International Statistical Review

*International Statistical Review* (2020), 88, 2, 398–418 doi:10.1111/insr.12400

# Small Area Estimation for Disease Prevalence Mapping

**Jonathan Wakefield[1,2], Taylor Okonek[1] and Jon Pedersen[3]**

[1]*Department of Biostatistics, University of Washington, Seattle, USA*

[2]*Department of Statistics, University of Washington, Seattle, USA*

[3]*Fafo AVF, Oslo, Norway*
*E-mail: jonno@uw.edu*

## Summary

**Small area estimation (SAE) entails estimating characteristics of interest for domains, often geographical areas, in which there may be few or no samples available. SAE has a long history and a wide variety of methods have been suggested, from a bewildering range of philosophical standpoints. We describe design-based and model-based approaches and models that are specified at the area level and at the unit level, focusing on health applications and fully Bayesian spatial models. The use of auxiliary information is a key ingredient for successful inference when response data are sparse, and we discuss a number of approaches that allow the inclusion of covariate data. SAE for HIV prevalence, using data collected from a Demographic Health Survey in Malawi in 2015–2016, is used to illustrate a number of techniques. The potential use of SAE techniques for outcomes related to coronavirus disease 2019 is discussed.**

*Key words*: area-level models; Bayesian methods; complex surveys; design-based inference; direct estimation; indirect estimation; model-based inference; spatial smoothing; unit-level models; weighting.

## 1 Introduction

Small area estimation (SAE) describes the endeavour of producing estimates of quantities of interest, such as means and totals, for domains (usually areas) that have sparse or non-existent response data. SAE is carried out in many fields including health, demography, agriculture, business, education and environmental planning. In this article, we focus on binary health outcomes, for which SAE aids in highlighting geographical disparities and is broadly useful for health planning, resource allocation and budgeting. SAE allows fundamental questions such as 'how many people in my area have condition X or need treatment Y?' Disease mapping is traditionally based on a complete enumeration of disease cases and differs from SAE, which is typically based on a subset of individuals, selected via a survey, which may have a complex design. The standard reference on SAE is Rao & Molina (2015), while an excellent review is provided by Pfeffermann (2013).

In survey sampling, inference has often focused on the *design-based* (or *randomisation*) approach. This focus has carried over into SAE. This approach is quite distinct from the *model-based* approaches that are the bread and butter of mainstream spatial statistics. Both inferential

approaches are discussed in Skinner & Wakefield (2017). Design-based methods assess the frequentist properties of estimators, averaging over all possible samples that could have been drawn, under the specified sampling design. Under this paradigm, the values of the responses in the population are viewed as fixed rather than random. Model-based approaches can be either frequentist or Bayesian. If a hypothetical infinite population model-based approach is taken, a probabilistic model is specified for the responses, which are now viewed as random variables. Modelling may be carried out within the design-based paradigm via model-assisted approaches (Särndal *et al.*, 1992), in which a model is specified but desirable design-based properties are retained, even under model misspecification. A cautious view (Lehtonen & Veijanen, 2009) is that design-based (including model-assisted) inference may be reliable in situations when there are large or medium samples in areas, while if data are sparse, a model-based approach may be a necessity. In a companion article, Datta (2009) reviews, and is more enthusiastic towards, model-based approaches. Random effects modelling is popular under the model-based approach to SAE, and inference for these models is often frequentist through empirical Bayes estimation—this is what we refer to as frequentist model-based inference, rather than the predictive approach described in Valliant *et al.* (2000).

If a model-based approach is taken, a key element of model specification (and a source of contention) is determining how to account for the sampling design. Models may be specified at the *area level* or at the *unit level*. For the former, an important reference is Fay & Herriot (1979), which introduced the idea of modelling a weighted estimate of an area-level characteristic. For the latter, Battese *et al.* (1988) describe a nested error regression model at the level of the sampling unit.

Design-based inference from sample surveys aims to estimate finite population quantities using information about the sampling process. In the model-based framework described above, the finite population is itself a sample from a superpopulation, a random process that can be described by some model. If, in an actual survey, the realised sample and the finite population can be described by the same model, then the sample design is ignorable. A simple random sample (SRS) is ignorable in this sense. However, most real-life household surveys have non-ignorable designs, and most SAE models depend on assumptions specific to the design used. When the design is not ignorable, one needs to incorporate the design into the model. Ideally, such incorporation would include the relevant aspects of the design including design weighting, clustering, non-response corrections and weight adjustments (see Section 2). While this information may be available, many aspects of the sampling frame (such as the locations of all clusters in a design with cluster sampling) are typically not available, at least in sufficient detail to be useful. For surveys such as the Demographic and Health Surveys (DHS), which are extensively carried out in low-income and middle-income countries (LMIC), stratification, clustering and estimation weights will typically be accessible. For surveys in developed countries, little information may be available.

Prevalence mapping (Wakefield, 2020) is a name that has been given to the production of maps displaying the prevalence of health and demographic outcomes, and this endeavour clearly has large overlaps with SAE. While SAE smoothing methods often use area-level models, prevalence mapping exercises often use model-based geostatistics (MBG) methods in which a continuous spatial model is specified. Examples of prevalence mapping using area-level SAE techniques include HIV prevalence (Gutreuter *et al.*, 2019) and the under-5 mortality rate (U5MR) (Dwyer-Lindgren *et al.*, 2014; Mercer *et al.*, 2015; Li *et al.*, 2019). Examples of prevalence mapping with MBG include HIV prevalence (Dwyer-Lindgren *et al.*, 2019), malaria (Gething *et al.*, 2016), U5MR (Golding *et al.*, 2017) and vaccination coverage (Utazi *et al.*, 2018; Mosser *et al.*, 2019).

As we discuss in more detail in Section 6, given the current state of data availability and quality, the use of SAE for coronavirus disease 2019 (COVID-19) endpoints is limited. Seroprevalence data will provide an important resource in the future, if they arise from a carefully designed study, including probability sampling from a well-defined population. In the absence of COVID-19 data, we take as a motivating example SAE of HIV prevalence among women aged 15–29, in districts of Malawi, using data from the 2015–2016 Malawi DHS. We use these data to demonstrate the use of various SAE techniques. This endpoint has some similarities with the prevalence of SARS-CoV-2 (the virus that causes COVID-19), as both AIDS and COVID-19 are infectious diseases for which knowing the geographic prevalence is an important public health endpoint. Nevertheless, there are significant differences that are relevant for how one should build SAE models for COVID-19 and for disease prevalence more generally. HIV spreads dominantly by intimate, direct contact in relatively private settings. In contrast, SARS-CoV-2 spreads dominantly by droplets and aerosols released by breathing or coughing, both in public and private spaces. SARS-CoV-2 appears to spread more than HIV by so-called superspreaders, individuals that infect a high number of others. Spatial smoothing models are not good at picking up very localised local troughs or peaks, and so, SARS-CoV-2 SAE models may be somewhat more dependent on covariates. In the case of SARS-CoV-2, due to the public nature of the spread, a relatively large range of covariates are candidates for use, such as population density, registered deaths and mobility data. Individuals with HIV or SARS-CoV-2 can be asymptomatic; with HIV, this is a state on the path to AIDS, which is associated with high infectivity, while with SARS-CoV-2, individuals can recover and the levels of infectivity (at time of writing) are not precisely known.

We will refer to the Malawi districts as admin-2 areas; there are 3 admin-1 areas, 28 admin-2 areas and 243 admin-3 areas—here, we are using the GADM (database of global administrative areas) classification (https://gadm.org/download_country_v3.html). In the 2015-2016 Malawi DHS, two-stage stratified cluster sample was implemented, with the sampling clusters (enumeration areas) being stratified by district and urban/rural. The Malawi Population and Housing Census, conducted in Malawi in 2008, provided the sampling frame for the survey. The sample for the 2015–2016 Malawi DHS was designed to provide estimates of key indicators for the country as a whole, for urban and rural areas separately and for each of the 28 districts. The sampling frame contained 12 558 clusters and our analyses use data from 827 sampled clusters (the supplementary materials give more details). In the 2015–2016 DHS survey for Malawi, 8 497 women in the age range 15–49 were eligible for HIV testing, and 93% of them were tested. HIV prevalence data were obtained from voluntarily taken blood samples from survey respondents. Testing is anonymous, and as such, respondents cannot receive their test results directly from DHS. Instead, they are given educational materials and referrals to voluntary, free counselling and testing. Blood samples are collected on filter paper via finger pricks before being sent to a laboratory for testing (Malawi DHS, 2016).

Urban clusters were oversampled, relative to rural clusters (details are in the supplementary materials). We do not include the island district of Likoma in analyses, because it is spatially distinct (and quite culturally different from the nearest points on the mainland) and has a very small population. In the top left panel of Figure 1, we plot the weighted prevalence estimates at the district level and see great variation across the country and what appears to be an increasing gradient in HIV prevalence from north to south. However, the uncertainty (shown in the supplementary materials) is relatively large, particularly in the south. Ignoring the survey design gives a national prevalence estimate (standard error) of 6.28% (0.37%) for women aged 15–29, while the weighted estimate is 6.18% (0.48%). The weighted estimate is smaller because the weights correct for the urban oversampling (HIV is more prevalent in urban areas) and the

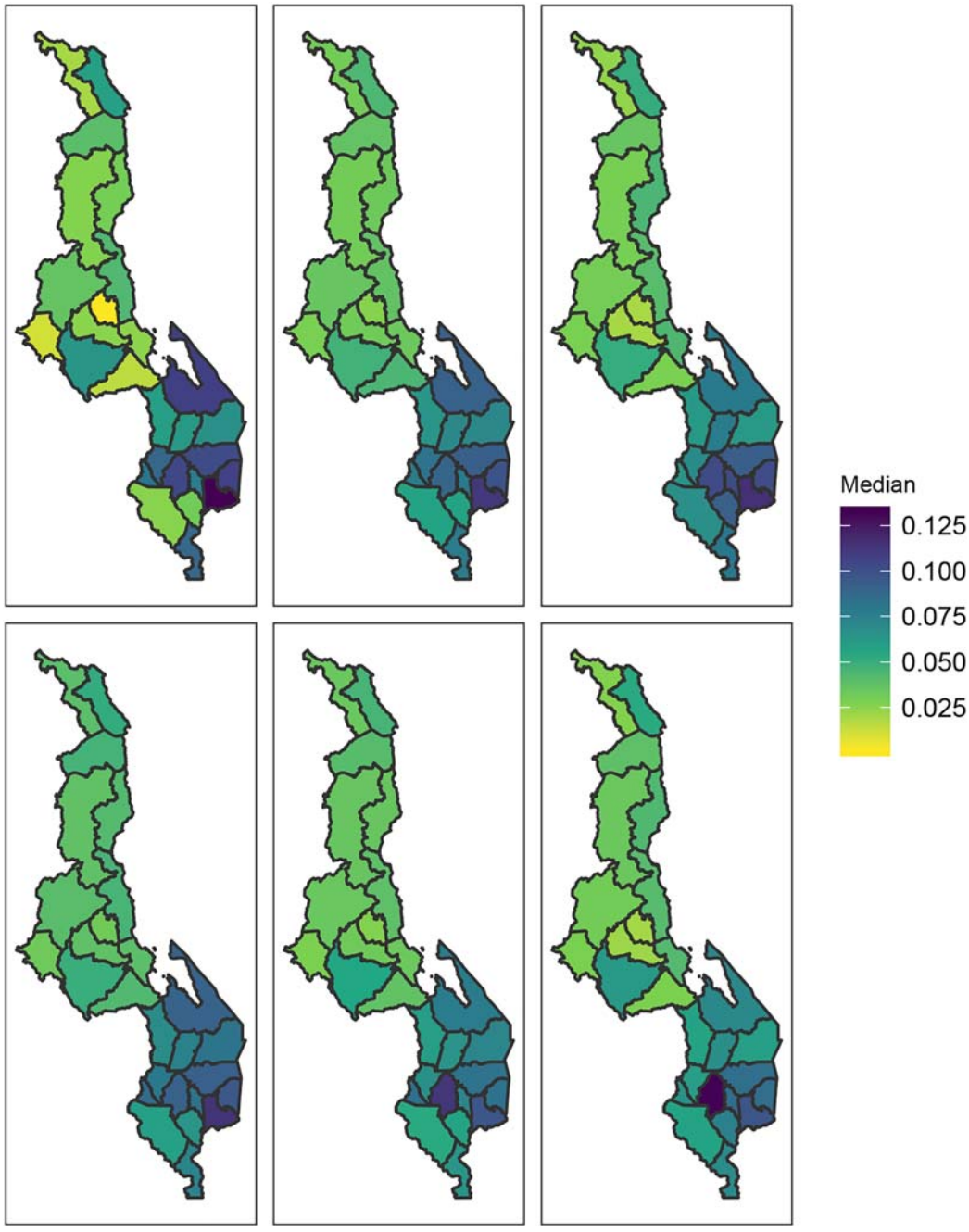


**Figure 1.** *Estimates of HIV prevalence among women aged 15–29 in districts of Malawi in 2015–2016. Top row estimates are from area-level models: direct estimates; smoothed direct estimates; and smooth direct estimates with antenatal care (ANC) HIV prevalence covariate. Bottom row estimates are from unit-level models: no urban/rural adjustment and no covariate; urban/rural adjustment only; and urban/rural adjustment and ANC HIV prevalence covariate. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

increased standard error is due to the variance calculation acknowledging the clustering of sampled women.

The structure of this article is as follows. In Section 2, notation and basic ideas are presented before traditional methods of direct and indirect estimation are described in Section 3. We focus on spatial methods that have not been covered in recent SAE reviews in Section 4, with both area-level and unit-level models being described. Models of each kind are applied to the HIV prevalence data in Section 5. The potential application of SAE methods to COVID-19 modelling is discussed in Section 6, and the paper concludes with a discussion in Section 7.

## 2 Notation and Overview

We let $i$ index the areal units for which estimation is required, with $m$ areas in total. Assume there are $N_i$ individuals in the population with responses $y_{ik}$, $k = 1, \ldots, N_i$, in area $i$. We let $S_i$ represent the set of indices of the selected individuals in area $i$, with $n_i = |S_i|$ the number sampled in area $i$. From a design-based perspective, $n_i$ may or may not be random, depending on the survey plan—for example, if the small areas (domains) correspond to the strata in a stratified design (sometimes called *primary domains*), then the $n_i$ will be non-random. It is more typical for this not to be the case, and in this situation, we have *secondary domains*. Under a model-based view, $n_i$ is conditioned upon and therefore fixed. We take as target of inference $m_i = \frac{1}{N_i}\sum_{k=1}^{N_i} y_{ik}$, the empirical mean of the response in the finite population. As examples of binary outcomes, interest may focus on the proportions of a population in an area that have a disease, have been vaccinated or who are practicing social distancing (a much softer endpoint). This is a finite population characteristic, and such targets are common in survey sampling. Another common target is the total $\sum_{k=1}^{N_i} y_{ik}$. In the model-based approach, the hypothetical infinite population mean in area $i$ is denoted $p_i$ (because our article focuses on prevalence, our notation suggests a probability). This can be interpreted as the mean of the distribution from which the finite population was (hypothetically) drawn.

We will focus on household surveys that are extremely popular, both in high-income countries (examples in the United States include the National Crime Victimization Survey and the American Community Survey) and in LMIC (e.g. the DHS and Multiple Indicator Cluster Surveys). Sampling of respondents for household surveys based on face-to-face interviewing is usually conducted via stratified multistage cluster sampling. A sample of primary sampling units (PSUs) is drawn from what is typically an area-based sampling frame, such as enumeration areas in a population census. Most surveys employ stratification, usually by residence (urban and rural) and administrative units. Because stratification usually has very low cost and somewhat reduces the variance of estimates, there is usually no reason to avoid stratification. From the point of view of estimating totals across the whole sample, an allocation of sampling units to strata that is proportional to the population size is preferable. Optimal allocation (Lohr, 2019, Section 3.4.2), where allocation of sampling units is proportional to the standard deviation of the estimator and size of the stratum, is preferable if the variances of the target estimator in each stratum are known and there is a single target estimator. In the case of most household surveys, the aim is to provide estimates of many different population quantities that all have different variances. Therefore, optimal allocation is not possible. Proportional allocation ensures that the variances will not be worse than if the sample was by SRS. However, when the $n_h$ are chosen, many surveys are disproportionally sampled across strata because of the need

to report on particular domains with similar precision, such as reporting by urban and rural or individual provinces.

Given the $n_h$ (the number of clusters to sample in strata $h$), nearly all household surveys select the $n_h$ PSUs by probability proportional to size (PPS), where the size measure is the number of households in each PSU, as recorded in the sampling frame. A linear systematic PPS sample (Murthy & Rao, 1988, Section 8) can be obtained if the sampling units are arranged contiguously along the real number line, each unit occupying a space equal to its size $N_i$. Let $T_i = T_{i-1} + N_i$, $i = 1, \ldots, m$, with $T_0 = 0$ and the sampling interval be $t = T_m/n$. Then, select a random point $r$ from 1 to $t$ and select unit $i$ if

$$T_{i-1} < r + jt \leq T_i, \qquad j = 0, 1, \ldots, n-1.$$

If the size measure is constant for all sampling units, the sample becomes a linear systematic SRS.

In a two-stage sample, which is the most common design, a list of households in each selected PSU is constructed, usually by mapping the cluster and listing every household by visiting every dwelling. A fixed number of households are then drawn using SRS or linear systematic SRS. This method has the benefit that the total sample size, both in total and in each cluster, is fixed by design. The fixed sample sizes thus reduce variance and ease the planning of the fieldwork. A key element in the design-based approach to inference are the *design weights*, which are the reciprocals of the *inclusion probabilities*. We now derive these probabilities for a stratified two-stage cluster design. Suppose there are $N_h$ households in strata $h$ with $N_{hc}$ being the number of households in cluster $c$, $c = 1, \ldots, C_h$, $h = 1, \ldots, H$. Hence, $H$ is the number of strata, and there are $C_h$ clusters in strata $h$. Power considerations lead to choices for the sample sizes $n_h$, of clusters to sample in strata $h$ using PPS sampling, so that the first stage inclusion probability for cluster $c$ in strata $h$ is $\pi_{1hc} = n_h N_{hc}/N_h$. Consequently, every cluster has a non-zero probability of being sampled, which, all things being equal, ensures that the weighted estimator (defined in Equation 2 in Section 3.1) is design unbiased, that is, unbiased when averaging over all possible samples that could be taken. The number of households to sample in cluster $c$ of strata $h$ is $n_{hc}$. The second stage inclusion probabilities for sampling are therefore $\pi_{2hc} = n_{hc}/N_{hc}$. The overall inclusion probability of cluster $c$ in strata $h$ is

$$\pi_{hc} = \pi_{1hc} \times \pi_{2hc} = \frac{n_h N_{hc}}{N_h} \times \frac{n_{hc}}{N_{hc}} = \frac{n_h n_{hc}}{N_h}. \tag{1}$$

If the number of households sampled from each cluster at the second stage is constant within strata, that is, $n_{hc} = m_h^\star$, then we write $m_h = n_h \times m_h^\star$ and obtain $\pi_{hc} = m_h/N_h$, independent of $c$. Thus, inclusion probabilities are equal within strata (self-weighting), provided that the size measure (number of households) is used for each PSU in the first and second stages, so that the $N_{hc}$ in (1) cancel. In sampling with more than two stages, for example, in the case when PSUs are districts, second stage units are enumeration areas, and third stage units are households—the second stage units are again selected by PPS to achieve equal probabilities of selection. Multistage sampling tends to increase variance substantially, because most of the variance accrues at the first stage of selection (because typically, a small number of first stage units are sampled). Therefore, multistage sampling is usually avoided if possible. For SAE, surveys that use more than two stages are not ideal, because the sample becomes concentrated in the relatively few geographic areas that have been selected in the first stage.

In stages selected with PPS, surveys in developing countries usually employ linear systematic PPS. This method has the benefit of being simple and allows for implicit stratification if desired. Implicit stratification involves sorting the sampling frame by some variable, usually geographic

location. If the selection is systematic, it is likely that adjacent selections will be more similar to each other than to non-adjacent selections. Pairs of adjacent clusters are therefore defined as implicit strata. The DHS used implicit stratification for early surveys but have discontinued this practice.

A great amount of energy is expended on calculating the variances of estimators. The drawback of linear systematic PPS is that the resulting selections of PSUs are dependent on each other. The lack of independence precludes the use of some variance estimators, such as the Yates–Grundy–Sen estimator (Yates & Grundy, 1953), that require the joint inclusion probabilities of PSUs, which cannot be defined, because all pairs of clusters no longer have non-zero probability of being selected. Therefore, for variance calculation purposes, it is generally assumed that PSUs were sampled with replacement, even though this is not the case under linear systematic PPS. This assumption leads to a generally insignificant inflated variance estimate. Alternative PPS selection methods do exist (see Brewer & Hanif, 2013) but are seldom used even though statistical suites such as SPSS and SAS provide them, as do several R packages. Surveys in developed countries often employ different designs because more information is available to the sampler.

Design-based weights express the contribution of each selected sampling unit to the estimate of a population parameter. The starting point for weight construction is the design weights. They are the inverse of the product of the inclusion probabilities at each stage of sampling. The design weights thus directly reflect the sampling process. Imperfections, such as non-response, mar most surveys. A common way to deal with non-response is to adjust the weights, by upweighting sampling units that are assumed to be similar to the non-responding sampling unit. The adjustment is typically carried out using neighbouring units, for example, a group of geographically adjacent survey clusters, as adjustment cells. The estimation weight is then the product of the design weight and the inverse of the non-response rate for the cell. An alternative is predicting the response propensity of each sampling unit using, for example, logistic regression. The estimation weight then becomes the product of the design weight and the inverse of the predicted response probability for each sampling unit. Methods for adjusting for non-response are described in Lohr (2019, Chapter 8).

Estimates of totals from sample surveys may also differ from known totals. For example, the age and gender distribution from a survey may be different from reliable population registration. To correct for the difference, post-stratification, raking and calibration are sometimes used to adjust the weights so that the survey estimates match the known totals. Post-stratification divides the population into groups, for example, by age and sex, and adjusts the weights within each group, so that known population totals are recovered. Raking does the same adjustment but to the marginals of the table formed by the classification variables. Calibration is similar but uses continuous variables instead of discrete. Post-stratification, raking and calibration ideally impart information to the sample and therefore reduce bias. These methods are commonly used in countries with good sources of information (the DHS do not use these procedures). They are sometimes used when the analyst believes that some sources, such as projections from a population census, are better than the survey estimates, but that is a practice that easily goes awry.

In general, weight variation in household surveys increases variance by a factor of $1 + \mathrm{CV}^2$, where CV is the coefficient of variation of the weights (Kish, 1992). To avoid the variance penalty from weight variation, most household surveys strive for equal probability of selection, at least within each stratum, and therefore constant weights within strata (as we saw with self-weighting). Nevertheless, there may be good reasons to accept unequal weights, such as representing particular domains or optimising for a single estimator. Equal weights within strata are difficult to achieve in practice. For SAE, the equal probability of selection method results

in high variance in low population density areas. One might think that a design that spreads the sample equally across a country would solve that problem. However, it would allocate too much of the sample to where few people live, and too little to population-dense areas, making the estimates worse overall.

Survey weight variation comes from three principal sources. The first source is disproportional allocation of sampling units to strata, which is a part of the design itself. The second source is differences that arise because of differences between estimates of the number of sampling units in one stage of the sampling process and subsequent ones. For example, in a typical two-stage cluster sample where the clusters are selected with PPS, inclusion probabilities will be equal for all households in a stratum if the number of households in each cluster used for selection in the first stage is the same as the number of households found in the listing of the clusters. Because the first stage numbers come from the sampling frame, which may be somewhat outdated, and the listing is carried out immediately before the survey, the difference between the two numbers is often large and leads to substantial weight variation. Third, weight variation may arise from non-response corrections, post-stratification, raking and calibration. Because the weight variation increases the variance, large weights are truncated in many surveys. The truncation trades the cost of a (hopefully) small bias for an often substantial reduction in variance and therefore also in the mean squared error (MSE).

As a result, the final data (including the weights) are the product of several random processes (Pfeffermann, 2011). The first is the generation of population units. The second is the selection resulting from the application of the sampling design. The third is the selection of responding units, given that they have been included in the sample. Finally, there is usually an ad hoc modelling step that is used to force sample statistics to resemble population parameters and reduce variance.

## 3 Traditional Methods

### *3.1 Direct Estimation*

A direct estimate in a specific area only uses response data on the variable of interest from that area. For simplicity of explanation, we assume that the strata correspond to the small areas of interest and that the weights are simply the reciprocal of the inclusion probabilities. We will also not explicitly index the clusters in this section, because the discussion is relevant for general designs. Let $d_{ik}$ be the design weight associated with individual $k$ in area $i$, whose response is $y_{ik}$. Within area $i$, the design-based weighted (direct) estimator (Horvitz & Thompson, 1952; Hájek, 1971) is

$$\widehat{m}_i^{\mathrm{HT}} = \frac{\sum_{k \in S_i} d_{ik} Y_{ik}}{\sum_{k \in S_i} d_{ik}}. \tag{2}$$

In many instances, a useful interpretation for a binary response is that the numerator is estimating the total number of events in the area, while the denominator is estimating the population size (unfortunately, this view is less useful in the DHS, because the weights are normalised to sum to the sample size). The variance of the estimator, $V_i^{\star}$, may be calculated using standard methods. Variances for estimates from complex surveys do not have exact formulas and are usually computed using either linearisation or replication methods (Wolter, 2007). The variance of the Horvitz–Thompson estimator can easily be computed using linearisation methods, provided that information on the weights and the sample structure is available. Estimators from health surveys, such as mortality rates, often have a complex structure, which makes standard

linearisation methods of variance correction difficult, however. Therefore, Jackknife methods are commonly used (Pedersen & Liu, 2012). Some public use surveys, such as the US National Health and Nutrition Examination Survey, hide design variables for confidentiality reasons and instead publish a set of weight variables that can be used for Jackknife estimates. For SAE, the design variables used in the sampling process are needed.

A starting point for SAE analysis is to map the weighted estimates. These weighted estimates have excellent properties (so long as the weights are reliable and stable), for example, design consistency. If abundant data are available, no further modelling is needed. In general, the variance is $O(n_i^{-1})$, and so, for small $n_i$, this approach will not be reliable. In this case, models are necessary to allow for incorporation of covariate information and/or leveraging of spatial dependence between areas.

### *3.2 Indirect Estimation*

We briefly describe a number of traditional indirect estimators that are constructed to reduce the MSE, calculated under the randomisation distribution. These estimators are generally design based, but achieve a favourable MSE through variance reduction. Suppose we have covariates $\boldsymbol{x}_{ik}$ that are available on the complete population $k = 1, \ldots, N_i$, $i = 1, \ldots, m$. In a conventional regression analysis, attention focuses on the slope coefficients, but in SAE, we are interested in the empirical population means $m_i$, and so, it makes sense that we would want to make inferences about the responses of unobserved individuals. A *synthetic estimator* is

$$\hat{m}_i^{\text{SYN}} = \frac{1}{N_i}\sum_{k=1}^{N_i} \boldsymbol{x}_{ik}^{\text{T}}\widehat{\boldsymbol{B}}, \tag{3}$$

where

$$\widehat{\boldsymbol{B}} = \left[\sum_{i=1}^{m}\sum_{k\in S_i} d_{ik}\boldsymbol{x}_{ik}^{\text{T}}\boldsymbol{x}_{ik}\right]^{-1}\sum_{i=1}^{m}\sum_{k\in S_i} d_{ik}\boldsymbol{x}_{ik}^{\text{T}}y_{ik} \tag{4}$$

are the slope estimates derived from the set of samples across *all* areas. The success of this estimator depends on how appropriate the regression model is for all areas, although this may be relaxed to have common slopes for a collection of contiguous areas, rather than *all* areas. The variance of the estimator given by (3) and (4) will be $O(1/n)$, where $n = \sum_{i=1}^{m} n_i$ is the total sample size, but there is the possibility of large bias. Fitting separate models in *each* area is appealing, but then, the small $n_i$ problem re-emerges. As a side note, a possible compromise approach that we are investigating is to use spatially varying coefficient models in which each area has its own slope, but these slopes are smoothed across areas using a random effects model. In spite of the potential for large bias, synthetic estimators are popular, in part because they are intuitive and can be used for areas in which no response data were collected.

To deal with the potential large bias of (3), the bias may be estimated to give the *survey regression estimator*,

$$\begin{aligned}\hat{m}_i^{\text{SR}} &= \frac{1}{N_i}\sum_{k=1}^{N_i}\boldsymbol{x}_{ik}^{\text{T}}\widehat{\boldsymbol{B}} + \frac{1}{N_i}\sum_{k\in S_i} w_{ik}(y_{ik} - \boldsymbol{x}_{ik}^{\text{T}}\widehat{\boldsymbol{B}})\\ &= \hat{m}_i^{\text{HT}} + (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_i^{\text{HT}})^{\text{T}}\widehat{\boldsymbol{B}},\end{aligned}$$

where $\hat{m}_i^{\mathrm{HT}}$ and $\bar{\boldsymbol{x}}_i^{\mathrm{HT}}$ are the Horvitz–Thompson estimates of $m_i$ and $\bar{\boldsymbol{x}}_i$, respectively. This estimator is approximately design unbiased, but the variance is unfortunately back to $O(1/n_i)$. A *composite estimator* is of the form

$$\hat{m}_i^{\mathrm{COM}} = \delta_i \hat{m}_i^{\mathrm{SR}} + (1 - \delta_i)\hat{m}_i^{\mathrm{SYN}},$$

with $0 \leq \delta_i \leq 1$ estimated in such a way that for larger $n_i$, we have $\delta_i$ closer to 1. This estimator is intuitively appealing and when motivated by a random effects model leads to a principled approach to estimation of $\delta_i$.

## 4 Spatial Smoothing Models

### *4.1 Area-Level Models*

In this section and the next, we focus on binary outcomes, because this is a common situation for health and demographic responses and the Malawi HIV prevalence example falls under this umbrella. Spatial smoothing methods often use random effects modelling, and so the first hurdle to overcome when using such approaches in an SAE setting is how to acknowledge the survey design. In a major advance, Fay & Herriot (1979) introduced a very clever approach that models a transform of the weighted estimate, in order to gain precision by using a random effects model. For binary outcomes, one choice of transform is $Z_i = \mathrm{logit}\left(\widehat{m}_i^{\mathrm{HT}}\right)$. We denote the associated design-based variance by $V_i$. An *area-level model* is

$$Z_i \sim \mathrm{N}(\theta_i, V_i) \tag{5}$$

$$\theta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i + S_i, \tag{6}$$

where $\theta_i$ is the logit of the true proportion in area $i$. Area-specific deviations from the regression model are modelled using a pair of random effects. The independent and identically distributed (iid) terms are $e_i \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2)$ while the collection $\boldsymbol{S} = [S_1, \ldots, S_n]$ are assigned a spatial distribution. The Fay–Herriot model did not include the spatial random effects $\boldsymbol{S}$, but the iid random effects only. Choices for the spatial distribution are described in Banerjee *et al.* (2015), with common forms being the conditional autoregressive (CAR) and intrinsic CAR (ICAR) models. Both of these choices capture the concept that, in general, outcomes are likely to be similar in locations that are close by. Mercer *et al.* (2015) and Li *et al.* (2019) used these *smoothed direct* models in a space–time context, using spatial ICAR and temporal random walk components, along with a space–time interaction term. The inclusion of iid and ICAR random effects corresponds to the popular Besag, York, Mollié model introduced by Besag *et al.* (1991). Both the CAR and the ICAR choices are Markov random field models, which offer computational advantages (Rue & Held, 2005). Markov random field models may be fit using either a frequentist approach or a Bayesian approach in which priors are placed on the fixed effects (the $\boldsymbol{\beta}$s) and on the variances/spatial dependence parameters. Design-based consistency is achieved (so long as the priors do not assign zero mass to the true $\theta_i$) because the $V_i$ term will tend to zero and the bias due to the random effects smoothing disappears asymptotically.

Two practical difficulties with this approach are that the direct estimates may be on the boundary for a summary parameter that is not on the whole real line. For example, in the binary case, we may have $\widehat{m}_i^{\mathrm{HT}}$ equal to 0 or 1. In this case, $Z_i$ will be undefined. Further, a transform of the weighted estimator may not share the same design-based properties as the untransformed estimator, such as being design unbiased. These problems may be alleviated by using an unmatched

sampling and linking model (You & Rao, 2002). A second difficulty is that reliable variance estimates $V_i$ may be unavailable, particularly for areas with few samples. In this case, variance smoothing models can be used (Rao & Molina, 2015, Section 6.4.1).

### *4.2 Unit-Level Models*

In this section, we assume that the units of analysis are clusters within a multistage cluster design. We let $\boldsymbol{s}_{ic}$ represent the geographical location of cluster $c$ in area $i$ and explicitly index the counts and sample sizes as $Y_{ic}$ and $n_{ic}$, respectively, for $c = 1, \ldots, C_i$, $i = 1, \ldots, m$. A crucial assumption here (Rao & Molina, 2015, Section 4.3) is that the probability of selection, given covariates, does not depend on the values of the response (as discussed in the context of *ignorability* in Section 2). This implies that if stratified random sampling is used, stratification variables must be included in the model. One would expect cluster sampling to lead to correlated responses within clusters, and cluster-level random effects are introduced to accommodate this aspect (Scott & Smith, 1969). Another situation in which care is required is when PPS sampling is carried out. If the model used misspecifies the relationship between the response and size, then incorrect inference will result. To provide some protection against this, Zheng & Little (2003) and Chen *et al.* (2010) model the response as a spline function of size for continuous and binary outcomes, respectively.

For binary responses, a common model (Diggle & Giorgi, 2019) is

$$Y_{ic}|p_{ic} \sim \text{Binomial}(n_{ic}, p_{ic}). \tag{7}$$

Care must be taken to acknowledge the design, when modelling $p_{ic} = p(\boldsymbol{s}_{ic})$, the response probability associated with location $\boldsymbol{s}_{ic}$. For a stratified design, it is, in general, important to include fixed effects for each strata (Paige *et al.*, 2020). In practice, there are a number of options when we have a design with strata that consist of urban/rural crossed with geographical districts. If there were no random effects, then strictly, we would need interactions between district and urban/rural. However, the models we describe include spatial random effects (to leverage spatial dependence) and interactions would produce an unidentifiable model. Consequently, we include a single urban/rural effect and so are tacitly assuming that the association between the response and the urban/rural variable is approximately constant across areas. Another possibility would be to include a spatially varying urban/rural coefficient, or model the associations separately in larger regions, in the spirit of what is sometimes done with synthetic estimation.

One candidate model to accompany (7) is

$$p_{ic} = \text{expit}\left(\beta_0 + \boldsymbol{x}(\boldsymbol{s}_{ic})^{\mathrm{T}}\boldsymbol{\beta}_1 + z(\boldsymbol{s}_{ic})\gamma + S(\boldsymbol{s}_{ic}) + \epsilon_{ic}\right), \tag{8}$$

where $z(\boldsymbol{s}_{ic})$ represents the strata within which cluster $c$ lies, $\exp(\gamma)$ is the associated odds ratio and $\boldsymbol{x}(\boldsymbol{s}_{ic})$ are covariates available at location $\boldsymbol{s}_{ic}$, with odds ratios $\exp(\boldsymbol{\beta}_1)$. The spatial random effect $S(\boldsymbol{s}_{ic})$ is associated with cluster location $\boldsymbol{s}_{ic}$ and may be continuous or discrete. The cluster-level error $\epsilon_{ic} \sim \text{N}(0, \sigma^2_\epsilon)$ is the so-called *nugget*, which is traditionally vaguely specified as representing short-scale variation and/or 'measurement error'. Measurement error in a binary outcome would correspond to the misclassification of the response (e.g. by the respondent or the interviewer), and so, labelling the nugget as such is assuming that the cumulative misclassification of the samples can be approximately modelled by $\epsilon_{ic}$.

An MBG model takes $S(\boldsymbol{s}_{ic})$ as a realisation of a zero-mean Gaussian process. Gaussian process models are common choices for continuously indexed spatial models and imply that any collection of spatial random effects has a multivariate normal distribution. A popular choice

for the variance–covariance is the Matérn covariance function (Stein, 1999), for which the covariance is

$$\operatorname{cov}(S(\boldsymbol{s}_1), S(\boldsymbol{s}_2)) = \sigma_s^2 \frac{2^{1-\nu_s}}{\Gamma(\nu_s)} \left( \sqrt{8\nu_s} \frac{||\boldsymbol{s}_2 - \boldsymbol{s}_1||}{\rho_s} \right) \mathrm{K}_{\nu_s} \left( \sqrt{8\nu_s} \frac{||\boldsymbol{s}_2 - \boldsymbol{s}_1||}{\rho_s} \right),$$

where $\rho_s$ is the spatial range corresponding to the distance at which the correlation is approximately 0.1, $\sigma_s$ is the marginal standard deviation, $\nu_s$ is the smoothness (which is usually fixed, because it is difficult to estimate) and $\mathrm{K}_{\nu_s}$ is a modified Bessel function of the second kind, of order $\nu_s$. When the number of clusters $C = \sum_{i=1}^{m} C_i$ is large, computation is an issue, because we need to manipulate $C \times C$ matrices, which involves $O(C^3)$ operations (Rue & Held, 2005). Various approximations have been proposed to overcome this problem, for example, the stochastic partial differential equations (SPDEs) approach pioneered by Lindgren *et al.* (2011)—this is the approach we use in Section 5. Other approaches are described by Heaton *et al.* (2018).

Aggregation to the area level is carried out via

$$p_i = \int_{A_i} p(\boldsymbol{s}) \times q(\boldsymbol{s})\, d\boldsymbol{s} \quad \approx \quad \sum_{l=1}^{\mathrm{M}_i} p(\boldsymbol{s}_l) \times q(\boldsymbol{s}_l), \tag{9}$$

where $p(\boldsymbol{s}) = \operatorname{expit}(\beta_0 + \boldsymbol{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta}_1 + z(\boldsymbol{s}_{ic})\gamma + S(\boldsymbol{s}))$ is the risk at location $\boldsymbol{s}$ (the nugget is, for better or worse, frequently left out, because it is viewed as measurement error) and $q(\boldsymbol{s})$ is the population density at $\boldsymbol{s}$, which is needed at all locations on the approximating mesh, $\boldsymbol{s}_l$, $l = 1, \ldots, M_i$. Cluster-level covariate models are appealing when compared with area-level covariate alternatives, because they are closer to the mechanism of action and reduce the possibility of ecological bias (Wakefield, 2008) though if prediction is the aim this is less important. A large disadvantage in an SAE context is that to aggregate to the area level, at which inference is required, one needs to know the values of the covariates, including the design variables, on the mesh. Obtaining reliable urban/rural classifications, population density and covariate surfaces is not straightforward. The direct and smoothed direct approaches avoid the need for a population surface, because the weights implicitly include population information from the original sampling frame.

An alternative, overdispersed binomial, unit-level model that we use for the HIV prevalance data is

$$Y_{ic} \mid p_{ic}, \lambda \sim \text{BetaBinomial}(n_{ic}, p_{ic}, \lambda) \tag{10}$$

$$p_{ic} = \operatorname{expit}\left(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + z_{ic}\gamma + e_i + S_i\right), \tag{11}$$

where $\lambda$ is the overdispersion parameter and we have taken the spatial random effect to be decomposed as $S(\boldsymbol{s}_{ic}) = e_i + S_i$, with $e_i$ and $S_i$ as in (6). We also have covariates that are specified at the area level, rather than the cluster level. The marginal variance is $\operatorname{var}(Y_{ic}) = n_{ic}p_{ic}(1-p_{ic})[1 + (n_{ic}-1)\lambda]$ so that small values of $\lambda$ correspond to little overdispersion. The aggregate risk is far easier to evaluate when compared with the model with continuous spatial random effects and cluster-level covariates, that is calculated via (9), because all clusters in the same strata have identical risk. We still need to know $q_i$, the proportion of the population in the area that is urban, but this is preferable to needing a pointwise classification of urban/rural. The aggregate risk is therefore

$$p_i = (1-q_i) \times \operatorname{expit}(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + e_i + S_i) + q_i \times \operatorname{expit}(\beta_0 + \gamma + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + e_i + S_i). \tag{12}$$

Dyed-in-the-wool,design-based aficionados are wary of models such as the ones described in this section, because establishing design consistency is very difficult.

## 5 Small Area Estimation for HIV Prevalence Mapping

We return to the HIV prevalence example and describe the fitting of both area-level and unit-level models. All fitting was done in the `R` programming environment. The smoothing models were fit using the `INLA` package, which provides Bayesian inference via a fast implementation using the method described in Rue *et al.* (2009) and Rue *et al.* (2017). To obtain the weighted estimates and variances for the areas, the `survey` package (Lumley, 2004) was used. A number of the models can be conveniently fitted in the `SUMMER` package (Martin *et al.*, 2018; Li *et al.*, 2020). For the discrete spatial model, we use the BYM2 parameterisation (Riebler *et al.*, 2016) in which we have $b_i = e_i + S_i$ (in Equation 6) and an overall variance parameter $\sigma_b^2$ for $b_i$ and a parameter, $\phi$, that represents the proportion of this variance that is spatial. In all analyses, we use penalised complexity priors (Simpson *et al.*, 2017), details of which may be found in the supplementary materials.

### *5.1 Area-Level Models*

In high-income countries, the census (and other sources) provide information on all of the population in small areas of interest. In this case, weighted estimates can be obtained from model-assisted procedures, such as those described in Section 3.2, and can be smoothed, if needed. Unfortunately, in LMIC, there are few candidate variables, and so, we instead use the smoothed direct model (5–6) and include as an area-level covariate the (logit of) HIV prevalence estimates obtained from antenatal care (ANC) facilities. The covariate corresponds to the HIV prevalence of pregnant women attending health facilities to receive ANC, aggregated to the district level. A map of this covariate is provided in the left panel of Figure 2. The specific data we use are from Malawi Ministry of Health quarterly integrated HIV programme reports in 2016. The right panel of Figure 2 shows the logits of the direct estimates of the HIV prevalence against the logits of the ANC prevalence estimates, with a reference line added. Not surprisingly, we see a strong association.

The smoothed direct model estimates without the ANC covariate are shown in the top middle panel of Figure 1. The shrinkage (overall via the iid random effects and locally via the ICAR random effects) is apparent, with a flatter map compared with the direct estimates. The posterior uncertainty is reduced, because the totality of data is used in the smoothing. The supplementary materials contain comparison plots of the estimates and their uncertainties and a table that summarises the parameter estimates.

When we add the ANC covariate to the smoothed direct model, the posterior median of $\sigma_b$ (which reflects the amount of residual variation on the logit scale) is reduced from 0.40 to 0.14, and $\phi$ (the proportion of spatial variation) goes from 0.56 to 0.26. The posterior median of the odds ratio (95% credible interval) associated with the logit ANC covariate is 2.7 (1.8, 4.0) so the association is very strong, which explains why the across-district residual variability is so reduced when the covariate is added. The strong spatial structure of the covariate explains why $\phi$ is reduced. The dream of spatial modelling is to find covariates that make the area-level random effects 'go away'. The top right panel of Figure 1 shows the estimates from the ANC covariate smoothed direct model. The pattern of lower prevalence in the north, with higher levels in the south, is still apparent in the smoothed estimates. The average standard error of the

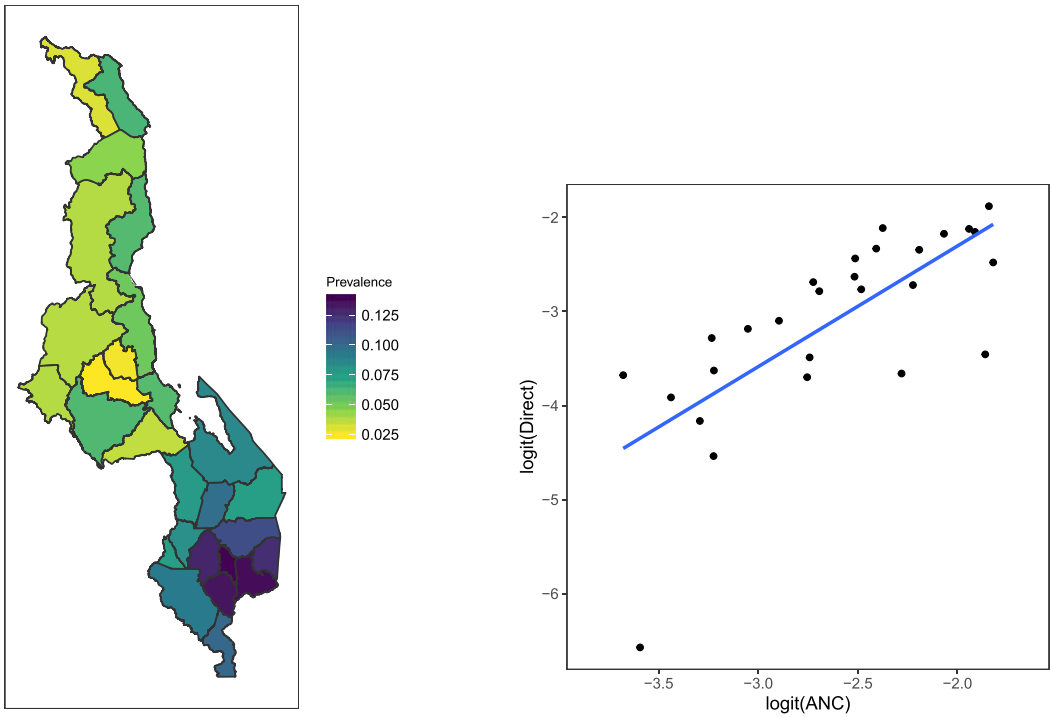


**Figure 2.** *Left: map of ANC HIV prevalence for districts of Malawi. Right: logit of direct prevalence estimates versus logit of antenatal care (ANC) prevalence estimates—we see a strong association. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

direct estimates is 0.018 and the average posterior standard deviations of the smoothed direct area estimates are 0.014 (no covariate) and 0.010 (ANC covariate), illustrating the benefits of spatial smoothing and the addition of covariates. Even with the spatial smoothing and use of the covariate, the uncertainty in the ranking of areas is still large. The supplementary materials contain distributions on the rankings under six different models. Mulanje consistently comes out as the district with the highest HIV prevalence among women aged 15–29.

## *5.2 Unit-Level Models*

We fit three betabinomial models (10–11) with area-level random effects: (A) no urban/rural stratification and no covariates, (B) urban/rural stratification and no covariates, and (C) urban/rural stratification and the ANC covariate. Model (A) should not be used, as it does not account for the urban/rural stratification; we include it to demonstrate the bias this introduces. Models (B)–(C) produce two prevalence estimates for each of the 27 admin-2 areas, one for urban women and one for rural women. To produce the admin-2 estimates, we need to weight the two estimates, via (12), by the proportions of women aged 15–29 who live in rural and urban areas. From the report (Malawi DHS, 2016), we know the number of households, by admin-2 area, that are urban and rural. The report also states that the average number of women aged 15–49 per household is 0.94 in rural areas and 1.09 in urban areas. We scale the household numbers by these values to approximate the proportions of women by rural/urban in each admin-2

area. The bottom row of Figure 1 shows the subsequent fits from the three models—the fits are quite similar to the smoothed direct models. Closer examination reveals differences, however. In the no covariate model, the odds ratio associated with being urban is 2.3 (1.8, 2.9), and this hardly changes when the ANC variable is added. This more than doubling of the odds is a cluster-level association. The supplementary materials give comparisons of all the estimates. It is clear that leaving the urban/rural covariate out of the model leads to bias. A figure in the supplementary materials shows that estimates from the model without stratification are generally higher than those with stratification, because of the oversampling of urban areas, within which HIV prevalence is higher.

The odds ratio associated with the ANC covariate is 2.3 (1.7, 3.0), so the association is also strong. The standard deviations of the total variation, $\sigma_b$, were estimated as 0.37, 0.36 and 0.14 under models (A), (B) and (C), respectively, while the proportions of the variance that are spatial, $\phi$, were estimated as 0.62, 0.71 and 0.22. Hence, we see a large reduction in both total residual variation and proportion spatial when the ANC covariate is added. In terms of the area-level averages, recall that the average standard error of the direct estimates is 0.018. The average posterior standard deviations of the admin-2 averages, under unit-level models (A), (B) and (C), are 0.014, 0.012 and 0.010, showing the benefits of the smoothing in space and via the covariate.

We also calculate estimates for the 243 admin-3 areas, which is more difficult with the smoothed direct model, because of data sparsity (26 of the areas contain no data). The supplementary materials contain summaries of this analysis. These results should be viewed with some caution, because model checking at this level is very difficult and the aggregation depends on knowing the proportion of each admin-3 area that is urban. For the latter, we use information from the DHS report and WorldPop population density estimates (Stevens *et al.*, 2015; Wardrop *et al.*, 2018). Specifically, we threshold the WorldPop population density surface at the admin-2 level to get a proportion of the population in urban and rural areas that agrees with the proportions from the DHS report. We then apply these admin-2 specific thresholds to the constituent admin-3 areas, to get the proportions of each admin-3 area that are urban/rural.

Finally, we fit three betabinomial MBG models (i.e. models with Gaussian process spatial random effects and a Matérn covariance function), using the SPDE method. Details are in the supplementary materials. The three models are: without an urban/rural fixed effect or the ANC covariate, with an urban/rural fixed effect but without the ANC covariate and with both the urban/rural fixed effect and the ANC covariate. Because of the strong north–south and east–west trends in prevalence, all models also included linear terms in latitude and longitude. Care must be taken with prior specification for the range parameter, because we want the spatial model to pick up spatial dependence at shorter distances than is being modelled with the latitude and longitude contributions.

The results are provided in the supplementary materials and should be viewed with caution, particularly for the admin-2 summaries, because the aggregation required a mesh-specific estimate of urban/rural, to use in (9). We use a similar procedure to the previous, but assign each pixel in each admin-2 area to urban/rural based on the admin-2 population density thresholds. There are a number of outstanding issues with the unit-level model-based analyses that need further consideration. In general, the weights have three components—design, non-response and post-stratification/raking—and one needs to consider all three. We have discussed the stratification aspect, but further thought is required for non-response. Post-stratification from a modelling perspective has been considered (Gelman, 2007). In practice, the urban/rural covariate is not known with any great geographical accuracy, and the impact of this aspect on inference needs to be investigated.

### 5.3 Model Assessment

One of the hardest parts of model-based approaches to SAE is assessment of model assumptions. A cross-validation strategy is to systematically remove one area at a time and then obtain a prediction of the missing area's (logit of the) direct prevalence estimate, based on the remaining areas. The asymptotic distribution of this direct estimate is $\text{logit}(\hat{m}_i^{\text{HT}}) \sim \text{N}(\text{logit}(m_i), V_i)$. We simulate samples from the approximation to the posterior of $\text{logit}(p_i)$ that is provided by INLA (Rue *et al.*, 2017) and add iid $\text{N}(0, V_i)$ errors to each sample. The result is the predictive distribution of what the model thinks the direct estimate will be in the area for which the data were removed. We then plot representations of these 27 predictive distributions and compare with the observed points $\text{logit}(\hat{m}_i^{\text{HT}})$. The supplementary materials show these plots for the two smoothed direct models and three betabinomial models, with interval widths of 50% and 80%. The smoothed direct models have more severe undercoverage than the betabinomial models. The three betabinomial models are difficult to distinguish via this exercise, but the interval widths are narrowest for the model with a fixed effect for urban/rural and the ANC covariate. Checking the admin-3 estimates is far more difficult because there are few reliable direct estimates for comparison in a cross-validation exercise.

## 6 Application of Small Area Estimation Techniques to Coronavirus Disease 2019 Modelling

COVID-19 is an infectious disease caused by the SARS-CoV-2 virus. There are many endpoints that are of interest in geographical studies including disease prevalence, disease incidence, the number and fraction of people who have been infected with SARS-CoV-2 (via an antibody test), measures of social distancing and excess mortality. Estimating each of these responses over time and by age, gender, race and occupation is also highly desirable. Many study/data types would not fall under our definition of SAE because they are based on routine surveillance data and are nominally a complete enumeration. Surveillance data are often examined via space–time *disease mapping* or *cluster detection* techniques (Waller & Gotway, 2004).

Seroprevalence studies, in which blood samples are taken from sampled individuals and examined for antibodies, have been carried out to estimate the number of infections due to SARS-CoV-2. With an appropriate sampling design and a reliable test, such studies can be used to estimate the proportion of the population, by demographic groups, who have been previously infected by SARS-CoV-2. These studies can be used for SAE, if the numbers sampled are sufficiently large. Stringhini *et al.* (2020) describe a seroprevalence study carried out in Geneva, Switzerland, with participants taken from an existing study that contained a stratified sample from the study population (the canton of Geneva). The selection of individuals is key because, for example, those reporting at health facilities (and are subsequently tested) are not representative of the population. Unfortunately, many of the early seroprevalence studies came (justifiably) under fire for data quality issues and for sampling from ill-defined populations, which leads to great difficulties with interpretation. This is well documented by Vogel (2020), who gives a number of examples. Health researchers are under intense pressure, given that they are working in the context of the political and press feeding frenzy for results to be made available, which has lead to inadequate validation of the antibody tests used and limited scientific scrutiny of the design of studies through peer review.

It has been documented that certain subgroups (e.g. based on age, sex, race) are disproportionately affected by COVID-19, and as such, a stratified analysis may be used to reliably

estimate outcomes of interest. But the numbers required will be large if a small area analysis is envisaged, given the relative rarity of death from COVID-19, in particular.

A potential drawback with traditional survey sampling and SAE in the context of COVID-19 is the possibility that seroprevalence or incidence is highly clustered. There are two notions of clustering that are important here. The first is infectious network clustering, that is, that an infected person infects others in his or her network. Such clustering, while intrinsic to any infectious disease, may or may not lead to the other form of clustering, namely, geographic clustering. Geographic clustering of COVID-19 is common, both in communities and in institutions (there are many factors that may lead to this, including the propensity for individuals of common demographic groups, or having high-risk occupations, tending to live in close proximity). When geographic clustering is present, a large proportion of the infected population may be located in relatively few survey clusters, many of which may be geographically adjacent. Because a typical cluster sample design spreads out the sample clusters over a geographic region, there is a large chance that the sample will not pick up the high-prevalence areas. One may solve the problem by increasing the sample size, but that rapidly becomes costly.

An alternative is adaptive cluster sampling (Thompson, 1992; Thompson & Seber, 1996). An ordinary cluster sample is first drawn. Then, each selected cluster is listed and target individuals identified (e.g. persons with a positive test for Sars-CoV-2). If the number of target individuals is larger than a given threshold, then clusters that share a border with the initial cluster are included in the survey. The process is repeated for each of the adjacent clusters, until a network of clusters is formed that has no adjacent cluster that satisfies the threshold.

If the threshold is well selected, then the adaptive cluster sample will pick up geographically based networks of infection well. If it is set too low, then all clusters in the frame will be included. If the threshold is too high, the sample limits itself to the initial sample. With a target population that is clustered and judicious setting of the threshold value, adaptive cluster sampling can be very efficient. Unfortunately, there is no optimal sampling strategy for adaptive cluster sampling (Turk & Borkowski, 2005; Thompson & Seber, 1996), so the exact choice of design must be piloted.

A potential disadvantage of adaptive cluster sampling is that while it may be efficient for the estimator that informs the design, such as the prevalence of a disease, it is not necessarily efficient for estimators of other population quantities. Thus, if the survey aims to estimate additional quantities other than the one designed for, one might have to consider other designs. This problem is, to some extent, mitigated by the fact that many other quantities may be both easy to estimate and not require narrow confidence bands. For example, the proportion of people using face masks may not require as precise an estimate as a seroprevalence estimate.

Looking into the future, given the successful development of a vaccine, one could use SAE techniques to estimate vaccination coverage rates. Such an approach has been used in many different settings: for bacille Calmette–Guérin; diphtheria, pertussis and tetanus; oral polio; and measles vaccines in India (Pramanik *et al.*, 2015) and diphtheria, pertussis and tetanus in Africa (Mosser *et al.*, 2019). In the United States, in the absence of a national register, particular outcomes can be geographically examined using survey data. Albright *et al.* (2019) carried out SAE at the county level in Alabama for human papillomavirus using data from the National Immunization Survey.

To get at transmission dynamics, one needs to fit more complex models such as susceptible–exposed–infected–removed models. Unfortunately, there are a number of serious drawbacks for using such models in an SAE enterprise. The first is that the data required are not routinely available, with the usual problems of underreporting and misreporting being a serious problem. The second is that fitting stochastic models is very challenging (see the references in Fintzi *et al.*, 2017 for details). To avoid the computational difficulties, one may use

deterministic compartmental models (Anderson & May, 1991) and then map summary statistics such as the reproductive number $R_0$. In large populations, such models may work well, but such models may perform poorly when the system is far from its deterministic limit (Andersson & Britton, 2000). Inference is also difficult because error terms are 'added on' rather than having the stochastic and deterministic parts of the models be entwined, as is the ideal for a bona fide statistical model. However, compartmental models are by far preferable to simple curve-fitting exercises in which a generic function is fit to the data—such approaches are likely to perform very poorly for predictions of complex phenomena because they (basically) contain little of the science or relevant context. The impact of interventions, such as social distancing and vaccination campaigns, is also impossible to determine with such approaches.

## 7 Discussion

SAE is a huge topic, and we have been selective in this review, focusing on the use of Bayesian spatial models for analysing survey data. There is a disconnect between the spatial statistics and SAE research communities. The former are very cavalier about ignoring the survey design and the latter tend to use simple discrete spatial models (or include iid spatial random effects only) and empirical Bayes estimation. Regular spatial modellers who analyse binary data do not emphasise consistency, perhaps because the parameters of non-linear random effects models are not generally consistent, unless the model is correctly specified. Outside of the linear model, results on frequentist (model-based) consistency for the parameters of spatial models are few and far between.

It is desirable that when area-level estimates are aggregated, they are in agreement with estimates for larger areas within which they are nested. Such *benchmarking* is an important endeavour in many applications. For example, Li *et al.* (2019) produced small area estimates of under-5 mortality in 35 countries using DHS data and benchmarked to the official United Nations national estimates, which use far more extensive data sources. The estimates in larger areas are better estimated due to larger sample sizes and additional data. Many approaches to benchmarking have been suggested. In particular, we find the Bayesian method of Zhang & Bryant (2020) very appealing, and it could be used with the area-level and unit-level models we have described.

We have discussed both discrete and continuous spatial models. The former have dominated the SAE literature and have many appealing features including computational efficiency, a wealth of experience in their use and robustness to different data-generating mechanisms (Paige *et al.*, 2020). However, the discrete spatial models always have an ad hoc neighbourhood specification, which is unfortunate. Continuous spatial models are far more appealing in this respect and allow data that are aggregated to different levels to be combined (Wilson & Wakefield, 2020). But continuous spatial models have greater computational challenges, and careful prior specification is required (Fuglstad *et al.*, 2019). There is also the temptation when viewing a pixel-level map to believe that you know the value of the outcome of interest at a fine scale, when often there is huge uncertainty associated with estimation at the pixel level. If pixel maps are displayed, they should be accompanied by a map of uncertainty. Different methods for showing uncertainty are described in Dong & Wakefield (2020). We prefer to view the continuous spatial priors as a way of inducing spatial dependence between areas. Local hotspots of disease are unlikely to be apparent when spatial smoothing models are used, unless the number of excess cases is large, because spatial models smooth to the risk in surrounding locations. This is why local covariates are important for modelling a continuous risk surface. Area-level models do not pretend to pick up small-scale variation, because it is an area-average spatial parameter

that is being modelled, which is why area-based maps should be interpreted ecologically—the risk estimate is not associated with any one geographical location, but is an average, and within an area, there is heterogeneity.

There is a growing SAE literature on accounting for errors in the covariates used for modelling, with various errors-in-variables models being considered (Ybarra & Lohr, 2008, Barber *et al.*, 2016, Burgard *et al.*, 2019); see Section 6.4.4 of Rao and Molina, for a review.

Finally, as more and more data streams become available, there is a growing need for methods that combine data in appropriate ways, and this is especially true of SAE. For example, in an LMIC context, data may be available from a variety of surveys with different designs, sample vital registration systems and censuses. Synthesising such data is not straightforward but can offer great benefits, if carried out successfully.

## Acknowledgements

The authors would like to thank the Malawi Ministry of Health for allowing the use of the ANC data, Jeff Eaton for helpful comments on the manuscript and Geir-Arne Fuglstad for assistance with the SPDE modelling.

## References

Albright, D.L., Lee, H.Y., McDaniel, J.T., Kroner, D., Davis, J., Godfrey, K. & Li, Q. (2019). Small area estimation of human papillomavirus vaccination coverage among school-age children in Alabama counties. *Public Health*, **177**, 120–127.

Anderson, R.M. & May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press: Oxford.

Andersson, H. & Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer Lecture Notes: New York.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition. CRC Press: New York.

Barber, X., Conesa, D., Lladosa, S. & López-Quílez, A. (2016). Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospatial Health*, **11**, 11–20.

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

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

Brewer, K.R.W. & Hanif, M. (2013). *Sampling with Unequal Probabilities*, Vol. **15**. Springer Science and Business Media: New York, NY.

Burgard, J.P., Esteban, M.D., Morales, D. & Pérez, A. (2019). A Fay–Herriot model when auxiliary variables are measured with error. *Test*, 1–30.

Chen, Q., Elliott, M.R. & Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, **36**, 23.

Datta, G.S. (2009). Model-based approach to small area estimation. In *Sample Surveys: Inference and Analysis, Handbook of Statistics 29b*, Eds. Pfeffermann, D. & Rao, C.R., North-Holland: Amsterdam, pp. 251–288.

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

Dong, T. & Wakefield, J. (2020). Modeling and presentation of health and demographic indicators in a low- and middle-income countries context. Submitted.

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

Dwyer-Lindgren, L., Kakungu, F., Hangoma, P., Ng, M., Wang, H., Flaxman, A.D., Masiye, F. & Gakidou, E. (2014). Estimation of district-level under-5 mortality in Zambia using birth history data, 1980–2010. *Spatial and Spatio-Temporal Epidemiology*, **11**, 89–107.

Fay, R.E. & Herriot, R.A. (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.

Fintzi, J., Cui, X., Wakefield, J. & Minin, V.N. (2017). Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics*, **26**, 918–929.

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

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153–164.

Gething, P.W., Casey, D.C., Weiss, D.J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K.E., Dalrymple, U., Rozier, J., Rao, P.C., Kutz, M.J., Barber, R.M., Huynh, C., Shackleford, K.A., Coates, M.M., Nguyen, G., Fraser, M.S., Kulikoff, R., Wang, H., Naghavi, M., Smith, D.L., Murray, C.J.L., Hay, S.I. & Lim, S.S. (2016). Mapping plasmodium falciparum mortality in Africa between 1990 and 2015. *New England Journal of Medicine*, **375**, 2435–2445.

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

Gutreuter, S., Igumbor, E., Wabiri, N., Desai, M. & Durand, L. (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 *Foundations of Statistical Inference*, Eds. Godambe, V.P. & Sprott, D.A., Holt, Rinehart and Winston: Toronto, pp. 326.

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

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, **8**, 183–200.

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

Li, Z.R., Hsiao, Y., Godwin, J., Martin, B.D., Wakefield, J. & Clark, S.J. (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., Martin, B.D., Dong, T.Q., Fuglstad, G.-A., Godwin, J., Paige, J., Riebler, A., Clark, S. & Wakefield, J. (2020). Space-time smoothing of demographic and health indicators using the R package SUMMER. Submitted.

Lindgren, F., Rue, H. & Lindström, J. (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.

Lohr, S.L. (2019). *Sampling: Design and Analysis:* Second Edition. CRC Press: New York.

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

Malawi DHS (2016). Malawi Demographic Health Survey 2016–16. In Technical report, ICF.

Martin, B.D., Li, Z.R., Hsiao, Y., Godwin, J., Wakefield, J. & Clark, S.J. (2018). Summer: Spatio-temporal under-five mortality methods for estimation in R. R package version 0.2.1.

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

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

Murthy, M.N. & Rao, T.J. (1988). Systematic sampling with illustrative examples. In *Handbook of Statistics, Volume 6*, Elsevier Science Amsterdam: Amsterdam, pp. 147–185.

Paige, J., Fuglstad, G.-A., Riebler, A. & Wakefield, J. (2020). Model-based approaches to analysing spatial data from complex surveys. *Journal of Survey Statistics and Methodology*. To appear.

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

Pfeffermann, D. (2011). Modelling of complex survey data: why model? why is it a problem? how can we approach it. *Survey Methodology*, **37**, 115–136.

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

Pramanik, S., Muthusamy, N., Gera, R. & Laxminarayan, R. (2015). Vaccination coverage in India: a small area estimation approach. *Vaccine*, **33**, 1731–1738.

Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation*, Second Edition. John Wiley: New York.
Riebler, A., Sørbye, S.H., Simpson, D. & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, **25**, 1145–1165.
Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press: Boca Raton.
Rue, H., Martino, S. & Chopin, N. (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., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P. & Lindgren, F.K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, **4**, 395–421.
Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer: New York.
Scott, A. & Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, **64**, 830–840.
Simpson, D., Rue, H., Riebler, A., Martins, T.G. & Sørbye, S.H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). *Statistical Science*, **32**, 1–28.
Skinner, C. & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, **32**, 165–175.
Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer: New York.
Stevens, F.R., Gaughan, A.E., Linard, C. & Tatem, A.J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, **10**, e0107042.
Stringhini, S., Wisniak, A., Piumatti, G., Azman, A.S., Lauer, S.A., Baysson, H., De Ridder, D., Petrovic, D., Schrempft, S., Marcus, K. & Arm-Vernez, I. (2020). Repeated seroprevalence of anti-SARS-Cov-2 IgG antibodies in a population-based sample from Geneva, Switzerland. *The Lancet*. Published online: June 11, 2020.
Thompson, S.K. (1992). *Sampling*. John Wiley and Sons: New York.
Thompson, S.K. & Seber, G.A.F. (1996). *Adaptive Sampling*. Wiley: New York.
Turk, P. & Borkowski, J.J. (2005). A review of adaptive cluster sampling: 1990–2003. *Environmental and Ecological Statistics*, **12**, 55–94.
Utazi, C.E., Thorley, J., Alegana, V.A., Ferrari, M.J., Takahashi, S., Metcalf, C.J.E., Lessler, J. & Tatem, A.J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, **36**, 1583–1591.
Valliant, R., Dorfman, A.H. & Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley: New York.
Vogel, G. (2020). First antibody surveys draw fire for quality, bias. *Science*, **368**, 350–351.
Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health*, **29**, 75–90.
Wakefield, J. (2020). Prevalence mapping. In *Wiley Statsref: Statistics Reference Online*, Wiley: New York, pp. 1–7.
Waller, L.A. & Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons: Hoboken, New Jersey.
Wardrop, N.A., Jochem, W.C., Bird, T.J., Chamberlain, H.R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V. & Tatem, A.J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, **115**, 3529–3537.
Wilson, K. & Wakefield, J. (2020). Pointless spatial modeling. *Biostatistics*, **21**, e17–e32.
Wolter, K. (2007). *Introduction to Variance Estimation*, Second Edition. Springer Science & Business Media: New York.
Yates, F. & Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, **15**, 253–261.
Ybarra, L.M.R. & Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919–931.
You, Y. & Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, **30**, 3–15.
Zhang, J.L. & Bryant, J. (2020). Fully Bayesian benchmarking of small area estimation models. *Journal of Official Statistics*, **36**, 197–223.
Zheng, H. & Little, R.J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–117.

[*Received June 2020, Revised July 2020, Accepted July 2020*]

## Supporting Information

Supporting information may be found in the online version of this article.