have the same geographical (recorded) cluster location but there is within-cluster dependence of units, which is why we specify an overdispersion model for the aggregated counts in each cluster. The cluster-specific random effects, δ_{ic}, account for the dependence that two individuals within the same cluster will display (Scott and Smith, 1969), which will induce *overdispersion* in the sampling model. If we take a beta model for the cluster random effects (as opposed to the normal model for δ_{ic}, assumed above), we obtain an alternative model for overdispersion. This leads to a betabinomial marginal sampling model, and this model is discussed more fully in Section 6 and in the Supplementary Materials. Marino et al. (2019) consider a logistic model with the distribution of the area-level random effects being unspecified, and illustrate its use with data on unemployment incidence.

Now now consider aggregation for the SAE unit-level model (13) and suppose δ_{ic} represents within-area variation. Then, under the superpopulation model, $r_i = \text{E}[P_i]$ and from (17), we obtain,

$$\begin{aligned} r_i &= \sum_{c=1}^{C_i} r_{ic} \times \frac{N_{ic}}{N_i} \\ &= \sum_{c=1}^{C_i} \text{expit}(\alpha + \boldsymbol{z}_{ic}^{\top}\boldsymbol{\beta} + u_i + \delta_{ic}) \times \frac{N_{ic}}{N_i}. \end{aligned}$$

In practice, we may sometimes know the number of clusters in the sampling frame, but not their locations or sizes. Dong and Wakefield (2021) estimate subnational vaccination prevalence in Nigeria using the 2018 DHS. The sampling frame is approximated by simulating locations $\boldsymbol{s}_{ic}$ from a population density surface, and then taking information available in the DHS survey manual to simulate cluster-level (child) population sizes, N_{ic}. Clearly this approach contains a number of approximations, but is more consistent with an SAE, rather than an MBG, approach, since it attempts to recover the final population. Samples are obtained from the posterior via

$$r_i^{(b)} = \sum_{c=1}^{C_i} \text{expit}(\alpha^{(b)} + \boldsymbol{z}(\boldsymbol{s}_{ic})^{\top}\boldsymbol{\beta}^{(b)} + u_i^{(b)} + \delta_{ic}^{(b)}) \times \frac{N_{ic}}{N_i}$$

with $\delta_{ic}^{(b)} \mid \sigma_\delta^{2(b)} \sim \mathrm{N}(0, \sigma_\delta^{2(b)})$ and with samples from the posterior, $\{\alpha^{(b)}, \boldsymbol{\beta}^{(b)}, \sigma_\delta^{2(b)}\}$ for $b = 1, \ldots, B$. This provides a method of accounting for within-area variation in risk, when a discrete spatial model is used. Here we have assumed that the δ terms should be included in the aggregation, a point we now discuss in more detail.

Model (13) describes between-cluster variation in the target population through covariates and captures unexplained between-area variation through the area-specific random effects. Models such as this were considered by Wakefield et al. (2020). The δ_{ic} terms may also represent "real" between-cluster signal, since the spatial random effect term u_i is constant within areas. Further discussion of δ_{ic} is postponed to Section 4.2.

The above cluster-level model may be described as a two-fold nested error regression unit-level model (Rao and Molina, 2015, Section 4.5.2) for binary data, with a binomial sampling model.

When estimating poverty, income or inequality measures, linear models like the basic unit-level model have been used and there is great interest in such approaches. Elbers et al. (2003) use a nested error regression model with random effects at the cluster level instead of at the area-level. Their method, commonly called the ELL approach (after the authors of the paper, Elbers, Lanjouw, and Lanjouw), has been widely used to estimate small area means of poverty-related quantities, especially by the World Bank. However, as pointed out by Das and Haslett (2019), the ELL model assumes that all between-area differences are explained by the covariates, which may not be reasonable when covariates are unavailable or weakly related to the outcome, see Diallo and Rao (2018) for an extended ELL model, see also Molina and Rao (2010). Molina et al. (2019) review further improvements and discuss other aspects of poverty mapping.

There are two major challenges to the use of unit-level prevalence models: 1) appropriately accounting for the design and 2) aggregating from clusters to areas. In the context of the household surveys that are common in LMICs, both aspects are even trickier, because of the lack of available information on the sampling frame. We next discuss 1), with 2) being delayed to Section 4.3.

3.5 Accounting for the Complex Design

Unlike Fay-Herriot area-level models, which incorporate sampling weights when computing direct estimates and their variances (and hence account for the design), unit-level models do not generally include sampling weights in the estimation procedure. Hence, it is crucial to include design information in the unit-level model specification.

One might naively include the weights in the regression model as a covariate, but aggregation would require the weights for the whole frame, and one is then still faced with the problem of getting an appropriate area-level variance estimate. There are more complex procedures that include incorporating functions of the selection probabilities in the model (Verret et al., 2014), or model the sampling weights to account for informative sampling, see Pfefferman (2007) for a frequentist approach and Si et al. (2015) for a Bayesian slant. Parker et al. (2023) review various approaches to account for informative sampling in the unit-level model.

It is (often implicitly) assumed that the sampling design is ignorable with respect to the model used for estimation. We need both (i) the specified risk model to be correct for all sampled and non-sampled individuals in both observed and unobserved clusters, and (ii) outcomes for any set of individuals from the target population to be independent, conditional on their risk parameters. Requirement (i) is often justified by including relevant design variables as covariates in the model, for example, including strata variables as covariates (Little, 2012). However, when some of the relevant design variables are unavailable or their functional relationship to the response is unknown, then this approach may not be successful.

For household surveys in LMICs the stratification is usually Admin-1 area crossed with urban/rural (with each cluster being classified as urban or rural). To account for this, one could include fixed effects for Admin-1 areas, urban/rural status, and their interaction. A more pragmatic and parsimonious approach is to include random effects (spatial or otherwise) at Admin-1 or Admin-2, and the main effect of urban/rural only (thereby assuming that the association with urban/rural is approximately constant across Admin-1 areas).

For many indicators, non-negligible associations between prevalence and urban/rural will be present. Including an urban/rural indicator in the model is straightforward, but to then aggregate, the urban and rural population sizes must be known for each area of interest. Typically these sizes may be available at the Admin-1 level (in the survey reports) but are usually unavailable at the Admin-2 level. Note also that the urbanicity status of a cluster may change over time due to urbanization, but when unit-level models procedures adjust for urbanicity, it is not the current urban/rural classification of the cluster that is relevant, but rather, the classification when the sampling frame was constructed. The recorded stratification variable should thus be viewed as a fixed partition over time. A method for constructing urban/rural fractions has recently been developed, by producing an urban/rural surface over the study region, based on a classification algorithm (Wu and Wakefield, 2024). In Section 6 we describe an approach to estimating urban/rural fractions at Admin-2, in the context of subnational HIV prevalence estimation.

If forming an urban/rural surface is deemed too difficult, one may include readily available variables such as

population density and/or nighttime lights in the regression model, as surrogates for urban/rural. The success of this strategy is difficult to anticipate. As a starting point, it is prudent to first estimate the association between prevalence and urban/rural. If the association is negligible, one might sleep a little better. One can also examine a sequence of cluster-level models that include urban/rural alone and also with population density.

Beyond stratification, the DHS typically implements PPS sampling of clusters, with size corresponding to the number of households in the cluster. However, the size of sampled clusters is not usually made public for privacy reasons and the size of non-sampled clusters is typically not known, meaning that cluster size may not be used as a covariate. As a result, if cluster size is associated with the outcome of interest, then the design is not ignorable. Flexible models to account for PPS sampling have been proposed (Zheng and Little, 2005).

Since accounting for stratification and clustering via model specification is delicate, other methods have been proposed for acknowledging the sampling design when using unit-level models. Many of these methods leverage the sampling weights during parameter estimation, including pseudo-EBLUP estimation (You and Rao, 2002b). Pseudo-Bayesian approaches that produce weight-adjusted posteriors are also under development, but are tricky to extend to random effects models (León-Novelo and Savitsky, 2019; Williams and Savitsky, 2021; Gao and Wakefield, 2023a).

Binary data unit-level models in the SAE literature are often used in a high income countries context in which there are high quality census data available. In this case, the logistic regression model may include demographic groups (e.g., by age, sex, race) as covariates. In this case, aggregation consists of averaging over the group-specific risks in each area, weighting by the population in each cross-classification cell (with these subpopulation sizes being available from the census). Both Malec et al. (1997) and Congdon and Lloyd (2010) follow this approach, which has strong connections with the multilevel regression and poststratification approach described in Section 5.5. Aggregation for cluster-level models will be considered in Section 4.3.

In Section 4 we describe further MBG cluster-level models which are particular unit-level models that have seen widespread use for analyzing household survey data in LMICs.

3.6 Computation

Fay-Herriot models are simply LMEMs and computation is straightforward. In general, maximum likelihood (ML), restricted maximum likelihood (REML) and method of moment estimators fall under the frequentist umbrella for estimating variance components in the model (Rao and Molina, 2015). Linear unit-level models are also similarly relatively easy to implement, since thay are also LMEMs. Non-linear (e.g., logistic) models, are examples of GLMMs and frequentist inference is more involved, particularly with spatial random effects, since unlike their linear counterparts, the integration over random effects cannot be carried out analytically. However, likelihood methods are still available.

Bayesian inference can be carried out for LMEMs and GLMMs with iid or spatial random effects using MCMC or INLA. There are many MCMC algorithms that could be used but spatial models, and the dependent posteriors they produce, can be problematic in the sense of displaying slow mixing (Rue and Held, 2005). While MCMC algorithms tailored for specific models can overcome some issues (Rue and Held, 2005, Chapter 4), INLA offers an accessible alternative. INLA is extremely fast and LMEMs and GLMMs with iid or spatial random effects are exactly the kinds of models that INLA was designed for. The accuracy of the INLA method has been demonstrated on many occasions, see for example Osgood-Zimmerman and Wakefield (2023).

Specific packages to implement area-level and unit-level models are described in Section 5.7.

4. MODEL-BASED GEOSTATISTICS

Like unit-level models from the SAE literature, MBG approaches treat individual outcomes as response data. Almost all unit-level SAE approaches assume iid or discrete spatial random effects, while in contrast MBG approaches assume continuous spatial surfaces that represent the risk of experiencing an outcome at any given location. These risk surfaces may be used to compute spatially averaged risks for areas of interest, and can be viewed as proxies for finite population prevalences.

4.1 Geostatistical Models

For prevalence mapping with MBG, the individual response data are typically aggregated to the cluster level, with each cluster representing a distinct spatial location.

The following model for mapping health outcomes at the Admin-2 level is typical (see for example Mayala et al. (2019)), and uses a binomial response model with a logit link and latent spatial random effects. The sampling model is again given by (12) with,

$$\text{logit}(r_{ic}) = \alpha + \boldsymbol{z}(\boldsymbol{s}_{ic})^\top \boldsymbol{\beta} + \omega(\boldsymbol{s}_{ic}) + \delta_{ic} \tag{14}$$

where $\boldsymbol{s}_{ic}$ denotes the cluster location, the logit baseline odds is α, $\boldsymbol{z}(\boldsymbol{s}_{ic})$ is a vector of cluster-specific covariates and the associated log odds ratios are $\boldsymbol{\beta}$. The spatially correlated random effects are denoted $\omega_{ic} = \omega(\boldsymbol{s}_{ic})$ and $\delta_{ic} \mid \sigma^2_\delta \sim_{\text{iid}} \text{N}(0, \sigma^2_\delta)$ are independent cluster random effects. The area-specific random effects used in the previously described SAE unit-level model (13) have been replaced by ω_{ic}, which in most MBG approaches is assumed

to arise from a GRF model, often with a Matérn correlation structure (Stein, 1999). The discussion on acknowledging the design in Section 3.5 carries over to MBG models, so that when samplng is informative, one is required to include design variables in the covariate model.

This model describes risk as a function of location (where the GRF is continuous, but covariates can be discontinuous), with δ_{ic} modeling "random deviations" from the risk surface at an observed cluster. In summary, the MBG form (14) aims to describe as much of the residual spatial variation as possible, while the SAE version (13) emphasizes the prevalences in the discrete *target areas* and the area-specific random effect does not try to explain residual spatial variation at a finer scale than the target areas. The former can lead to improved unit-level predictions compared to the latter, but for area-level summaries, it is not clear which model will be beneficial.

4.2 Interpretation of the Cluster-Specific Error Term

In Equation (14), as discussed in detail by Dong and Wakefield (2021), δ_{ic} may represent one of three distinct types of variation. It may be a surrogate for "measurement error", that is, the misrecording of responses. Given the binary nature of the responses, this would not be an ideal model, which would instead contain misclassification probabilities. If this is the interpretation given, then δ contributions would not be included in predictions. Some authors have not included the cluster-specific effect in prediction (Wakefield et al., 2019; Diggle and Giorgi, 2019).

Another, very plausible interpretation is that the terms are a mechanism to induce *overdispersion*, i.e., excess-binomial variation, due to within-cluster dependence in responses. In this case, the predicted risk at location $\boldsymbol{s}_{ic}$ is the marginal risk, averaged over δ_{ic}, so that for a generic cluster,

$$\begin{aligned}&\mathrm{E}[\, r_{ic} \mid \alpha, \boldsymbol{\beta}, \omega(\boldsymbol{s}_{ic}) \,] \\ &= \mathrm{E}[\, \mathrm{expit}(\alpha + \boldsymbol{z}(\boldsymbol{s}_{ic})^\top \boldsymbol{\beta} + \omega(\boldsymbol{s}_{ic}) + \delta_{ic}) \,], \end{aligned} \tag{15}$$

where the expectation is over $\delta_{ic} \mid \sigma_\delta^2 \sim \mathrm{N}(0, \sigma_\delta^2)$.

The third explanation is that δ_{ic} represents "true" variation in the risk surface, i.e., unstructured between-cluster variation. In this case, when one aggregates across an area, δ contributions must be included. Hence, for an unobserved cluster, operationally the risk is (15), regardless of whether we assume that δ is included to accommodate overdispersion or true signal, because in both cases we need to average over $\delta \mid \sigma_\delta^2$.

4.3 Estimating Prevalence across Administrative Areas: Aggregation

In the MBG spatial statistics formulation with a GRF model, an aggregated area-level risk target is conceived as,

$$r_i = \int_{A_i} r(\boldsymbol{s}) d_i(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s}, \tag{16}$$

where A_i is the geographical extent of area i and $d_i(\boldsymbol{s})$ is a normalized density surface which represents the locations of the relevant population. In contrast, the finite population prevalence target in the SAE world is,

$$P_i = \frac{Y_i^{\mathrm{POP}}}{N_i} = \sum_{c=1}^{C_i} \frac{Y_{ic}^{\mathrm{POP}}}{N_{ic}} \times \frac{N_{ic}}{N_i}, \tag{17}$$

where (recall) C_i is the total number of clusters in area i, and N_{ic} and Y_{ic}^{POP} are the population denominators and numerators, with

$$N_i = \sum_{c=1}^{C_i} N_{ic}, \qquad Y_i^{\mathrm{POP}} = \sum_{c=1}^{C_i} Y_{ic}^{\mathrm{POP}},$$

being the totals in area i, $i = 1, \ldots, m$.

The links between (17) and (16) are clear with the prevalence $Y_{ic}^{\mathrm{POP}}/N_{ic}$ being replaced by the risk $r(\boldsymbol{s})$ and the population fraction N_{ic}/N_i being replaced by the density $d_i(\boldsymbol{s})$. The numerator and denominator of the weighted estimator $\hat{p}_i^{\mathrm{w}}$, defined in (2), directly estimate the numerator (Y_{ic}^{POP}) and denominator (N_i) in (17). We also contrast with the unit-level model that lead to the averaged risk (10) in which a superpopulation model provides estimates for the set of N_i finite individuals in area i. In a perfect world, master sampling frame information would be available for each cluster, including the relevant population sizes and locations, along with all covariates used in the model. Unfortunately such information is never available.

Under the MBG approach, risk may be modeled via (14), yielding a surface representing risk at any location $\boldsymbol{s}$ in the study region, regardless of whether a cluster exists at the location. To approximate (16) in area i, a grid consisting of G_i cells with centroids $\boldsymbol{s}_{ig}^{\mathrm{G}}$, $g = 1, \ldots, G_i$, is created, for example, at resolution $1\,\mathrm{km} \times 1\,\mathrm{km}$ or $5\,\mathrm{km} \times 5\,\mathrm{km}$ (Utazi et al., 2020; Local Burden of Disease Vaccine Coverage Collaborators and others, 2021). The risk for area i is then estimated by the population-weighted estimated risk across the grid,

$$r_i = A \sum_{g=1}^{G_i} r(\boldsymbol{s}_{ig}^{\mathrm{G}}) d_i(\boldsymbol{s}_{ig}^{\mathrm{G}}) \tag{18}$$

where the factor A is defined as the area of the cells and $d_i(\boldsymbol{s}_{ig}^{\mathrm{G}})$ is the normalized population density at grid location $\boldsymbol{s}_{ig}^{\mathrm{G}}$. Current MBG approaches implicitly assume that the population in the administrative area is so large that the difference between prevalence and risk is minor. With reference to (17), this implies changing the target of inference to $r_i = \mathrm{E}[P_i]$.

Following the discussion in Section 4.2, under either overdispersion or true signal interpretations, we should include δ_{ic} contributions to the area-level averages, i.e., with δ^{G}_{ig} terms in the predictions at each grid cell (Burstein et al., 2019; Utazi et al., 2020, 2021). For unobserved clusters, we use (15) and it is this expectation which is being implicitly used when aggregation is carried out. However, since grid cells are not matched to clusters, there is an arbitrary element here, since the results will change when different gird sizes are used. This argues for taking $G_i = C_i$, though the latter may not be known and it still leaves the problem of choosing locations, $\boldsymbol{s}^{\mathrm{G}}_{ig}$, $g = 1, \ldots, G_i$.

It is often not explicitly stated whether the cluster random effects δ are included in the aggregation. Paige et al. (2022) discuss population-weighted spatial aggregation in detail. Issues related to interpreting the cluster-specific non-spatial variation have received little discussion in recent efforts to use MBG for prevalence mapping of demographic and health indicators (Giorgi and Diggle, 2017; Burstein et al., 2019).

Chan-Golston et al. (2019) take a somewhat different geostatistical view, explicitly acknowledging the complex design. They develop theory for a continuous response, with available covariates assumed to acknowledge any informative sampling. Follow up work includes Finley et al. (2024) and Chan-Golston et al. (2022) who take a preferential sampling viewpoint with a selection model that depends on location through covariates. Complex survey designs lead (implicitly) to spatially varying sampling efforts. This resembles geostatistical preferential sampling (Diggle et al., 2010), but in survey sampling, the inclusion probabilities are not directly linked to the level of risk in the clusters. The inclusion probabilities are instead determined by a survey design that aims to reach the desired statistical power for areas of interest in a cost efficient way. This does not mean that the survey design should be haphazardly ignored, however, as we have emphasized.

In many applications that we have looked at using MBG approaches (and the SPDE approximation) we have found that the reported prediction intervals are narrower than the corresponding area-level intervals. The MBG approaches make more assumptions and so, if those assumptions are met, can result in more precise estimation. However, the reported uncertainty could be artificially low because of missing sources of error, errors in the aggregation process, or other model failures.

4.4 Computation

MBG analyses usually adopt a Bayesian approach to inference, in contrast to the frequentist approaches which are dominant in the SAE literature. Samples may be obtained from the posterior distribution of $r(\boldsymbol{s})$ for any location $\boldsymbol{s}$, or for the spatially averaged area-level risks r_i, via (18). The posterior uncertainty for any variable may be quantified using credible intervals or by reporting the posterior variance or posterior coefficient of variation.

When using a continuous spatial model, the number of observed clusters $n = n_1 + \cdots + n_m$, is large, and computation is an issue because we need to manipulate $n \times n$ matrices, which involves $O(n^3)$ operations (Rue and Held, 2005). Many approximations have been proposed to overcome this problem, see Heaton et al. (2018). The stochastic partial differential equation (SPDE) approach pioneered by Lindgren et al. (2011) is particularly popular for MBG approaches in LMICs. SPDE models, implemented using INLA, have been used in LMICs to produce high resolution maps of health and demographic outcomes, including vaccination coverage (Utazi et al., 2020, 2021), neonatal, infant, and child mortality (Golding et al., 2017; Burstein et al., 2019), childhood malnutrition (Kinyoki et al., 2020), contraceptive usage rates (Bosco et al., 2017), and malaria prevalence rates (Bhatt et al., 2015). The SPDE method is still seeing active development (Lindgren et al., 2022). The INLA software implementation has restrictions on the models that can be fitted, and so Template Model Builder (TMB) (Kristensen, 2014) has been used an alternative for GRF models (Burstein et al., 2019), when more flexibility in modeling is required. TMB and INLA are compared in Osgood-Zimmerman and Wakefield (2023).

5. ADDITIONAL ISSUES

In this section, we discuss a number of other issues that are pertinent to prevalence mapping.

5.1 Geographical Issues

Although the clusters in DHS and MICS surveys are typically census enumeration areas, GPS coordinates are supplied for the center of each cluster only, and we treat these as points in our analyses, following convention. For confidentiality, urban/rural cluster locations are displaced by up to $2\,\mathrm{km}/5\,\mathrm{km}$ with the locations of a random sample of a further 1% of rural clusters being jittered by up to $10\,\mathrm{km}$. In many cases, the jittering algorithm is constrained to keep the location in the same sampling stratum (i.e., Admin-1 crossed with urban/rural). As a result, some clusters may be randomly relocated to an Admin-2 area that does not correspond to the true area. The displacement has implications for modeling and has been investigated methodologically, as we detail shortly, and from an applied perspective, see for example Tonye et al. (2024).

For continuous spatial models, the jittering distances are short compared to the typical effective range parameter for GRFs in geostatistical models and the displacement has been demonstrated to have only a minor impact on parameter estimation and predictive accuracy (Wilson and Wakefield, 2021; Altay et al., 2022). However, rasters

can show large variation in values within typical jittering distances and failure to account for this local variability reduces predictive power (Altay et al., 2024). With respect to covariate modeling, Perez-Heydrich et al. (2013) and Warren et al. (2016) have suggested using a buffer zone of $5\,\mathrm{km}$ within which to average the covariates. However, these methods still exhibited reduced predictive power compared to methods fully accounting for covariate uncertainty and complicate the interpretation of covariate models (Altay et al., 2024).

In older surveys (especially MICS surveys), the geographical coordinates of the clusters may not be available, with only the areas within which the cluster lies being accessible. Marquez and Wakefield (2021) develop a model for data in which polygon information only is available, by averaging over potential locations, in contrast to previous approaches that proposed ad hoc algorithms (Golding et al., 2017; Reiner et al., 2018; Utazi et al., 2018).

In this paper, we argue that in many situations we may include discrete spatial random effects at a level that corresponds to the level at which inference is required, particularly at coarser geographical levels. However, one situation in which continuous spatial models are more appealing is when multiple datasets are available over areas with different boundaries. Continuous spatial models provide a natural way to model risk at each spatial location, particularly when the boundaries shift over time, or the different geographies are not nested.

5.2 Auxiliary Variables

To produce reliable subnational estimates, at fine geographical resolution in particular, requires auxiliary variables with good predictive power. One simple viewpoint is that SAE is a predictive exercise and so one can be flexible in the strategy used for constructing a covariate model. But auxiliary information may be available from a variety of sources, and some are more preferable than others. The use of census variables is common in high income countries, but in LMICs censuses are less frequently carried out and the data they produce are less reliable. Administrative source data such as from health facilities offer great potential but are currently not routinely available. Satellite data, such as night time lights and vegetation indices, are becoming widely available, and are being extensively used, see Section 5.6. Meteorological and climate variables may be useful, though are perhaps less obviously related to demographic and health variables, and in some cases will show little variation over the study region. Economic, household, cultural and lifestyle variables (for example, dietary variables) may show strong associations in regression settings. However, to obtain these variables at finer geographical resolutions requires modeling, based on survey data (Gething et al., 2015), which opens up a whole can of worms. The utilization of mobile phone and social media data are exciting prospects but their use is currently in its infancy.

One should also be careful in selecting variables to include in the model. In an extreme case, suppose we are modeling child mortality and there is an available covariate that is a composite measure of poverty. We would clearly not want to use such a variable if it were constructed using a measure of child mortality.

When modeling at a fine pixel-level scale there are a number of issues that require consideration. The displacement of cluster locations is relevant to the use of pixel-level covariates, as discussed in Section 5.1. For a fine-scale pixel surface, the associated covariates on the pixel grid may take a large range of values, but the cluster-level data from which the covariate model is estimated may not adequately cover this range, particularly when there are a large number of covariates. This may result in aberrant predictions or pixels with sets of covariates that are not adequately represented in the observed clusters.

Model choice is in its infancy in SAE, but is important, particularly if there are a large number of covariates. Machine learning methods can deal with large numbers of variables, but have their own challenges, as discussed in Section 5.6.

5.3 Uncertainty Measures

The presence of random effects in SAE models complicates the interpretation of uncertainty intervals. Under the frequentist paradigm, bias and variance are calculated with respect to repeated sampling of both the data and the random effects. Consider a linear model with area-level random effects, $u_1, \ldots, u_m$. Then the BLUP $\hat{u}_i$, for a generic area i, is unbiased in the sense that $\mathrm{E}[\hat{u}_i - u_i] = 0$, where the expectation is with respect to the distributions of both the distributions of y_{ij}, and u_i. However, in a real population, $\hat{u}_i$ is biased in the sense that $\mathrm{E}[\hat{u}_i - u_i \mid u_i] \neq 0$, $i = 1, \ldots, m$, due to shrinkage. Hence, the "Unbiased" in BLUP is deceptive and is not relevant for making inference about *any particular area* because the area-level predictor $\hat{u}_i$ is clearly biased.

When frequentist analyses are carried out, the mean squared error (MSE) is often used to quantify the uncertainty for each area, and is defined as $\mathrm{E}[(\hat{u}_i - u_i)^2] = \mathrm{var}(\hat{u}_i - u_i)$ where, again, the expectation is taken over both data and random effects distributions. As with the first moment discussion above, $\mathrm{var}(\hat{u}_i - u_i)$ differs from $\mathrm{var}(\hat{u}_i - u_i \mid u_i)$. Conditional MSE has been considered, Rao and Molina (2015, Section 6.2.7).

The posterior variance $\mathrm{var}(u_i \mid \boldsymbol{y})$ is a Bayesian measure of uncertainty which represents the variance of the random variable u_i given the data and chosen priors, and conditional on the model being correct. Credible intervals can be constructed from the posterior for any parameter of interest when samples from the posterior are available.

For prevalences close to zero and with small samples in particular, the posterior will be skewed and the posterior intervals will reflect this, which is desirable. Such intervals are also exact, up to the numerical approximation that is used for computation. If the posterior approximations can be improved, these intervals can be made arbitrarily close to exact. In contrast, frequentist intervals are based on asymptotic normality, which is somewhat ironic given the "small" in SAE.

Direct estimates have asymptotically correct coverage for every area since they use data from the relevant area only, and do not employ shrinkage. Depending on one's inferential persuasions, one may be comforted when Bayes intervals coincide with frequentist intervals in large samples. However, for random effects models, the interpretation of both frequentist confidence intervals and Bayesian credible intervals is fraught with difficulties. The discussion that follows is true for more complex random effects models (including unit-level models), but for simplicity, consider a simple Fay-Herriot model with $\mathrm{E}[\hat{\theta}_i \mid u_i] = \alpha + u_i$ and to keep the discussion simple, assume α is known. The 95% prediction interval is, $(\alpha + \hat{u}_i) \pm 1.96 \times \sqrt{\mathrm{MSE}(\hat{u}_i)}$. For this area, and assuming a fixed u_i, if we repeatedly sample data and produce intervals, those intervals will not contain the true u_i 95% of the time. What we can say is that the coverage will be approximately 0.95 if we average over *all* possible areas, that is, we need to also average over the random effects distribution. Due to the shrinkage, areas which have extreme, u_i, in the tails (away from zero) will have coverage below the nominal 0.95 while areas in the center of the distribution, will have coverage above 0.95. This realization is deeply disturbing, since it will often be areas with relatively low or high prevalence that will have the lowest coverage.

The silence on these issues in both the SAE and MBG literatures has been deafening. A notable exception is Burris and Hoff (2020), who discuss this issue in detail, and introduce a new method for constructing intervals that borrow strength across areas but retain the correct frequentist coverage for each area. The average width of the intervals is narrower than the width of direct estimates, but not as narrow as from the random effects model.

5.4 Model Checking

We briefly discuss model checking for each of weighted, area-level, unit-level, and MBG approaches. One of the beauties of the weighted (direct) estimator is the minimal assumptions under which it is based. However, achieving nominal coverage requires a sufficiently large sample size to ensure the accuracy of the normal sampling model for the estimator. Instability of the variance estimator can reduce the accuracy. As mentioned in Section 1, agencies produce guidelines for reporting weighted estimates (often in terms of coefficients of variation), which is presumably at least in part implicitly based on considerations of accuracy of the normality assumption and variance stability.

Beyond the appropriateness of the sampling distribution, Fay-Herriot models assume a particular random effects model on a particular linking scale. In our experience, the particular form of discrete spatial model (e.g., BYM2, CAR, SAR), will not be critically important in many examples, as all perform some form of local smoothing. Plotting the smooth small areas estimates versus the direct estimates (when available) allows the shrinkage to be examined. Deciding on when there is "too much shrinkage" is a critical question, with currently no clear guidelines.

SAE unit-level models are far more dependent on model assumptions. Unfortunately, the sparsity of data within each cluster (since there are usually a small number of units within each) makes critiquing the sampling model difficult.

MBG approaches are the most dependent on assumptions, with all difficulties associated with unit-level models being present, along with a more delicate spatial model. Non-parametric exploration of the correlation function through, for example, semi-variograms, is challenging since we have sparse observations and count data. In general, in the context of the survey sample sizes in LMICS, spatial parameters (such as the total variance and proportion spatial parameter in the BYM2 model, and the spatial variance and range parameter in the GRF model) can be poorly estimated, because of the sparsity of the survey data and paucity of information in spatially dependent data (Zhang, 2004). Comparing the posteriors to the priors, and examining sensitivity of predictions (point and interval estimates) to spatial hyperpriors is recommended.

Cross-validation provides an obvious approach for model assessment, though there is no consensus on how it should be performed in a spatial setting with complex survey data. The former aspect is discussed in Roberts et al. (2017), particularly in the context of blocking strategies where care should be taken with the manner in which the data are split. We routinely carry out leave-one Admin-1 area out cross-validation, and then predict the direct estimate, using the remainder of the areas. This can be useful, but when there are a small number of Admin-1 areas (such as in our Zambia example) model assessment is difficult. This procedure is also not directly answering the question of interest, "Is the model adequate at Admin-2?". Clearly, it will be more difficult to assess model performance when the data are sparse. There are many choices to make when cross-validation is carried out, beyond the splitting strategy, including which metrics to use to compare left out data with predictive distributions. The procedure for leaving out data should also be consistent with the survey design. In Section 6 we examine a number of different options.

5.5 Multilevel Regression and Poststratification

Multilevel Regression and Poststratification (MRP) is a technique that was introduced by Gelman (1997). Under one version, a multilevel regression model is used to model a binary outcome at the unit level, as a function of a potentially large set of categorical variables. The population is partitioned into cells representing each possible combination of categorical variables and units within the same cell are aggregated. In the poststratification step, the population prevalence in each area is computed as a weighted average of the cell estimates with weights corresponding to the population fractions in the area. In this weighting step, there are close links with Generalized Regression (GREG) estimators; see Rao and Molina (2015, Section 2.3.2).

A key component of the approach is the random effects prior on the parameters corresponding to each categorical variable. The existence of a large number of cells means that the use of hierarchical smoothing is essential to overcome data sparsity. Many of the variables, including space, time, age and education levels, are candidates for smoothing priors. To ensure that no bias results from informative sampling, the relevant design variables must be included in the set of variables that are included in the model. MRP is particularly popular among political scientists, see Park et al. (2004) for an early example and Wang et al. (2015) for a successful use of the method when the original data (which were based on a sample of Xbox users) were not representative. Valliant (2020) reports a simulation study in which MRP does not perform particularly well in a situation in which a limited number of covariates were available. MRP can be viewed as a unit-level SAE model in which individual responses within the same cross-classification cell of (say) demographic subgroups are combined and then modeled across areas, with averaging over these subgroups corresponding to the aggregation step.

5.6 Spatial Machine Learning

Given the abundance of geospatial covariates now available, in the context of prevalence mapping in LMICs (as discussed in Section 5.2), it is natural that machine learning (ML) methods are growing in popularity. ML approaches are appealing in principle, as they can provide flexible modeling (to capture nonlinearities and interactions, for example) when there are a large number of auxiliary variables, while potentially avoiding a difficult model selection process. ML approaches have been used both with SAE area-level and unit-level models, and with MBG models.

While it is relatively straightforward to use generic ML algorithms (once the covariate data have been massaged into a usable form), there are a number of issues that require careful thought, including accounting for the survey design, choosing tuning parameters in the ML algorithm, and, most importantly, obtaining valid interval estimates. There are serious complications with the latter since great care is required to find a justifiable approach with a vast literature available on resampling methods such as the bootstrap, jackknife, infinitesimal jackknife and split conformal prediction methods. For many ML algorithms it is not clear which, if any, of these methods will give a valid prediction interval in any given context. Dezeure et al. (2015) provide a discussion of various aspects, including describing ML algorithms for which the vanilla bootstrap does not work, because the limiting distribution of the predictor is a complicated object which may not be continuous. The validity of a particular approach in a mapping context may also depend on an assumption of iid outcomes which will not hold for data from surveys, in which there are dependencies due to both the sampling design and the spatial nature of the data. In addition, a recurring theme is that the binary nature of the data is rarely acknowledged when ML techniques are applied.

Jean et al. (2016) were early proponents of the use of convolutional neural network (CNN) approaches for poverty mapping, but notably there was little discussion of the validity of uncertainty estimates. Burke et al. (2021) provide a review of ML techniques in the context of the SDGs, but again avoid the thorny question of uncertainty quantification.

We highlight a number of applications of different classes of ML algorithms, beginning with random forests (RFs). Georganos et al. (2021) describe a "geographical random forest" in which a local RF is constructed and illustrated by modeling population density in Dakar, using satellite data. Krennmair and Schmid (2022) examine the use of a RF covariate model embedded within a linear mixed effects model and describe a bootstrap procedure to produce measures of uncertainty (while noting that more theory is required on the procedure). Bilton et al. (2017) apply classification trees to poverty mapping at a low geographic level using survey data from Nepal. The authors use weights both in the classification algorithm and in calculating measures of uncertainty using a bootstrap.

Stevens et al. (2015) use RFs to produce high-resolution gridded maps of population, which is an extremely important endeavor. Ratledge et al. (2022) apply a CNN with satellite data to attempt to assess the causal impact of electrification. Newhouse (2023), in a wide-ranging review critiques and compare various approaches (including CNNs and tree-based methods) for wealth mapping using geostatistical data.

Bosco et al. (2017) compare artificial neural networks (NNs) with a MBG for estimating a number of development indicators in Bangladesh, Kenya, Nigeria and Tanzania, using a range of geospatial coordinates. They report mixed success in prediction across indicators and countries.

A quite different use of NNs is described by Semenova et al. (2022) who use variational autoencoders to create a class of Gaussian process priors, which can then be used to model spatial random effects such as the ω process used in (14). Wikle and Zammit-Mangion (2023) provide a review of deep NNs for spatial and spatiotemporal data, describing many approaches to constructing flexible spatial models. The review highlights difficulties with the use of deep learning: large datasets and high computational burden; the black box nature of the fitting that makes interpretation hazardous; and the difficulties in obtaining measures of uncertainty for the predictions. The second and third of these flaws is common to many spatial ML approaches.

Chi et al. (2022) link DHS data on household wealth variables to high-dimensional data from a variety of sources including satellite imagery, mobile phone networks, topographic maps and Facebook. A data reduction exercise is carried out and then a gradient-boosted regression tree is used for prediction on a $2.4\,\mathrm{km} \times 2.4\,\mathrm{km}$ grid for 135 LMICs. They acknowledge that providing a measure of the uncertainty of the map is important, but their approach to computing this uncertainty is ad hoc. Yeh et al. (2020) model a wealth index using satellite data and DHS data, with a CNN. As is typical the complex survey design is ignored and uncertainty estimates are not reported. However, an interesting aspect of this paper, is attempting to look at changes in wealth over time. Corral et al. (2025) examine the performance of ML approaches to poverty mapping. They compare validation techniques that compare predictions with direct estimates, rather than census-based measures, showing that the former can be misleading while the latter can provide a more realistic measure of the performance of an algorithm. Newhouse et al. (2025) examine poverty mapping using poverty data from a survey and compare CNNs using satellite covariate data with traditional census-based small area estimates. The model they develop uses the (scaled) design weights and is a unit-level model with two levels of random effects.

Bhatt et al. (2017) use an ensemble method known as stacked generalization to provide a more flexible mean function within an MBG model. Unfortunately, there is no theory to give valid inference with such an approach but it has still been used by IHME on numerous occasions, see for example, Golding et al. (2017); Osgood-Zimmerman et al. (2018); Browne et al. (2021). Lloyd et al. (2020) outline a stacked generalization approach to predicting the residential status of buildings in urban areas.

As pointed out above, a large impediment to the routine use of ML techniques in a spatial context is constructing valid interval estimates on predictions. A modified split conformal procedure is used by Michal et al. (2024) to obtain intervals for lasso and RF models. Wieczorek (2024) describes the general use of conformal prediction in a survey sampling setting. Mao et al. (2024), in the context of continuous response data, develop methods for applying conformal prediction algorithms to spatial data by exploiting approximate local exchangeability. McConville et al. (2017) describe a model-assisted regression model with a lasso, and derive the variance of the estimator (to give an uncertainty estimate), and illustrate its use in the context of tree canopy cover estimation. In general, asymptotic arguments are used to derive variance estimates, which again should be taken in the context of sparse data situations. Conformal prediction approaches (at least in their current guise) also have sample size limitations, which may lead to interval estimates that are not practically useful. The same drawback is also true for prediction-powered inference (PPI) approaches to forming intervals (Angelopoulos et al., 2023), which have strong links with a number of literature strands including model-assisted estimation (Särndal et al., 1992) and semiparametric inference (Robins and Rotnitzky, 1995).

Saha et al. (2023) and Zhan and Datta (2024) consider GRF models with RF and NNs regression models, respectively. Rigorous frequentist analysis of point predictions is carried out, but predictive intervals are more difficult to calculate, as they point out, for example in the discussion of Zhan and Datta (2024). More recent related work on modeling binary data with RFs is reported in Saha and Datta (2025).

There are Bayesian implementations of ML models, which offer the possibility of obtaining valid uncertainty intervals, under correct model specification. MacBride et al. (2025) embed a RF in a Bayesian hierarchical discrete spatial smoothing algorithm, but the frequentist RF implementation does not give a procedure that fully acknowledges the total uncertainty. Jiang and Wakefield (2023) combine a GRF with a Bayesian adaptive regression tree component, so the uncertainty quantification is fully Bayesian, but computational complexity is a serious deterrent to this approach and the simulations show that achieving nominal frequentist coverage with such an approach is challenging.

Though ML approaches for prevalence mapping are very much in their infancy, they provide an exciting prospect for fully exploiting auxiliary information and providing estimates at granular levels, though assessing the accuracy and calibration of predictions and intervals is likely to remain a challenge.

5.7 SAE `R` Packages

There are many R packages implementing variations of Fay-Herriot area-level models and unit-level linear models and here we mention only a subset. The `sae` package (Molina and Marhuenda, 2015) uses frequentist inference, and allows for spatial random effects, via a SAR

model and space-time estimation with an autoregressive regression model of order 1 (AR1) model. The `emdi` package (Kreutzmann et al., 2019) also includes methods for area-level and unit-level modeling, allowing for SAR spatial random effects with linear unit-level models. The possibility of a Box-Cox transformation on the response may be integrated within the modeling procedure. The package `PovMap` (Edochie et al., 2024) adds features to `emdi`. The `tipsae` package (De Nicolò and Gardini, 2024) allows the fitting of area-level beta models.

Many Bayesian implementations of SAE models use the `R` package `INLA` (Lindgren and Rue, 2015; Rue et al., 2017; Bakka et al., 2018), which implements approximate full Bayesian inference using the INLA method (Rue et al., 2009), and can be used for both discrete and continuous spatial models. The package is a popular choice for those using BYM2 or SPDE models. The `TMB` package (Kristensen et al., 2016) provides empirical Bayesian inference that allows fast inference for spatial models (Osgood-Zimmerman and Wakefield, 2023).

Implementations of unit-level logistic models are less common. The `RiskMap` package (Giorgi, 2024) allows unit-level logistic modeling, using GRF spatial models with Matérn correlation functions, and is the successor to the `PrevMap` package (Giorgi and Diggle, 2017). An SPDE implementation of geostatistical models is available in the `mbg` package (Henry and Mayala, 2025). Altay et al. (2025) provide an `R` package for MBG models, accounting for DHS jittering.

The `SUMMER` package (Li et al., 2025) allows Bayesian spatial and space-time modeling for area-level and unit-level models, with the latter offering linear and logistic choices. In the space-time formulation, the BYM2 spatial model is combined with AR1 or random walk models of either order 1 or 2 (RW1 or RW2) temporal models, along with a variety of space-time interaction models. Implementation is based on INLA. The `SUMMER` package was used in Li et al. (2019) and UN IGME (2021) for subnational child mortality modeling over time using area-level and unit-level betabinomial models. A number of the SAE models that are available in the `SUMMER` package have also been incorporated in the `survey` package (Lumley, 2004, 2024). The `surveyPrev` package (Dong et al., 2025) further streamlines the full workflow of spatial prevalence estimation, from data preparation to modeling and visualization, and focuses on DHS and MICS data. It has an accompanying shinyApp[3].

6. TWO CULTURES IN PRACTICE: HIV PREVALENCE MAPPING FOR ADULT WOMEN IN ZAMBIA

6.1 SAE and MBG Approaches

We return to the 2018 Zambia DHS data introduced in Section 1, in which the aim is to estimate the HIV prevalence/risk, among women aged 15–49, across Admin-1 and Admin-2 areas. Zambia has one of the highest HIV burdens in the world, as highlighted by Dwyer-Lindgren et al. (2019) and the subnational distribution of prevalence has been the subject of a number of studies, see for example, Mweemba et al. (2022) and Cuadros et al. (2023). Our analyses are not intended to be definitive or comprehensive, but rather to compare and contrast SAE and MBG approaches using various models.

Below we distinguish between methods that estimate the prevalence, p_i, versus those that estimate the risk, r_i, in areas $i = 1, \ldots, m$. We now summarize the seven models we use for comparison. Additional details on each model are provided in the Supplementary Materials, where we also report on a number of other models.

- M1 **Direct:** We use the Hájek estimator $\hat{p}_i^{\mathrm{w}}$, as defined in (2), for weighted (direct) estimation.
- M2 **Fay-Herriot BYM2 without covariates:** Given direct estimates $\hat{p}_i^{\mathrm{w}}$, we compute $\hat{\theta}_i^{\mathrm{w}} = \mathrm{logit}(\hat{p}_i^{\mathrm{w}})$ and estimate the associated sampling variances, V_i. These are used as inputs to an area-level model following (3) and (4) with a linking model that includes a BYM2 spatial prior on the area-level random effects, but no covariates.
- M3 **Fay-Herriot BYM2 with covariates:** As the previous model, except with area-level covariates.
- M4 **Betabinomial BYM2 without covariates:** A betabinomial unit-level sampling model that allows for overdispersion at the cluster level, with a linking model that includes BYM2 spatial prior on the area-level random effects, but no covariates apart from an urban/rural indicator, which we discuss more fully below.
- M5 **Betabinomial BYM2 with covariates:** As the previous model, except with unit-level covariates.
- M6 **Betabinomial GRF without covariates:** A betabinomial unit-level sampling model that allows for overdispersion at the cluster level, with a GRF spatial prior on the area-level random effects and without covariates, apart from an urban/rural indicator.
- M7 **Betabinomial GRF with covariates:** As the previous model, except with unit-level covariates.

Models M1–M3 estimate prevalence, while models M4–M7 estimate risk. Models M2–M5 are more in the spirit of an SAE approach, while M6 and M7 are examples of an MBG formulation.

For the unit-level models (M4–M7) we need to account for the design. Since the stratification is urban/rural crossed with Admin-1 regions, there is a justification for Admin-1 crossed with urban/rural interactions. However, for reasons of parsimony, we assume a single fixed effect term for urban/rural combined with spatial random effects. Under all approaches, we wish to produce estimates

[3] https://sae4health.stat.uw.edu

at both Admin-1 and Admin-2 levels. The GRF continuous spatial models give both levels under a single model, suitably aggregated. We could fit Admin-2 spatial BYM2 models only, and then aggregate up to Admin-1, but we prefer to pick as simple a model as possible to achieve estimates at the desired Admin level. Hence, for each of the BYM2 approaches (models M2–M5), we fit two models one with BYM2 spatial effects at Admin-1 and one with these effects at Admin-2.

At the Admin-2 level, out of the 115 areas, there are 3 areas with no data, and 14 for which the variance estimates are either zero or (inaccurately) close to zero because of sparse data. For all 17 areas, when carrying out the Fay-Herriot analyses at Admin-2, we include these areas as missing data. We comment further on this aspect in Section 6.2.

A model that we have found useful for estimation at the Admin-2 level in other contexts is a nested BYM2 model with fixed effects at the Admin-1 level and *nested* BYM2 models within each Admin-1 area (with a sum-to-zero effect on the random effects within each Admin-1 area). We fitted this model also, but the results were very similar to the non-nested versions (as can be seen in the Supplementary Materials), and so we do not include the results in the main paper. The similarity of results in this example is due to the relative abundance of data at Admin-1 (10 areas only) and the non-rarity of the outcome.

The betabinomial model with overdispersion parameter d, for a generic cluster with sample size n and risk r, has variance,

$$nr(1-r)[1+(n-1)/(d+1)].$$

Hence, the excess binomial variation is $1+(n-1)/(d+1)$, which helps in the interpretation of d.

We use three covariates: *access*, based on estimated travel times to cities in 2015 (Weiss et al., 2018); *malaria incidence*, as estimated for 2018 by the Malaria Atlas Project (Hay and Snow, 2006) and *night time light intensity*, as observed via satellite imagery in 2016 (Román et al., 2018). We transform the night time lights variable to be $\log(1+x)$ where x is a measure of the intensity of night time lights, so that the distribution of these variables are not too skewed which could result in a small number of points influencing the predictions heavily. All covariates are standardized to have zero mean and unit variance. We extract all covariate values on a $1\,\mathrm{km} \times 1\,\mathrm{km}$ grid of locations based on the population raster used by WorldPop. For the unit-level models, we assign covariate values to each sampled cluster based on the centroid of the cluster location.

In order to perform aggregation with the unit-level models we need to evaluate the area-level estimate (18). The values of the three covariates and the population density values are all routinely available at the grid locations. In addition, we require aggregation over urban/rural. The complexity of this aggregation depends on the model fitted, as detailed in the Supplementary Materials. In the most complex case, we require the urban/rural indicator at each grid point, which is not routinely available. We use the following approach to produce pixel-level maps of urban/rural status.

We assume that we know the correct fraction of urban to total population in each Admin-1 area, which will be used to ensure the pixel urban/rural map is consistent with the area fraction. When no changes have been made to the sampling frame, the latter may be found in the survey reports; for this study we were able to obtain the fraction of the urban population at Admin-1, directly from the census. When the sampling frame has been changed, an updated list must be acquired. The urban/rural pixel map is produced using the following steps:

1. From WorldPop, download $100\,\mathrm{m} \times 100\,\mathrm{m}$ pixel maps of population density for the year matching the list of known urban proportions (this is 2010 for Zambia, the year of the last census), and age-specific population density maps for the year for which estimates are desired (this is 2018 for Zambia).
2. For each Admin-1 area, select a threshold and set pixels with population above the threshold as urban and values below the threshold as rural. Using the all-age population density map, the thresholds are set to the level for which the resulting urban proportions are equal to the known urban proportion for each Admin-1 area.
3. This Admin-1 consistent threshold is used across the whole map to obtain a grid-level set of urban/rural indicators.

In general, step 1. can introduce error, as the populations may neither be accurate, nor match exactly (in terms of age) the target population. Quantifying this uncertainty is difficult, however.

All computations were implemented in `R` (R Core Team, 2024). We conduct approximate Bayesian inference via INLA using the `SUMMER` and `surveyPrev` packages for models M2–M5, and our own INLA code with an SPDE representation of the GRF for models M6 and M7. For all covariate models, we adopt default priors on the intercept and slope parameters (independent Gaussian with zero mean and precision 0.001). For models with BYM2 random effects, we use penalized complexity (PC) priors for the variance parameters, as suggested by Riebler et al. (2016), with a scaled precision matrix. For the GRF models with SPDE random effects, we also use PC priors, as suggested by Fuglstad et al. (2019). See the Supplementary Materials for specific hyperparameter choices for these priors.

6.2 Results

Ninety-three percent of women who were eligible for HIV testing and were interviewed, consented and provided a blood specimen for HIV testing. Removing observations without geographic locations, the national weighted estimate of HIV prevalence for women 15–49 is 14.3% (95% interval: 13.2–15.4%) and is more than twice as high in urban areas, 20.4% (18.6–22.2%) as in rural areas, 8.9% (8.0–9.9%). Based only on direct point estimates, $\hat{p}_i^{\mathrm{w}}$, (see Figure 1 for area names), Copperbelt has the highest female HIV prevalence among Admin-1 areas, with 19.9% (16.5–23.4%) and is more than three times greater than Muchinga, the lowest, with prevalence 6.3% (4.6–8.0%).

Figures 2 and 3 provide Admin-1 and Admin-2 level HIV prevalence point estimates and coefficients of variation (CV). To keep the figures at a readable size, we report on the direct estimation model and the covariate models only, as described in Section 6.1.

We first discuss the Admin-1 level results. At this level, all point estimates produce visually similar maps. However, a more detailed comparison included in the Supplementary Materials indicates that the four unit-level models produce nationally aggregated estimates that show slight downward bias when compared with the weighted estimate, indicating that the design is not being fully accounted for in the unit-level models, and/or there is an aggregation issue. Interestingly, the covariate model results show slightly more bias than the no covariate results, which hints at aggregation being at least part of the issue.

The top half of Table 1 contains a range of inferential summaries for the Admin-1 analyses. The unit-level betabinomial BYM2 and GRF models shows noticeable excess-binomial variation. The overdispersion factor, for the median cluster sample size of 24, gives increases of 46–60% increase over the nominal binomial variation. The overdispersion decreases a little when covariates are added to the models, but there is little change in the spatial standard deviation parameters. The spatial standard deviation drops a little when covariates are added to the model, but in general, the changes are small.

There are marked differences in the CVs at the Admin-1 in the top half of Figure 3 and in the interval widths in Table 1. As expected, the direct estimates have the greatest uncertainty, which is reduced a little in the Fay-Herriot BYM2 models. At Admin-1, adding covariates reduces the uncertainty in general, though not by much, and the GRF models produce the narrowest intervals.

For Admin-1, we would use the weighted or Fay-Herriot BYM2 results, given the fewer assumptions, as compared to the unit-level models. The GRF models have potential aggregation issues and the narrow intervals may not correctly reflect uncertainty, since there is evidence

TABLE 1

Model summaries, with Cov = No/Yes being a label for presence of covariates in the model. Average interval width (expressed as a percentage) and variance components, at Admin-1 (top half) and Admin-2 (bottom half). Interval Width is for a 95% interval, and is expressed as a percentage. The excess binomial variation (Excess Binom) is the increase over the binomial model for a cluster of median size. In the BYM2 models the spatial Std Dev is the standard deviation of the BYM2 random effects, and in the GRF models it represents the Matérn standard deviation (so these are not comparable). Note that the same GRF model is used at both Admin-1 and Admin-2 levels, so that the spatial Std Dev estimates are the same; the Admin-1 and Admin-2 results are obtained through aggregation from the continuous surface. The Interval Width summaries for the direct model in the bottom half are calculated over the Admin-2 areas with data and valid variance estimates (98 districts versus the full set of 115).

Model	Cov	Interval Width	Excess Binom	Spatial Std Dev
Direct	–	5.07	–	–
Fay-Herriot BYM2	No	4.80	–	0.37
Fay-Herriot BYM2	Yes	4.82	–	0.33
Betabinomial BYM2	No	4.00	1.60	0.26
Betabinomial BYM2	Yes	3.97	1.53	0.26
Betabinomial GRF	No	3.64	1.48	0.40
Betabinomial GRF	Yes	3.51	1.46	0.41
Direct	–	12.75	–	–
Fay-Herriot BYM2	No	11.72	–	0.64
Fay-Herriot BYM2	Yes	10.61	–	0.53
Betabinomial BYM2	No	8.69	1.48	0.34
Betabinomial BYM2	Yes	7.93	1.44	0.33
Betabinomial GRF	No	6.38	1.48	0.40
Betabinomial GRF	Yes	5.59	1.46	0.41

that these intervals have low frequentist coverage, at least at Admin-2 (as we discuss shortly, see Table 2).

Turning to the Admin-2 point estimates, and referring to Figures 2 and 3, we see greater differences between methods, due to the sparsity of data at this level. The greatest variation across areas occurs with direct estimation (recall that three areas have no clusters, and therefore no estimates, and 14 additional areas have insufficient data to compute variance estimates). The least variation across areas (i.e., the greatest shrinkage) is seen with the GRF models. In terms of the magnitude of spatial variation, the BYM2 models lie between the direct and GRF approaches. In the bottom half of Table 1 we see that, as we would expect, the covariate models have reduced interval estimates, when compared to the no covariate models. Details on the parameter estimates are in the Supplementary Materials, but here we note that the only variable that had a significant impact was *access*. The GRF models have the narrowest intervals but, as we discuss further in Section 6.3, these intervals appear poorly calibrated.

The decision as to which model we prefer at Admin-2, is more difficult than at Admin-1. Due to the poor calibration of unit-level models, and aggregation issues, we

would choose the Fay-Herriot model with BYM2 random effects and covariates. One change we would make, if we were involved in a substantive collaboration, would be to use some form of variance smoothing, as described in Section 3.3, for the 14 areas with data but problematic variance estimates.

In Zambia, the data are relatively abundant at the Admin-2 level, but this is frequently not the case and in other analyses in LMICs, for which the data are more sparse, we have used unit-level models, because the amount of missing data meant that the fitting of Fay-Herriot models was no longer tenable. For unit-level models (at Admin-2 or lower) we have found it useful to include fixed effects at higher levels, with spatial random effects nested within these levels.

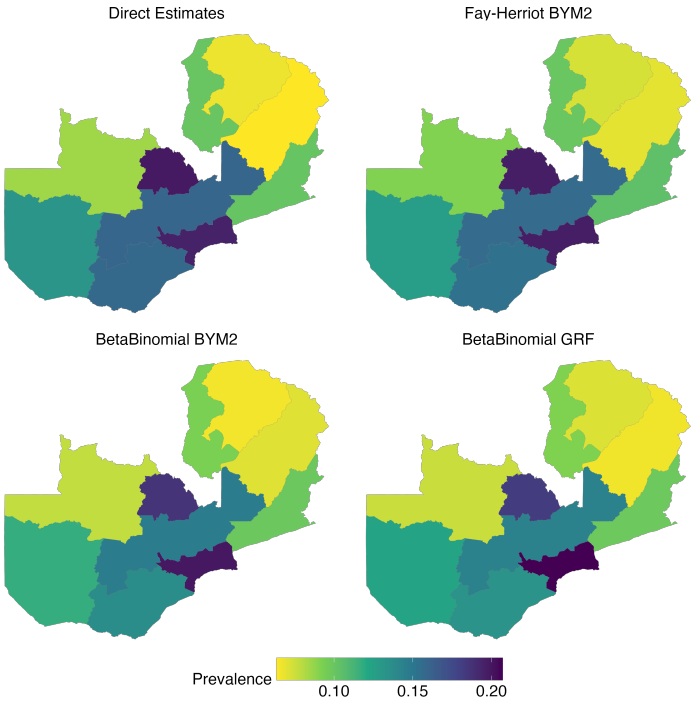
The Supplementary Materials contain a variety of graphical summaries. These include ranking plots, which can be very informative and are straightforward to construct, given samples from the posterior distribution.

6.3 Model Comparison

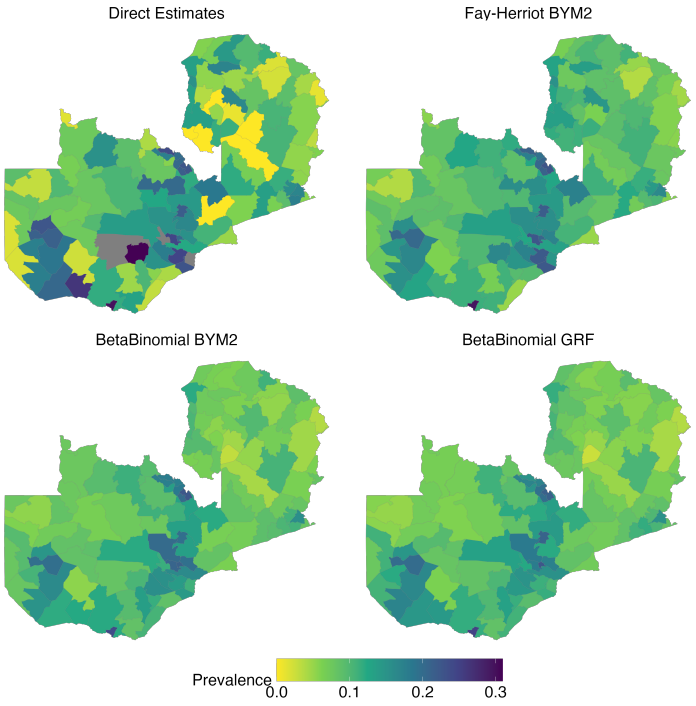
To assess the different models, we carried out a leave-one-out (LOO) cross validation procedure. We do not directly observe prevalence p_i (or risk r_i) and the input data for the direct and Fay-Herriot approaches is different to the input data for the unit-level models. Consequently, we predict the weighted estimates for Admin-2 areas, based on datasets in which we systematically leave out each Admin-2 area in turn. We then predict the logit of the direct estimate $\text{logit}(\hat{p}_i^{\text{w}})$ using the remaining data (all but seventeen areas contained a viable direct estimate and standard error). A variety of metrics were considered to assess the consistency between the left out data and the model predictions. In Table 2 we show the results for a subset of metrics, namely the Mean Absolute Error (MAE), the coverage of an 80% predictive intervals and the log score (which is sometimes referred to as the log of the conditional predictive ordinate). We emphasize that all metrics are computed after logit-transformation of the estimates, i.e., we predict $\text{logit}(\hat{p}_i)$. A full description of the metrics are given in the Supplementary Materials.

We emphasize that these metrics provide only part of the consideration on choosing a model and should not be taken as definitive. Cross-validation for dependent data is currently a topic of great interest, and LOO has been critiqued but as yet no consensus on the best approach has emerged (Bürkner et al., 2020; Adin et al., 2024).

In Table 2, we see that in terms of coverage and the log score, the Fay-Herriot BYM2 model with covariates performs well. A notable feature is the poor coverage of the unit-level models, with the betabinomial GRF with covariates model having a coverage of just 56%. All of these results are consistent with the average interval widths in Table 1 and we conclude that, in this example at least,



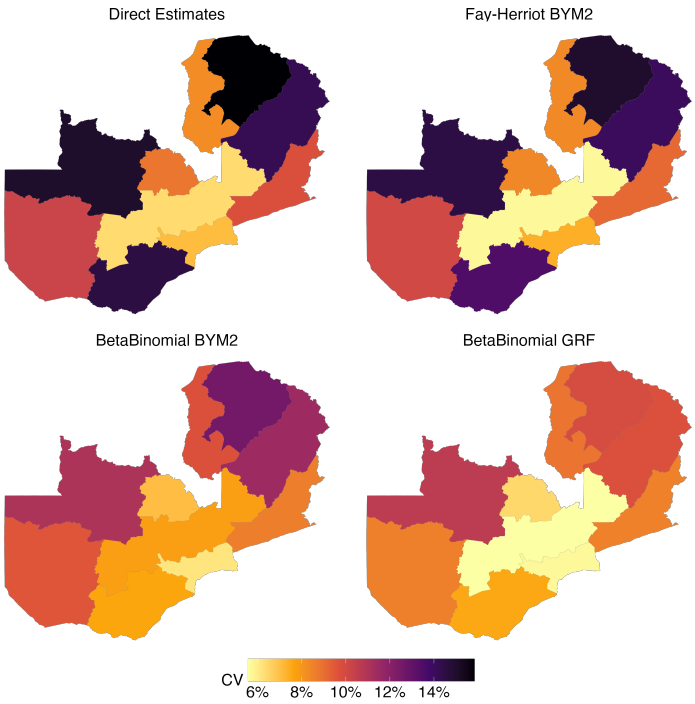
(a) Admin-1 prevalence.



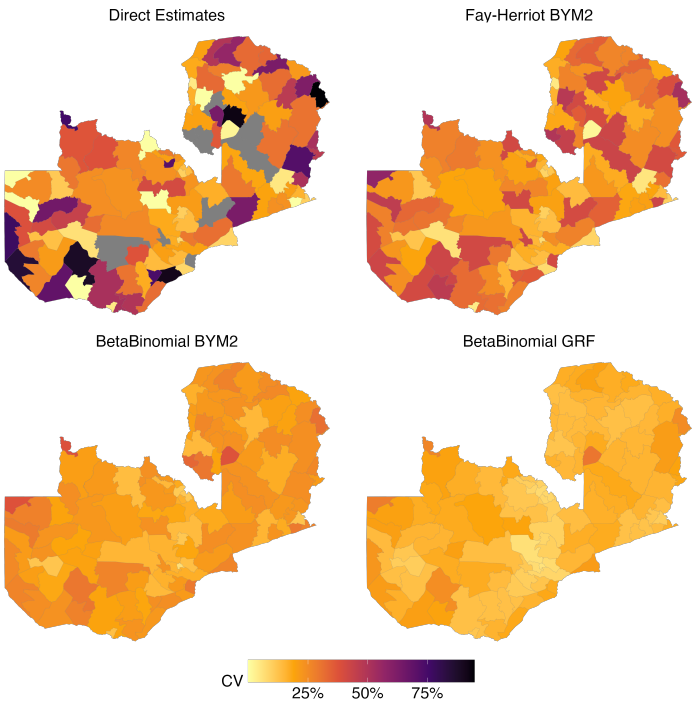
(b) Admin-2 prevalence.

FIG 2. *HIV prevalence/risk estimates for women aged 15–49 in Zambia in 2018: (a) Admin-1 areas, and (b) Admin-2 areas. The BYM2 and GRF models contain covariates.*

the unit-level models are all producing interval estimates that are overly-optimistic, a dangerous phenomena. The log score results favor the Fay-Herriot models over the betabinomial models but all models are close, apart from



(a) Admin-1 CV.



(b) Admin-2 CV.

FIG 3. *Coefficients of variation, expressed as a percentage, for prevalence/risk estimates for women aged 15–49 in Zambia in 2018: (a) Admin-1 areas, and (b) Admin-2 areas. The BYM2 and GRF models contain covariates.*

the betabinomial GRF model with covariates, which has significantly poorer performance. For the MAE, the unit-level models provide the lowest values, with the GRF models giving slighter lower values than the BYM2 models.

TABLE 2

Cross validation summary measures over left-out Admin-2 areas, with Cov = No/Yes being a label for presence of covariates in the model. Mean Absolute Error (MAE) is on the logit scale, nominal coverage is at 80% level and log score is the log of the predictive density measured at the observed data, which is logit$(\hat{p}_i^w)$. The best scores are highlighted in **bold**.

Model	Cov	MAE	Coverage	log score
Fay-Herriot BYM2	No	0.601	74	-1.033
Fay-Herriot BYM2	Yes	0.578	**82**	**-1.003**
Betabinomial BYM2	No	0.552	62	-1.062
Betabinomial BYM2	Yes	0.555	64	-1.150
Betabinomial GRF	No	**0.545**	60	-1.180
Betabinomial GRF	Yes	0.546	56	-1.492

7. DISCUSSION

Prevalence mapping is a crucial aid to revealing inequalities in LMICs. A beneficial approach to modeling is to bridge the rich literatures of survey sampling and spatial statistics, while avoiding a dogmatic approach in one's thinking. To elaborate, one should not expect one approach to be universally best, but rather the models chosen should depend on the geographical target level, the data sparsity, the informativeness of the design, and the availability of useful auxiliary information.

Weighted estimates are very appealing, since they are robust and design consistent, which is very desirable when the target is an empirical prevalence. If they produce estimates with acceptable precision, they should be used. Unfortunately, in many situations the data will be too sparse for successful use of weighted estimation. Area-level Fay-Herriot models retain design consistency properties, but use the totality of data to increase precision. Including spatial random effects at the target level of inference will be advantageous when strong covariates are not available, though one may want to include fixed effects at higher levels, to guard against over-smoothing. For example, if the target is Admin-2, one may model with fixed effects at Admin-1 and nested spatial random effects for each collection of Admin-2 areas within each Admin-1. This approach will also provide an approximate form of benchmarking (McGovern et al., 2024) and has an appealing link with the design since stratification is often carried out at Admin-1. For Fay-Herriot approaches, modeling the variances of the weighted estimates to provide usable estimates and robustness may be advantageous (Maples et al., 2009; Gao and Wakefield, 2023b).

The inferential path followed (frequentist or Bayes) and the particular type of discrete spatial model used (ICAR, CAR, SAR,...) are of lesser importance. However, when the data are sparse, a full posterior distribution has advantages over frequentist point estimates and uncertainty summaries, as it accounts for non-normality of sampling distributions and appropriately averages over nuisance parameters via marginalization.

Benchmarking subnational estimates to officially accepted national levels is desirable (Särndal, 2007; Zhang et al., 2014), but it is rare for accurate national prevalences to be available in LMICs. Generalized regression (GREG) approaches (Särndal, 2007) are appealing as a way of extending Fay-Herriot type models via the use of more complex models (Gao and Wakefield, 2024).

When weighted estimates, with reliable standard errors, are not available for sufficient areas, so that Fay-Herriot models are not an option, one may turn to unit-level models. We have a preference for discrete spatial models at the target geography, under the rationale, of not wanting to model at levels any lower than needed. As an extreme example, suppose we wanted a national estimate. If model assumptions are met and precise information required for aggregation is available then the theory tells us that a more precise estimate will be obtained from a spatial model. However, simply calculating a weighted estimate will be preferable, since it is based on few assumptions, avoids difficult aggregation steps, and will account for the survey design (and will often be precise enough for most purposes). A simple sanity check for any unit-level model is aggregating to the national level and comparing with the weighted estimate and its uncertainty interval.

Spatially continuous MBG models are aesthetically appealing, since they can produce estimates at any required geographical level. However, the aggregation step may inject bias. MBG approaches may produce estimates with the lowest uncertainty (because we are putting in the most assumptions), but one must then question whether these estimates are accurate measures of the true uncertainty. The issue is that the full uncertainty in the data, including the inputs to the aggregation process, may not be being fully acknowledged. We elaborate on this point shortly.

Aggregation can be very challenging with unit-level models. For example, if there is informative sampling over urban/rural strata, we need to investigate whether ignoring this aspect will lead to bias. Estimating an urban/rural association is relatively uncomplicated, but producing reliable fine-scale maps of urban/rural, which are needed for spatial aggregation, is an ongoing research problem. Treating space as continuous increases the challenge as one needs to rely on fine-scale population density rasters to perform aggregation to the areal level. Both population and covariate rasters are estimated based on other data sources, and there is a need for further investigation to understand the consequences of not acknowledging the inherent uncertainty when used in an aggregation scheme. The result of aggregation is typically an areal risk and examining the distinction between risk and prevalence at different spatial scales is worthwhile.

This all being said, MBG models may be the only viable option when data are sparse relative to the geographical target scale, particularly when strongly predictive covariates are available. In this case, one should, however, view the results more cautiously.

There are cases in which a continuously-indexed model is necessary, for example, to give a principled way of handling data that is available with limited geographical information (e.g., area only information, rather than points) or under different geographical partitions (Marquez and Wakefield, 2021; Wilson and Wakefield, 2020).

A fundamental issue with nonlinear unit-level models, including MBG approaches, is that they do not produce design consistent estimates. The only way to guarantee such consistency is to use the weights, but these are not utilized in standard unit-level models. You and Rao (2002b); Pfeffermann and Sverchkov (2007); Guadarrama et al. (2018) and Cho et al. (2024) have made some progress in achieving design consistency for linear unit-level models, but it is not straightforward to extend to logistic models, let alone continuous spatial models, especially since DHS and other surveys only report the final weights and not the constituent parts, which are required for areas with no data. Unit-level models may neglect to account for all aspects of the sampling design and frequently make the strong assumption that the design is ignorable. Often in LMICs the design is not ignorable because of urban/rural stratification and the strong association between many outcomes and this classification.

Section 6 demonstrates the clear importance of model choices both in terms of central predictions and uncertainty. Since it will always be possible to argue for many reasonable models, there is a strong need for formal validation techniques to compare different models in terms of their ability to estimate areal prevalences. However, the spatial misalignment between the point-referenced observations and the desired areal quantities makes validation a challenging task. One method for model validation is an approach that frames direct estimates as observations and scores different models based on their ability to predict direct estimates, but there are many choices (such as which data to leave out, and which metric to use for comparison of predictions and left-out data), and it is not an ideal approach since it is only possible at a large geographical scale which will often not coincide with the inferential target. One should be wary of using more complex models unless one can show they are better than simpler models through validation. Cross-validation can also be carried out at the unit-level, but there is little experience with this

enterprise, which is clearly an important area for future research.

The whole consistency of estimates argument is best framed within a spectrum of approaches, with weighted estimation at one end and MBG models at the other, and it is natural to change the approach as the sparsity of data changes. A very important research question is when, as a function of sample size, to transition between approaches.

In our experience, for estimation when the data are rich, for example (usually) when constructing Admin-1 estimates using DHS or MICS surveys, direct (weighted) estimation is a safe choice. If the survey does not provide precise enough Admin-1 estimates using weighted estimation, one should turn to models. Our preference is to begin with simple Fay-Herriot area-level models. For finer levels such as Admin-2 or Admin-3 (depending on the country, survey size and rarity of outcome), one can use spatial statistical unit-level models, but ensure that the model contains terms to acknowledge the complex design. In general, one should use models that one is comfortable with, whether they be traditional SAE style unit-level models or MBG versions. We are cautious with unit-level approaches, particularly with respect to acknowledging the design and aggregation. A critical evaluation of the latter is crucial. A key point is that the goal of the analysis should determine the approach, and different goals may call for different approaches, and one should not expect a single model to provide the best choice for all questions. Corral et al. (2021) reached a similar conclusion. When producing official statistics using SAE, Tzavidis et al. (2018) suggest that one consider progressively finer geographies while at each level assessing the feasibility of producing SAE estimates for the current level. This is broadly consistent with our recommendations.

ACKNOWLEDGMENTS

We are grateful to the Space Time Analysis Bayes (STAB) working group for discussion and feedback on the paper. We also acknowledge the DHS for access and use of the data, and for permission to make available the cluster aggregates and their displaced locations.

SUPPLEMENTAL MATERIAL

Supplement to "The Two Cultures of Prevalence Mapping: Small Area Estimation and Model-Based Geostatistics", contains further details on the application in Section 6, including additional results.

REFERENCES

Adin, A., E. T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics 62*, 100843.

Alkema, L., G. W. Jones, and C. U. Lai (2013). Levels of urbanization in the world's countries: testing consistency of estimates based on national definitions. *Journal of Population Research 30*, 291–304.

Altay, U., J. Paige, A. Riebler, and G.-A. Fuglstad (2024). Impact of jittering on raster-and distance-based geostatistical analyses of DHS data. *Statistical Modelling 25*, 55–74.

Altay, U., J. Paige, A. Riebler, and G.-A. Fuglstad (2025). Geoadjust: Adjusting for positional uncertainty in geostatistical analysis of DHS data. *R-Journal 16*, 15–26.

Altay, U., J. Paige, A. Riebler, and F. GA (2022). Accounting for spatial anonymization in DHS household surveys. *arXiv preprint arXiv 2202*.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023). Prediction-powered inference. *Science 382*, 669–674.

Bakka, H., H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren (2018). Spatial modeling with R-INLA: A review. *WIREs Computational Statistics 10*, e1443.

Battese, G. E., R. M. Harter, and W. A. Fuller (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., W. Basel, and J. Maples (2016). An overview of the US Census Bureau's small area income and poverty estimates program. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation*, pp. 349–378. Wiley Online Library.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregressions. *Biometrika 82*, 733–746.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics 43*, 1–59.

Bhatt, S., E. Cameron, S. Flaxman, D. Weiss, D. Smith, and P. Gething (2017). Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of The Royal Society Interface 14*, 20170520.

Bhatt, S., D. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. Battle, C. Moyes, A. Henry, P. Eckhoff, et al. (2015). The effect of malaria control on plasmodium falciparum in Africa between 2000 and 2015. *Nature 526*, 207–211.

Bilton, P., G. Jones, S. Ganesh, and S. Haslett (2017). Classification trees for poverty mapping. *Computational Statistics and Data Analysis 115*, 53–66.

Bosco, C., V. Alegana, T. Bird, C. Pezzulo, L. Bengtsson, A. Sorichetta, J. Steele, G. Hornby, C. Ruktanonchai, N. Ruktanonchai, E. Wetter, and A. J. Tatem (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface 14*, 20160825.

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

Browne, A. J., M. G. Chipeta, G. Haines-Woodhouse, E. P. Kumaran, B. H. K. Hamadani, S. Zaraa, N. J. Henry, A. Deshpande, R. C. Reiner, N. P. Day, et al. (2021). Global antibiotic consumption and usage in humans, 2000–18: a spatial modelling study. *The Lancet Planetary Health 5*, e893–e904.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021). Using satellite imagery to understand and promote sustainable development. *Science 371*, eabe8628.

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

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

Burstein, R., N. J. Henry, M. L. Collison, L. B. Marczak, A. Sligar, S. Watson, N. Marquez, M. Abbasalizad-Farhangi, M. Abbasi, F. Abd-Allah, et al. (2019). Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature 574*, 353–358.

Burstein, R., H. Wang, R. C. Reiner Jr, and S. I. Hay (2018). Development and validation of a new method for indirect estimation of neonatal, infant, and child mortality trends using summary birth histories. *PLoS Medicine 15*, e1002687.

Chan-Golston, A., S. Banerjee, T. R. Belin, S. E. Roth, and M. L. Prelip (2022). Bayesian finite-population inference with spatially correlated measurements. *Japanese Journal of Statistics and Data Science 5*, 407–430.

Chan-Golston, A. M., S. Banerjee, and M. S. Handcock (2019). Bayesian finite population modeling for spatial process settings. *Environmetrics 31*, e2606.

Chi, G., H. Fang, S. Chatterjee, and J. E. Blumenstock (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences 119*, e2113658119.

Cho, Y., M. Guadarrama-Sanz, I. Molina, A. Eideh, and E. Berg (2024). Optimal predictors of general small area parameters under an informative sample design using parametric sample distribution models. *Journal of Survey Statistics and Methodology 12*, 1430–1463.

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

Cloutier, E. C. and É. Langlet (2014). *Aboriginal Peoples Survey, 2012: Concepts and Methods Guide*. Statistics Canada=Statistique Canada.

Congdon, P. and P. Lloyd (2010). Estimating small area diabetes prevalence in the US using the Behavioral Risk Factor Surveillance System. *Journal of Data Science 8*, 235–252.

Corral, P., H. Henderson, and S. Segovia (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics 172*, 103377.

Corral, P., K. Himelein, K. McGee, and I. Molina (2021). A map of the poor or a poor map? *Mathematics 9*, 2780.

Corral, P., I. Molina, A. Cojocaru, and S. Segovia (2022). *Guidelines to Small Area Estimation for Poverty Mapping*. World Bank.

Cuadros, D. F., T. Chowdhury, M. Milali, D. T. Citron, S. Nyimbili, N. Vlahakis, T. Savory, L. Mulenga, S. Sivile, K. D. Zyambo, et al. (2023). Geospatial patterns of progress towards UNAIDS '95-95-95'targets and community vulnerability in Zambia: insights from population-based hiv impact assessments. *BMJ Global Health 8*, e012629.

Das, S. and S. Haslett (2019). A comparison of methods for poverty estimation in developing countries. *International Statistical Review 87*, 368–392.

Datta, G. S., T. Kubokawa, I. Molina, and J. Rao (2011). Estimation of mean squared error of model-based small area estimators. *Test 20*, 367–388.

Datta, G. S. and A. Mandal (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association 110*, 1735–1744.

De Nicolò, S., M. R. Ferrante, and S. Pacei (2024). Small area estimation of inequality measures using mixtures of Beta. *Journal of the Royal Statistical Society Series A 187*, 85–109.

De Nicolò, S. and A. Gardini (2024). The R package tipsae: Tools for mapping proportions and indicators on the unit interval. *Journal of Statistical Software 108*, 1–36.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical science 30*, 533–558.

Diallo, M. S. and J. N. K. Rao (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics 45*, 1092–1116.

Diggle, P. and E. Giorgi (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association 111*, 1096–1120.

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

Diggle, P. J., R. Menezes, and T.-L. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C 59*, 191–232.

Dong, Q., Z. R. Li, Y. Wu, A. Boskovic, and J. Wakefield (2025). *surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas*. R package version 1.0.0.

Dong, T. and J. Wakefield (2021). Modeling and presentation of health and demographic indicators in a low- and middle-income countries context. *Vaccine 39*, 2584–2594.

Dwyer-Lindgren, L., M. A. Cork, A. Sligar, K. M. Steuben, K. F. Wilson, N. R. Provost, B. K. Mayala, J. D. VanderHeide, M. L. Collison, J. B. Hall, et al. (2019). Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature 570*, 189–193.

Edochie, I., D. Newhouse, N. Tzavidis, T. Schmid, E. Foster, A. L. Hernandez, A. Ouedraogo, A. Sanoh, and A. Savadogo (2025). Small area estimation of poverty in four West African countries by integrating survey and geospatial data. *Journal of Official Statistics 41*, 96–124.

Edochie, I., D. Newhouse, N. Würz, and T. Schmid (2024). *povmap: Extension to the 'emdi' Package*. R package version 1.0.1.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica 71*, 355–364.

Erciulescu, A. L., N. B. Cruze, and B. Nandram (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society, Series A 182*, 283–303.

Fay, R. and R. Herriot (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association 74*, 269–277.

Ferreira, L. Z., C. E. Utazi, L. Huicho, K. Nilsen, F. P. Hartwig, A. J. Tatem, and A. J. Barros (2022). Geographic inequalities in health intervention coverage–mapping the composite coverage index in Peru using geospatial modelling. *BMC Public Health 22*, 2104.

Finley, A. O., H.-E. Andersen, C. Babcock, B. D. Cook, D. C. Morton, and S. Banerjee (2024). Models to support forest inventory and small area estimation using sparsely sampled LiDAR: a case study involving G-LiHT LiDAR in Tanana, Alaska. *Journal of Agricultural, Biological and Environmental Statistics*, 1–28.

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

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association 114*, 445–452.

Gao, P. A. and J. Wakefield (2023a). Pseudo-Bayesian unit level modeling for small area estimation under informative sampling. *arXiv preprint arXiv:2309.12119*.

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

Gao, P. A. and J. Wakefield (2024). Smoothed model-assisted small area estimation of proportions. *Canadian Journal of Statistics 52*, 337–358.

Gelman, A. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology 23*, 127–135.

General Assemby of the United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1.

Georganos, S., T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, and S. Kalogirou (2021). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International 36*, 121–136.

Gething, P., A. Tatem, T. Bird, and C. Burgert-Brucker (2015). *Creating spatial interpolation surfaces with DHS data*. DHS Spatial Analysis Reports No. 11.

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

Ghosh, M., K. Natarajan, T. W. F. Stroud, and B. P. Carlin (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association 93*, 273–282.

Ghosh, M. and J. Rao (1994). Small area estimation: An appraisal. *Statistical Science 9*, 55–93.

Giorgi, E. (2024). *RiskMap: Geo-Statistical Modeling of Spatially Referenced Data*. R package version 0.1.0.

Giorgi, E. and P. J. Diggle (2017). Prevmap: An R package for prevalence mapping. *Journal of Statistical Software, Articles 78*, 1–29.

Giorgi, E., P. J. Diggle, R. W. Snow, and A. M. Noor (2018). Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys. *International Statistical Review 86*, 571–597.

Giorgi, E., S. S. Sesay, D. J. Terlouw, and P. J. Diggle (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society: Series A 178*, 445–464.

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

Graetz, N., J. Friedman, A. Osgood-Zimmerman, R. Burstein, M. H. Biehl, C. Shields, J. F. Mosser, D. C. Casey, A. Deshpande, L. Earl, R. Reiner, S. Ray, N. Fullman, A. Levine, R. Stubbs, B. Mayala, J. Longbottom, A. Browne, S. Bhatt, D. Weiss, P. Gething, A. Mokdad, S. Lim, C. Murray, E. Gakidou, and S. Hay (2018). Mapping local variation in educational attainment across Africa. *Nature 555*, 48–53.

Guadarrama, M., I. Molina, and J. Rao (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis 121*, 20–40.

Gutreuter, S., E. Igumbor, N. Wabiri, M. Desai, and L. Durand (2019). Improving estimates of district HIV prevalence and burden in South Africa using small area estimation techniques. *PloS One 14*, e0212445.

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*. Toronto: Holt, Rinehart and Winston.

Hay, S. I. and R. W. Snow (2006). The Malaria Atlas Project: developing global maps of malaria risk. *PLoS Medicine 3*, e473.

Heaton, M. J., A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, D. Nychka, F. Sun, and A. Zammit-Mangion (2018). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics 24*, 1–28.

Henry, N. and B. Mayala (2025). *mbg: Model-Based Geostatistics*. R package version 1.1.0.

Hirose, M., M. Ghosh, and T. Ghosh (2023). Arc-sin transformation for binomial sample proportions in small area estimation. *Statistica Sinica 33*, 1–23.

Hobza, T. and D. Morales (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics 32*, 661–692.

Hosseinpoor, A. R., N. Bergen, A. J. Barros, K. L. Wong, T. Boerma, and C. G. Victora (2016). Monitoring subnational regional inequalities in health: measurement approaches and challenges. *International Journal for Equity in Health 15*, 1–13.

ICF International (2012). *Demographic and Health Survey Samling and Household Listing Manual*. MD: Claverton: MEASURE DHS.

Janicki, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics-Theory and Methods 49*, 2264–2284.

Janocha, B., R. Donohue, T. Fish, B. Mayala, and T. Croft (2021). *Guidance and recommendations for the use of indicator estimates at the subnational administrative level 2*. DHS Spatial Analysis Reports No. 20. Rockville, Maryland, USA.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*, 790–794.

Jiang, A. Z. and J. Wakefield (2023). BART-SIMP: a novel framework for flexible spatial covariate modeling and prediction using bayesian additive regression trees. *arXiv preprint arXiv:2309.13270*.

Jiang, J. and P. Lahiri (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics 53*, 217–243.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation (with discussion). *Test 15*, 1–96.

Jiang, J. and E.-T. Tang (2011). The best EBLUP in the Fay–Herriot model. *Annals of the Institute of Statistical Mathematics 63*, 1123–1140.

Kinyoki, D. K., A. E. Osgood-Zimmerman, B. V. Pickering, L. E. Schaeffer, L. B. Marczak, A. Lazzar-Atwood, M. L. Collison, N. J. Henry, Z. Abebe, A. A. Adamu, V. Adekanmbi, K. Ahmadi, O. Ajumobi, and A. Al-Eyadhy (2020). Mapping child growth failure across low- and middle-income countries. *Nature 577*, 231–234.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine 19*, 2555–2567.

Krennmair, P. and T. Schmid (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society, Series C 71*, 1865–1894.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software 91*, 1–33.

Kristensen, K. (2014). `TMB`: General random effect model builder tool inspired by ADMB. `R` *package version*.

Kristensen, K., A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software 70*, 1–21.

Lehtonen, R. and A. Veijanen (2009). Design-based methods of estimation for domains and small areas. In *Handbook of Statistics*, Volume 29, pp. 219–249. Elsevier.

León-Novelo, L. G. and T. D. Savitsky (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics 13*, 1608–1645.

Li, Z. R., Y. Hsiao, J. Godwin, B. D. Martin, J. Wakefield, and S. J. Clark (2019). Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLoS One 14*, e0210645.

Li, Z. R., B. D. Martin, T. Q. Dong, G.-A. Fuglstad, J. Paige, A. Riebler, S. Clark, and J. Wakefield (2025). Space-time smoothing of demographic and health indicators using the R package SUMMER. *To Appear in the R Journal*.

Lindgren, F., D. Bolin, and H. Rue (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics 50*, 100599.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software 63*, 1–25.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B 73*, 423–498.

Little, R. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics 28*, 309–372.

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

Lloyd, C. T., H. J. W. Sturrock, D. R. Leasure, W. C. Jochem, A. N. Lázár, and A. J. Tatem (2020). Using GIS and machine learning to classify residential status of urban buildings in low and middle income settings. *Remote Sensing 12*, 3847.

Local Burden of Disease Vaccine Coverage Collaborators and others (2021). Mapping routine measles vaccination in low-and middle-income countries. *Nature 589*, 415.

Lohr, S. L. (2021). *Sampling: Design and Analysis, Third Edition*. Chapman and Hall/CRC.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software 9*, 1–19.

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

MacBride, C., V. Davies, and D. Lee (2025). A spatial autoregressive random forest algorithm for small-area spatial prediction. *The Annals of Applied Statistics 19*, 485–504.

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

Malec, D., J. Sedransk, C. L. Moriarity, and F. B. LeClere (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association 92*, 815–826.

Mao, H., R. Martin, and B. J. Reich (2024). Valid model-free spatial prediction. *Journal of the American Statistical Association 119*, 904–914.

Maples, J., W. Bell, and E. T. Huang (2009). Small area variance modeling with application to county poverty estimates from the American Community Survey. In *Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association*, pp. 5056–5067.

Marhuenda, Y., I. Molina, and D. Morales (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics and Data Analysis 58*, 308–325.

Marhuenda, Y., I. Molina, D. Morales, and J. N. K. Rao (2017). Poverty mapping in small areas under a two-fold nested error regression model. *Journal of the Royal Statistical Society: Series A 180*, 1111–1136.

Marino, M. F., M. G. Ranalli, N. Salvati, and M. Alfo (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics 13*, 1166–1197.

Marquez, N. and J. Wakefield (2021). Harmonizing child mortality data at disparate geographic levels. *Statistical Methods in Medical Research 30*, 1187–1210.

Mayala, B., T. Dontamsetti, T. D. Fish, and T. N. Croft (2019). *Interpolation of DHS Survey Data at Subnational Administrative Level 2*. DHS Spatial Analysis Reports No. 17. Rockville, Maryland, USA.

McConville, K. S., F. J. Breidt, T. C. Lee, and G. G. Moisen (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology 5*, 131–158.

McGovern, A., K. Wilson, and J. Wakefield (2024). Direct-assisted Bayesian unit-level modeling for small area estimation of rare event prevalence. *arXiv preprint arXiv:2408.16129*.

Mercer, L., J. Wakefield, A. Pantazis, A. Lutambi, H. Mosanja, and S. Clark (2015). Small area estimation of childhood mortality in the absence of vital registration. *The Annals of Applied Statistics 9*, 1889–1905.

Michal, V., J. Wakefield, A. M. Schmidt, A. Cavanaugh, B. E. Robinson, and J. Baumgartner (2024). Model-based prediction for small domains using covariates: A comparison of four methods. *Journal of Survey Statistics and Methodology 12*, 1489–1514.

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

Molina, I. and Y. Marhuenda (2015). R package sae: Methodology. *The R Journal 7*, 81–98.

Molina, I., J. Rao, and M. Guadarrama (2019). Small area estimation methods for poverty mapping: a selective review. *Statistics and Applications 17*, 11–22.

Molina, I. and J. N. Rao (2010). Small area estimation of poverty indicators. *Canadian Journal of statistics 38*, 369–385.

Morales, D., M. D. Esteban, A. Pérez, and T. Hobza (2021). *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*. Springer Nature.

Mosser, J. F., W. Gagne-Maynard, P. C. Rao, A. Osgood-Zimmerman, N. Fullman, N. Graetz, R. Burstein, R. L. Updike, P. Y. Liu, S. E. Ray, et al. (2019). Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: a spatial and temporal modelling study. *The Lancet 393*, 1843–1855.

Mweemba, C., P. Hangoma, I. Fwemba, W. Mutale, and F. Masiye (2022). Estimating district HIV prevalence in Zambia using small-area estimation methods (SAE). *Population Health Metrics 20*, 8.

Newhouse, D. (2023). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin 76*, 7–32.

Newhouse, D., A. Ramakrishnan, T. Swartz, J. Merfeld, and P. Lahiri (2025). Small area estimation of monetary poverty in mexico using satellite imagery and machine learning. *Oxford Bulletin of Economics and Statistics*. Published: 14 April.

Osgood-Zimmerman, A., A. I. Millear, R. W. Stubbs, C. Shields, B. V. Pickering, L. Earl, N. Graetz, D. K. Kinyoki, S. E. Ray, S. Bhatt, et al. (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature 555*, 41–47.

Osgood-Zimmerman, A., A. I. Millear, R. W. Stubbs, C. Shields, B. V. Pickering, L. Earl, N. Graetz, D. K. Kinyoki, S. E. Ray, S. Bhatt, A. Browne, R. Burstein, E. Cameron, D. Casey, A. Deshpande, N. Fullman, P. Gething, H. Gibson, N. Henry, M. Herrero, L. Krause, I. Letourneau, A. Levine, P. Liu, J. Longbottom, B. Mayala, J. Mosser, A. Noor, D. Pigott, E. Piwoz, P. Rao, R. Rawat, R. Reiner, D. Smith, D. Weiss, K. Wiens, A. Mokdad, L. S.S., C. Murray, N. Kassebaum, and S. Hay (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature 555*, 41.

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

Otto, M. C. and W. R. Bell (1995). Sampling error modelling of poverty and income statistics for states. In *American Statistical*

Association, Proceedings of the Section on Government Statistics, pp. 160–165.

Paige, J., G.-A. Fuglstad, A. Riebler, and J. Wakefield (2022). Spatial aggregation with respect to a population distribution: Impact on inference. *Spatial Statistics 52*, 100714.

Park, D. K., A. Gelman, and J. Bafumi (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis 12*, 375–385.

Parker, P. A., R. Janicki, and S. H. Holan (2023). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology 11*, 829–857.

Pedersen, J. and J. Liu (2012). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Medicine 9*, e1001289.

Perez-Heydrich, C., J. Warren, C. Burgert, and M. Emch (2013). *Guidelines on the use of DHS GPS Data*. DHS Spatial Analysis Reports No. 8. Rockville, Maryland, USA.

Petrucci, A. and N. Salvati (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics 11*, 169.

Pfefferman, D. (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 179–183.

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

Pfeffermann, D. (2022). Time series modelling of repeated survey data for estimation of finite population parameters. *Journal of the Royal Statistical Society Series A 185*, 1757–1777.

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

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons.

Pratesi, M. and N. Salvati (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications 17*, 113–141.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. and I. Molina (2015). *Small Area Estimation, Second Edition*. New York: John Wiley.

Ratledge, N., G. Cadamuro, B. De La Cuesta, M. Stigler, and M. Burke (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature 611*, 491–495.

Reiner, R. C., N. Graetz, D. C. Casey, C. Troeger, G. M. Garcia, J. F. Mosser, A. Deshpande, S. J. Swartz, S. E. Ray, B. F. Blacker, et al. (2018). Variation in childhood diarrheal morbidity and mortality in Africa, 2000–2015. *New England Journal of Medicine 379*, 1128–1138.

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

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40*, 913–929.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–129.

Román, M. O., Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, et al. (2018). NASA's Black Marble nighttime lights product suite. *Remote Sensing of Environment 210*, 113–143.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika 57*, 377–387.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Application*. Boca Raton: Chapman and Hall/CRC Press.

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

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

Saei, A. and R. Chambers (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. Southampton Statistical Sciences Research Institute, Project Report.

Saha, A., S. Basu, and A. Datta (2023). Random forests for spatially dependent data. *Journal of the American Statistical Association 118*, 665–683.

Saha, A. and A. Datta (2025). Random forests for binary geospatial data. *arXiv preprint arXiv:2302.13828*.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*, 99–119.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

Scott, A. and T. Smith (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association 64*, 830–840.

Semenova, E., Y. Xu, A. Howes, T. Rashid, S. Bhatt, S. Mishra, and S. Flaxman (2022). PriorVAE: encoding spatial priors with variational autoencoders for small-area estimation. *Journal of The Royal Society Interface 19*, 20220094.

Si, Y., N. Pillai, and A. Gelman (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis 10*, 605–625.

Slud, E. V. and T. Maiti (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B 68*, 239–257.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One 10*, e0107042.

Stukel, D. M. and J. N. K. Rao (1997). Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics and Probability Letters 35*, 401–407.

Sugasawa, S. and T. Kubokawa (2015, July). Parametric transformed Fay–Herriot model for small area estimation. *Journal of Multivariate Analysis 139*, 295–311.

Sugasawa, S., T. Kubokawa, and J. N. K. Rao (2018). Small area estimation via unmatched sampling and linking models. *TEST 27*, 407–427.

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

Tang, X. and M. Ghosh (2023). Global-local priors for spatial small area estimation. *Calcutta Statistical Association Bulletin 75*, 141–154.

Tang, X., M. Ghosh, N. S. Ha, and J. Sedransk (2018). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association 113*, 1476–1489.

Tonye, S. G. M., R. Wounang, C. Kouambeng, and P. Vounatsou (2024). The influence of jittering DHS cluster locations on geostatistical model-based estimates of malaria risk in cameroon. *Parasite Epidemiology and Control 27*, e00397.

Torabi, M. and J. Rao (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis 127*, 36–55.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics (with discussion). *Journal of the Royal Statistical Society, Series A 181*, 927–979.

UN IGME (2021). *Subnational Under-five Mortality Estimates, 1990–2019: Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation*. New York: United Nations Children's Fund.

UN System Chief Executives Board for Coordination (2017). *Equality and Non-Discrimination at the Heart of Sustainable Development: A Shared United Nations Framework for Action*. New York.

Utazi, C. E., K. Nilsen, O. Pannell, W. Dotse-Gborgbortsi, and A. J. Tatem (2021). District-level estimation of vaccination coverage: Discrete vs continuous spatial models. *Statistics in Medicine 40*, 2197–2211.

Utazi, C. E., J. Thorley, V. A. Alegana, M. J. Ferrari, S. Takahashi, C. J. E. Metcalf, J. Lessler, and A. J. Tatem (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine 36*, 1583–1591.

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

Valliant, R. (2020). Comparing alternatives for estimation from non-probability samples. *Journal of Survey Statistics and Methodology 8*, 231–263.

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

Verret, F., J. Rao, and M. A. Hidiroglou (2014). Model-based small area estimation under informative sampling. *Survey Methodology 41*, 333–347.

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

Wakefield, J., P. A. Gao, G.-A. Fuglstad, and Z. R. Li (2025). Supplement to "The two cultures of prevalence mapping: small area estimation and model-based geostatistics".

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

Wakefield, J. C., N. G. Best, and L. A. Waller (2000). Bayesian approaches to disease mapping. In P. Elliott, J. C. Wakefield, N. G. Best, and D. Briggs (Eds.), *Spatial Epidemiology: Methods and Applications*, pp. 104–27. Oxford: Oxford University Press.

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting 31*, 980–991.

Warren, J. L., C. Perez-Heydrich, C. R. Burgert, and M. E. Emch (2016). Influence of demographic and health survey point displacements on point-in-polygon analyses. *Spatial Demography 4*, 117–133.

Weiss, D. J., A. Nelson, H. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature 553*, 333–336.

Wieczorek, J. (2024). Design-based conformal prediction. *Survey Methodology 49*, 443–473.

Wikle, C. K. and A. Zammit-Mangion (2023). Statistical deep learning for spatial and spatiotemporal data. *Annual Review of Statistics and Its Application 10*, 247–270.

Williams, M. R. and T. D. Savitsky (2021). Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *International Statistical Review 89*, 72–107.

Wilson, K. and J. Wakefield (2020). Pointless spatial modeling. *Biostatistics 21*, e17–e32.

Wilson, K. and J. Wakefield (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-Temporal Epidemiology 37*, 100421.

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

Wu, Y., Z. R. Li, B. Mayala, H. Wang, P. Gao, J. Paige, G.-A. Fuglstad, C. Moe, J. Godwin, R. Donohue, B. Janocha, T. Croft, and J. Wakefield (2021). *Spatial Modeling for Subnational Administrative level 2 Small-Area Estimation*. DHS Spatial Analysis Reports No. 21. Rockville, Maryland, USA.

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

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications 11*, 2583.

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

You, Y. and J. Rao (2002a). 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 J. Rao (2002b). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics 30*, 3–15.

Zhan, W. and A. Datta (2024). Neural networks for geospatial data. *Journal of the American Statistical Association 120*, 535–547.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association 99*, 250–261.

Zhang, X., J. B. Holt, H. Lu, A. G. Wheaton, E. S. Ford, K. J. Greenlund, and J. B. Croft (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American Journal of Epidemiology 179*, 1025–1033.

Zheng, H. and R. Little (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics 21*, 1–20.