Spatial Statistics 52 (2022) 100714

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Spatial aggregation with respect to a population distribution: Impact on inference

John Paige [a,*], Geir-Arne Fuglstad [a], Andrea Riebler [a], Jon Wakefield [b]

[a] *Department of Mathematical Sciences, NTNU, Trondheim, Norway*
[b] *Department of Statistics and Biostatistics, University of Washington, Seattle, USA*

ARTICLE INFO

Article history:
Received 12 July 2022
Received in revised form 6 October 2022
Accepted 1 November 2022
Available online 9 November 2022

Dataset link: https://dhsprogram.com/, https://www.worldpop.org/, https://dataforgood.facebook.com/dfg/, https://github.com/paigejo/continuousNugget, https://github.com/paigejo/continuousNugget, https://github.com/paigejo/SUMMER, https://github.com/paigejo/SUMMER

Keywords:
Geostatistics
Small area estimation
Risk
Prevalence
Bayesian
Integrated nested Laplace approximations

ABSTRACT

Spatial aggregation with respect to a population distribution involves estimating aggregate population quantities based on observations from individuals. In this context, a geostatistical workflow must account for three major sources of *aggregation error*: aggregation weights, fine scale variation, and finite population variation. However, these sources of aggregation error are commonly ignored, and the population instead treated as a fixed population density surface. We improve common practice by introducing a *sampling frame model* allowing aggregation models to account for aggregation error simply and transparently. This preserves aggregate point estimates while increasing their uncertainties.

We compare the proposed and the traditional approach using two simulation studies mimicking neonatal mortality rate (NMR) data from the 2014 Kenya Demographic and Health Survey. In the traditional approach, undercoverage/overcoverage of interval estimates depends arbitrarily on the aggregation grid resolution, while the new approach is resolution robust. Differences between the aggregation approaches increase as an area's population decreases, and are particularly large at the second administrative level and finer, but also at the first administrative level for some population quantities. These findings are consistent with those of an application to the true NMR data. We demonstrate in a sensitivity analysis that burden estimates and their uncertainties are not robust to changes in population density and census information, while prevalence estimates and uncertainties seem stable.

[*] Corresponding author.
E-mail address: john.paige@ntnu.no (J. Paige).

https://doi.org/10.1016/j.spasta.2022.100714

1. Introduction

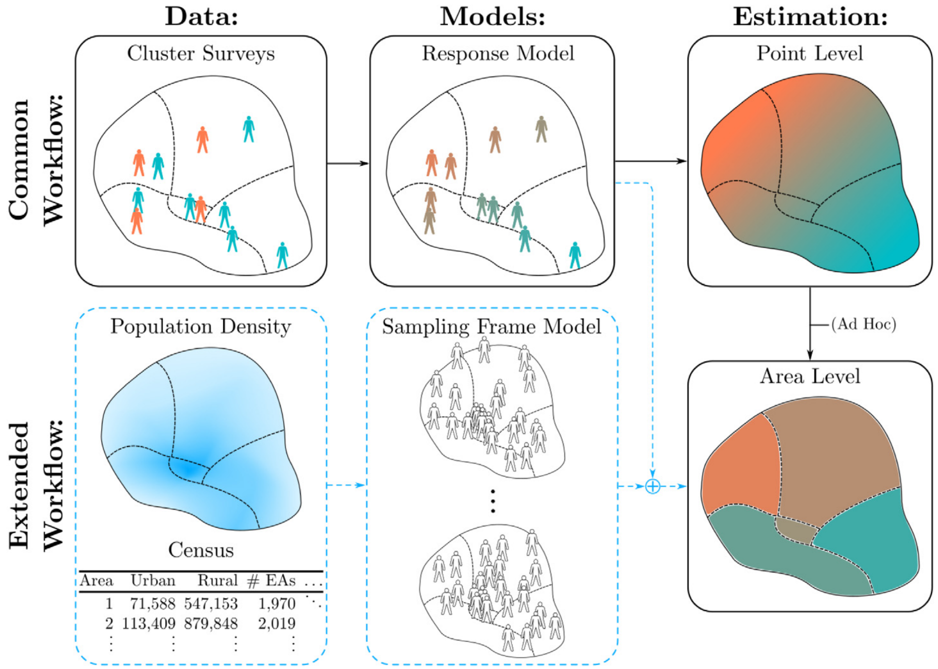
Spatial aggregation based on point-referenced observations is an important problem in spatial statistics (Gelfand et al., 2010). If the quantities of interest can be written as integrals of a spatial field, the desired posterior distributions can be computed by block kriging (Gelfand et al., 2010) or are directly available in basis decomposition methods such as fixed rank kriging (Cressie and Johannesson, 2008), the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011), and LatticeKrig and its extensions (Nychka et al., 2015; Paige et al., 2022a). However, in some cases, point-referenced measurements are collected from a 'target' population, a finite population of interest, that may consist of people (Kenya National Bureau of Statistics, 2014), plants, or animal species (Funwi-Gabga and Mateu, 2012; Ballmann et al., 2017; Laber et al., 2018). In these cases, aggregate estimates, often at multiple different areal resolutions, may be desired for the target population from which the observations were collected. We term this problem *spatial aggregation with respect to a population distribution*.

Our focus is small area estimation (SAE) of *prevalence*, i.e., the proportion of individuals with outcome 1, based on binary responses (0 or 1). Some approaches (Giorgi et al., 2018; Dwyer-Lindgren et al., 2019; Osgood-Zimmerman et al., 2018) approximate the prevalence with the *risk*, which is the expected number of individuals with outcome 1. It is worth emphasizing that even if we knew the risk in an area exactly, say $r = 0.7$, the prevalence p could vary widely around this number for a small population size just as an empirical binomial proportion might vary around its probability.

In this context, we identify three major sources of *aggregation error*: (1) aggregation weights, (2) fine scale variation, and (3) finite population variation. By aggregation weights we mean the weights used to take a weighted integral or average of point level estimates to produce areal estimates. These weights may involve population density, for example, or the proportion of population in the urban or rural part of an area. Fine scale variation is variability occurring at the finest modeled spatial scale, such as the scale of the response. Fine scale variability could be induced by unmodeled nonspatial or discrete spatial covariates, for example, or other local conditions. Finite population variation is variability caused by the finite size of the target population, and is the cause of variation in prevalence about the underlying risk.

Geostatistical models applied to, for example, neonatal mortality, women's secondary education, child growth failure, and vaccination coverage, routinely do not completely account for aggregation error when aggregating from point level predictions to the areal level (Paige et al., 2022b; Dong and Wakefield, 2021; LBD and others, 2020, 2021), as referenced in Fig. 1. For example, LBD and others (2020) and LBD and others (2021) do not include finite population variation, and include fine scale variation at the pixel level, which causes inference to depend on the grid resolution of the model. Also, Paige et al. (2022b) does not account for finite population variation and similarly only partially accounts for fine scale variation. Dong and Wakefield (2021) bases predictions on a single simulated population, which does not account for the full distribution of possible aggregation weights and target populations. We show in Section 5 that not (or even only partially) accounting for these sources of aggregation error can lead to poor predictive performance in some contexts.

We propose a spatial *aggregation model* where we combine a *response model* for the data with a *sampling frame model* that expresses uncertainty about the population distribution. The response model is a Bayesian hierarchical model with a data model, a process model, and a parameter model specifying priors. The sampling frame model by contrast describes the population distribution that is used when producing population aggregates. The term "sampling frame" is borrowed from survey statistics, and refers to the full list of the individuals and associated auxiliary information such as spatial locations and covariate values. The spatial aggregation model generalizes models



Area	Urban	Rural	# EAs	…
1	71,588	547,153	1,970	⋱
2	113,409	879,848	2,019	
⋮	⋮	⋮	⋮	

Fig. 1. Common geostatistical approach to spatial aggregation with respect to a population distribution (in black), with proposed additions (in dashed blue). While population data is sometimes used aggregating from point to areal level in common approaches, this procedure is ad hoc and does not account for several sources of aggregation error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proposed in Paige et al. (2022b) and Dong and Wakefield (2021), and is essential to account for aggregation error in the aggregate estimates. While some information about the sampling frame may be unknown, such as the exact number of unobserved individuals and their locations, external information such as population density maps provide a strong prior. Our proposed workflow is given in Fig. 1, which shows both the common geostatistical workflow as well as our proposed method for spatial aggregation with respect to a population.

We apply our proposed spatial aggregation model to neonatal mortality rates (NMRs), which is the prevalence of mortality among neonatals—children within 28 days of live birth. Data on NMR and neonatal mortality burden, the total neonatal mortality count, is observed from a Demographic and Health Surveys (DHS) survey. In DHS surveys, observations are made by selecting small areal units and sampling a subset of the individuals within. These areal units are called enumeration areas (EAs) and, e.g., the sampling frame in Kenya is divided into 96,251 EAs. We will treat the observations as point-referenced observations of clusters. The cluster level is a natural place to include the nugget effect and the spatial aggregation model should be constructed based on this choice. DHS surveys are conducted under complex survey designs where inference on population averages and totals in a given set of areas is of primary interest. However, producing estimates at multiple areal resolutions is also of interest. While parts of the design will be acknowledged in the proposed spatial model, as discussed in Paige et al. (2022b), survey statistics will not be a focus of this paper.

The observation of small areal units with the goal of making aggregate estimates for larger areas connects to the idea of basic areal units (Nguyen et al., 2012; Zammit-Mangion and Cressie, 2017) and the modifiable areal unit problem (MAUP) (Gehlke and Biehl, 1934; Openshaw and Taylor, 1979). The MAUP is still not well understood (Manley, 2014), and we are not proposing a general solution, but a practical way to address it in small area estimation based on cluster surveys, particularly in cluster surveys exhibiting fine scale variation.

Section 2 introduces the 2014 Kenya demographic health survey (KDHS2014), the dataset motivating this work. In Section 3 we define the statistical problem of interest along with common approaches of estimation, and discuss how they relate to the three considered major sources of aggregation error when aggregating point level predictions to the areal level for a population. We introduce a number of sampling frame models in Section 4, and show how the robustness of different aggregation models can depend on their ability to carefully account for aggregation error in Section 5. In Section 6, we conduct a simulation study to investigate the relative importance of different sources of aggregation uncertainty at different spatial scales. Then we investigate the practical implications for the analysis of NMR data from KDHS2014 in Section 7. In Section 8 we conduct a sensitivity analysis where we assess how changes in the population information used in the application affect the results of the analysis. The paper ends with discussion and conclusions in Section 9.

2. Neonatal mortality in 2010–2014 in Kenya

The 2014 KDHS provides information on a number of important health and demographic indicators including NMR. It consists of 1,582 clusters selected out of 96,215 EAs, where EAs are selected with probability proportional to the number of contained households. Within each selected EA, 25 households selected with simple random sampling from the associated cluster. The centroids of the sampled clusters are known up to a small amount of jittering, but the locations of the other EAs are unknown. Due to the stratification of the 2014 KDHS, the total number of urban and rural EAs is known within each 'Admin1' area, the administrative areas immediately below the national level.

Our aim is to estimate NMR and the burden of neonatal mortality in 2010–2014 for Kenya's 47 Admin1 and 301 Admin2 areas defined by Global Administrative Areas (GADM), where Admin2 areas are the administrative areas defined just below the Admin1 level. Of the 301 Admin2 areas originally defined, we combine the bordering 'unknown 8' and Kakeguria areas due to the small size and estimated population density of 'unknown 8'. We are also interested in the relative prevalence in Admin1 and Admin2 areas, which we define as the prevalence in the urban part of the area divided by the prevalence in the rural part.

Population density estimate maps are available from WorldPop (Stevens et al., 2015; Tatem, 2017) at 1 km resolution, and we produce urbanicity (adjusted population density) maps by thresholding (normalizing) population density to match the urban/rural population proportions (totals) at the Admin1 level based on the 2009 census (Kenya National Bureau of Statistics, 2014; KDHS, 2009), as in Paige et al. (2022b). More information on population density calculations and adjustments are provided in the supplement in Section S.2. Sampled NMRs are pictured for the 1,582 clusters in Fig. 2 along with the estimated 2014 population density. General population totals at the stratum (Admin1 × urban/rural) level are fixed at the 2014 estimates as given by the 2014 KDHS.

3. Sources of aggregation error in common practice

If the locations of each EA in area A with index $i = 1, \ldots, M$ were known along with the number of members of the target population that were born, N_i, and died Z_i in the time period, the prevalence and burden in A could be calculated as

$$p(A) = \sum_{i=1}^{M} \frac{N_i}{N} \frac{Z_i}{N_i}, \quad \text{and} \quad b(A) = \sum_{i=1}^{M} Z_i \tag{1}$$

respectively, where $N = \sum_{i=1}^{M} N_i$.

Since the locations of the EAs and since the N_i and Z_i are unknown, a number of methods have been used to estimate prevalence that avoid the formulation in (1). Typically, these methods begin

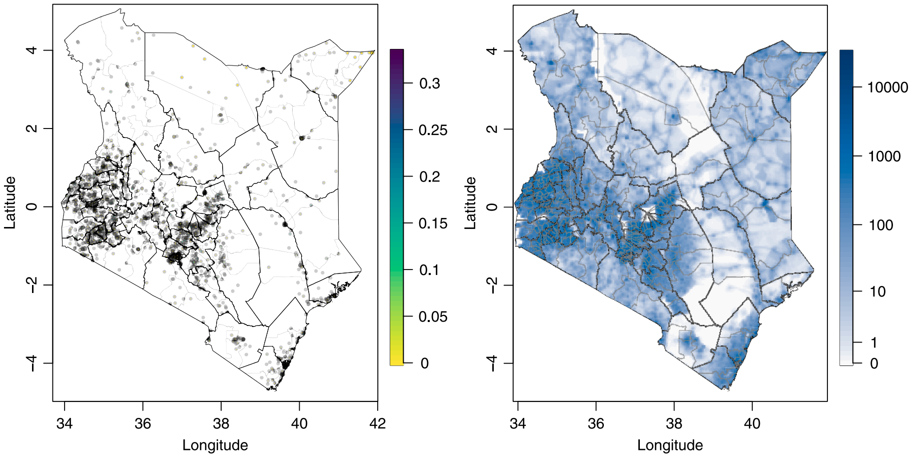


Fig. 2. Neonatal mortality rates in 2010–2014 as observed in 2014 KDHS sampled clusters (left), and estimated population density in people per km^2 in 2014 (right). Admin1 and Admin2 boundaries are shown as black and gray lines respectively.

with the following standard geostatistical prevalence sampling model with notation based on Diggle and Giorgi (2016):

$$\begin{aligned} y_c \mid r_c, n_c &\sim \text{Binomial}(n_c, r_c) \\ \text{logit}(r_c) &= \boldsymbol{d}(\boldsymbol{s}_c)^T \boldsymbol{\beta} + u(\boldsymbol{s}_c) + \epsilon_c, \quad c = 1, \dots, n. \end{aligned} \tag{2}$$

Here, the cluster level response y_c and cluster locations $\boldsymbol{s}_c$ are indexed by the cluster index c for n clusters in total. The cluster level risk is given by r_c, and the number of individuals in the target population in cluster c is given by n_c. The vector $\boldsymbol{d}(\boldsymbol{s}_c)$ contains the spatial covariates (possibly including an intercept) at the location of cluster c with associated effect sizes given by $\boldsymbol{\beta}$, and $\epsilon_c \sim N(0, \sigma_\epsilon^2)$ is independent and identically distributed (iid) variation at the level of the response, or the spatial nugget. The spatial effect $u = \{u(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$ is a centered stationary Gaussian random field with marginal variance σ_S^2.

There are various approaches for how to choose u, some of which are summarized in Paige et al. (2022b). For example, it is sometimes assumed that u follows a reparameterization of the Besag York Mollié model known as the BYM2 model (Riebler et al., 2016), a solution to a stochastic partial differential equation (SPDE) (Lindgren et al., 2011), or an extended LatticeKrig (ELK) model (Paige et al., 2022a). We will assume u is stationary with an isotropic Matérn covariance function $\text{Cov}(u(\boldsymbol{s}), u(\boldsymbol{s}')) = C(|\boldsymbol{s} - \boldsymbol{s}'|)$ (Matérn, 1986), and will approximate it using the SPDE approach, and will represent the effective spatial range as ρ.

Since (2) is only a model for the response, a second model is required for estimating areal prevalence. Slightly modifying notation from Keller and Peng (2019), a population quantity of interest (such as risk) for a spatial region A is sometimes estimated as,

$$r(A) = \int_A r(\boldsymbol{s}) \, \mathrm{d}Q(\boldsymbol{s}), \tag{3}$$

treating $r(\boldsymbol{s})$, the quantity of interest at location $\boldsymbol{s}$, as a spatial function, and where $Q(\boldsymbol{s})$ is a spatial distribution of interest. For example, the spatial density $q(\boldsymbol{s}) = \mathrm{d}Q(\boldsymbol{s})/\mathrm{d}\boldsymbol{s}$ is typically chosen as:

- $q(\boldsymbol{s}) = 1$ for **areal totals**,
- $q(\boldsymbol{s}) = 1/|A|$ for **areal averages**,
- population density for **population totals**, or
- population density normalized to have unit integral over A for **population averages** (such as for risk).

In the context of estimating population averages, sometimes areal averages are used to approximate population averages (e.g. Diggle and Giorgi 2016 and Giorgi et al. 2018), and population density is based on estimates that may oversmooth in urban areas unless this effect is corrected (Leyk et al., 2019), such as based on census population totals. Using areal averages in lieu of population averages, and not correcting for potential oversmoothing in population density are examples of potential 'aggregation weight errors', errors due to the choice in spatial distribution of interest Q or in its density q. An incorrect choice for Q and its density q results in biased estimates and larger mean squared error as discussed (and quantified under simplifying assumptions) in Section S.1 in the supplement.

Typically, (3) is approximated using a numerical integration/aggregation grid, although the resolution of the grid is not consistent. Models in this context may, for example, produce estimates at the 20 km (Lessler et al., 2018), 5 km (Osgood-Zimmerman et al., 2018; Graetz et al., 2018), 1 km (Utazi et al., 2018), or 100 m (Tatem, 2017) resolution, and the effects of grid resolution on model estimates is not always well understood. In fact, in Osgood-Zimmerman et al. (2018) and Graetz et al. (2018) a single nugget effect is included in each grid cell when aggregating to the areal level, weighted by the normalized population density. In that case, the number of nugget effects averaged over in an area's population average is equal to the number of grid points in the area, but the response model is the same regardless of the aggregation resolution. Hence, the variance of the population average depends arbitrarily on the aggregation grid resolution. We refer to that as the 'gridded' sampling frame model, with predicted risk and expected burden in area A denoted by $r_{\mathrm{grid}}(A)$ and $b_{\mathrm{grid}}(A)$ respectively. We call this source of aggregation error 'fine scale variation' since it relates to fine scale variation in the context of the MAUP, and the nugget effect, which is sometimes interpreted as fine scale variation.

If the ultimate goal is to estimate the population prevalence in (1), it should be emphasized that (3) is only an approximation in practice, since it typically takes the form of a continuous integral with respect to the spatial density q rather than summing over the inherently discrete individuals in the population, and since $r(\boldsymbol{s})$ is the risk rather than prevalence at a point, and so does not include all the variation of prevalence. We call this source of aggregation uncertainty 'finite population variation'. Despite the inherent difference between risk and prevalence, the terms are often used interchangeably (Giorgi et al., 2018; Dwyer-Lindgren et al., 2019; Osgood-Zimmerman et al., 2018), perhaps due to an unstated assumption that finite population variation is negligible.

4. Spatial aggregation with respect to a population distribution

Our goal will be to estimate NMR in 2010–2014 using the 2014 KDHS across a set of areas, with a generic area being denoted A. The number of EAs in A or in an area containing A is known, and the number of houses and neonatals born in the time period in A or in an area containing A is also approximately known based on census data. We will also assume each EA has at least 25 households, since 25 households are sampled in each cluster. Although it is not too difficult to relax these assumptions, we make them for the sake of simplicity. Since the totals in the urban and rural parts of each Admin1 area are known, we can think of area A as the urban or rural part of an Admin1 area, or the urban or rural parts of an Admin2 area. Given this information and the aggregation model, the joint distributions for the prevalence and burden in any set spatial region, such as for Admin2 areas or their urban or rural parts, is directly implied, even if the total target population, number of households, and number of EAs in those particular regions are all unknown.

We introduce three sampling frame models in Sections 4.1–4.3, for which the associated aggregation model is constructed by linking the sampling frame model to the response model. The link is formed by relating the logit risk of EA i (not necessarily observed) to (2) with $\mathrm{logit}(r_i) = \mathrm{logit}(r(\boldsymbol{s}_i)) = \boldsymbol{d}(\boldsymbol{s}_i)^T\boldsymbol{\beta} + u(\boldsymbol{s}_i) + \epsilon_i$ (with slight modifications in the case of one sampling frame model) for each i, and where ϵ_i is iid $N(0, \sigma_\epsilon^2)$. Given the aggregation models share the same response model, they have the same central predictions, varying only in their uncertainty, with the empirical aggregation model having the highest predictive variance and the smooth latent model having the lowest (see Section S.3 in the supplement for details). Fig. 3 provides an illustration summarizing how difference sources of uncertainty are incorporated into the main three considered aggregation models introduced in this section and their estimates of prevalence/risk.

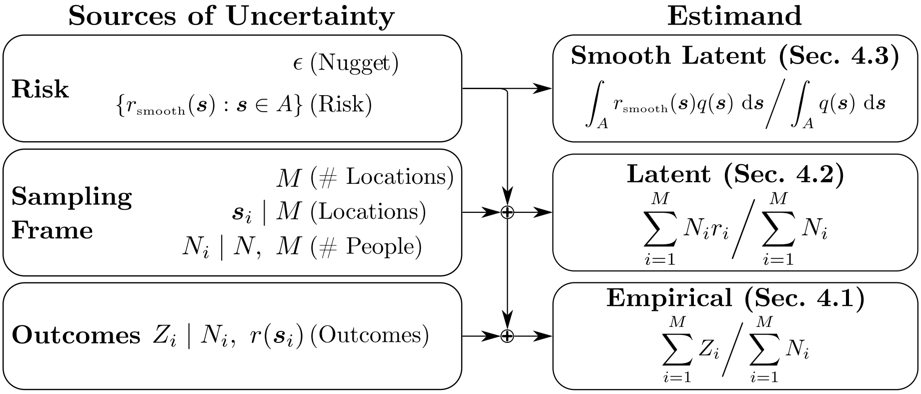

Sources of Uncertainty
Estimand
Risk
ϵ (Nugget)
$\{r_{\text{smooth}}(\boldsymbol{s}) : \boldsymbol{s} \in A\}$ (Risk)
Sampling Frame
M (# Locations)
$\boldsymbol{s}_i \mid M$ (Locations)
$N_i \mid N,\ M$ (# People)
Outcomes $Z_i \mid N_i,\ r(\boldsymbol{s}_i)$ (Outcomes)
Smooth Latent (Sec. 4.3)
$\int_A r_{\text{smooth}}(\boldsymbol{s})q(\boldsymbol{s})\ \text{d}\boldsymbol{s} \Big/ \int_A q(\boldsymbol{s})\ \text{d}\boldsymbol{s}$
Latent (Sec. 4.2)
$\sum_{i=1}^{M} N_i r_i \Big/ \sum_{i=1}^{M} N_i$
Empirical (Sec. 4.1)
$\sum_{i=1}^{M} Z_i \Big/ \sum_{i=1}^{M} N_i$

Fig. 3. Summary of how different sources of uncertainty are incorporated into the three main aggregation models considered: smooth latent, latent, and empirical. Model estimands for prevalence/risk along with the sections in which the models are introduced are also given.

4.1. Empirical model

Rather than approximating (1) with (3), it is possible to model the terms in (1) directly. Since the number of EAs in A is known, we will assume that the locations of the EAs, $\{\boldsymbol{s}_i\}_{i=1}^M$, follow a Poisson process with a possibly fixed number of points, M, and with rate proportional to the population density. Since N, the total target population and number of households in A, is approximately known from census data, we assume (as above) there are at least 25 households per EA, and that the rest of the total, known from census data, are distributed according to a multinomial distribution with equal probability per EA. If h_i is the number of households in EA i for each i, we distribute the total number of neonatals among EAs in A, $\{N_i\}_{i=1}^M$, according to a multinomial distribution with probability $h_i/\sum_i h_i$ of being in EA i for each i. Lastly, we assume $Z_i|N_i, r_i \sim \text{Binomial}(N_i, r_i)$. We can then calculate the population prevalence and burden conditional on the Z_i and N_i with (1). We will denote the prevalence and burden according to the empirical model as $p_{\text{emp}}(A)$ and $b_{\text{emp}}(A)$ respectively.

We call this the ‘empirical’ sampling frame model, since it directly models the quantities of interest, prevalence and burden, and since prevalence unlike risk is directly observable. Since population density is used in the point process model for the EA locations, and since census information is used to set the number of EAs, households, and neonatals in A, it correctly accounts for aggregation weights when producing the population estimate. Further, the nugget is included at the level of the EA, and N_i and Z_i is modeled directly for each EA, so fine scale and finite population variation are both accounted for as well.

4.2. Latent model

In this case, we make nearly all the same assumptions as in the empirical model, except we approximate the population prevalence with the risk, and so it is no longer necessary to assume any distribution for $Z_i|N_i, r_i$. To calculate the population risk and the ‘expected burden’ with the ‘latent’ sampling frame model, we use:

$$r_{\text{latent}}(A) = \sum_{i=1}^{M} \frac{N_i}{N} r_i, \quad \text{and} \quad b_{\text{latent}}(A) = \sum_{i=1}^{M} N_i r_i. \tag{4}$$

We call this the latent sampling frame model since it only models prevalence indirectly through the risk, which is a latent quantity and cannot be directly observed. The latent model accounts for population weights and fine scale variation for the same reasons as the empirical model, but it only partially accounts for finite population variation, since it includes variation in each N_i but not each Z_i (except through r_i). This is somewhat similar to the model in Dong and Wakefield (2021), except it proposes a full distribution for the EA locations and population denominators rather than fixing a single draw from the distribution for the EA locations and setting $N_i \approx N/M$ for each N_i, rounding to the nearest integer.

4.3. Smooth latent model

The 'smooth latent' sampling frame model is the same as the sampling frame model used in Paige et al. (2022b). The integral in (3) is approximated numerically on a grid using spatial density q equal to the normalized population density. Further, the 'smooth risk' at any each grid point is calculated by integrating out the nugget (i.e. the fine scale variation) as:

$$r_{\text{smooth}}(\boldsymbol{s}) = \int_{-\infty}^{\infty} \text{expit}\left\{\boldsymbol{d}(\boldsymbol{s})^T\boldsymbol{\beta} + u(\boldsymbol{s}) + \epsilon\right\} \frac{\phi(\epsilon/\sigma_\epsilon)}{\sigma_\epsilon}\, \mathrm{d}\epsilon, \tag{5}$$

where ϕ is the standard Gaussian density. This is fundamentally different from setting the nugget to 0 as would be the case if it were interpreted as measurement error, since the mean is shifted due to the nonlinear logit link function. The smooth risk for an area, $r_{\text{smooth}}(A)$, integrates $r_{\text{smooth}}(\boldsymbol{s})$ with respect to population density normalized to have unit integral over A, and the 'smooth burden', $b_{\text{smooth}}(A)$, can be calculated by integrating $r_{\text{smooth}}(\boldsymbol{s})$ with respect to the unnormalized population density.

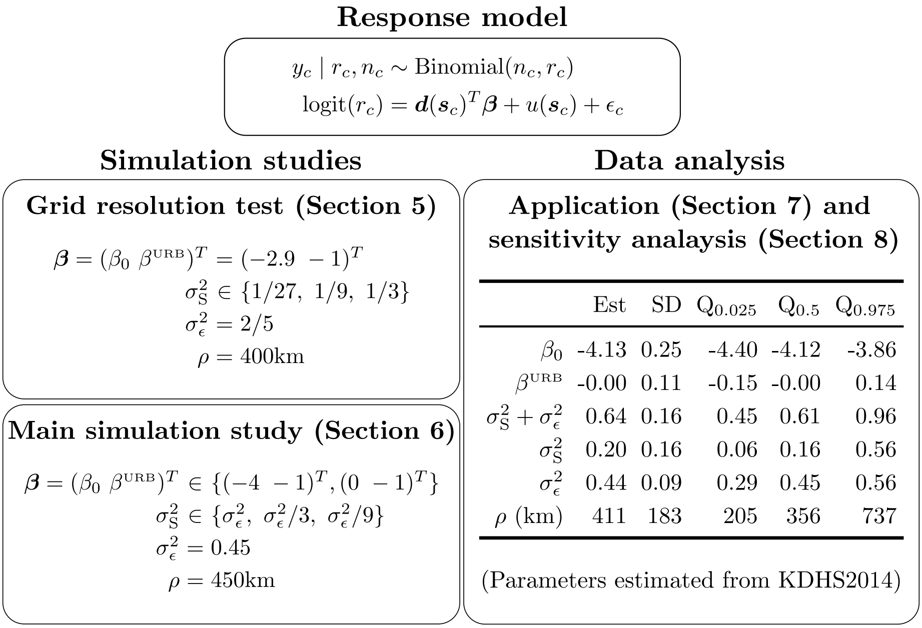
We call this the smooth latent model because, rather than summing over discrete risks as in the latent model, the smooth risk, a continuously indexed (although possibly discontinuous) spatial function, is integrated continuously. Since the smooth latent model accounts for population density, which can be scaled to match population totals using census data as in Paige et al. (2022b), this sampling frame model does account for aggregation weight error. By integrating out the nugget effect, the smooth latent model ensures its central predictions account for fine scale variation even if it does not fully account for fine scale variation, as described in detail in Section S.3 in the supplement. Since smooth risk is a risk rather than a prevalence, the smooth latent model does not account for finite population variation.

5. Importance of aggregation assumptions: Grid resolution test

We run a simple simulation study of Nairobi County NMRs based on the KDHS2014 and using a continuous spatial response model based on Paige et al. (2022b) to illustrate how different areal aggregation assumptions can yield substantively different areal predictive distributions, and to show that sampling frame models that are not chosen carefully can lead to poor properties in their predictions. We do this by comparing areal predictions of four different aggregation models that share a response model, but differ in their sampling frame model and therefore in their method of producing areal estimates.

We take $\boldsymbol{d}(\boldsymbol{s}) = [1\ I(\boldsymbol{s} \in U)]^T$ where $I(\boldsymbol{s} \in U)$ is an indicator that is 1 when $\boldsymbol{s}$ is in an urban area and 0 otherwise. We choose Nairobi County since Nairobi is the capital of Kenya, and so it is especially important to estimate its neonatal mortality prevalence accurately. In addition, it is one of the smallest and most densely populated counties in Kenya, so it is possible to increase the integration grid resolution with less computational expense.

We simulate 100 neonatal populations with associated mortality risks and prevalences down to the EA level across all of Kenya, generating one simulated survey per population, and basing simulation parameters and the survey design on the KDHS2014. This procedure is repeated for 4 grid resolutions: 200 m, 1 km, 5 km, and 25 km. In each case, we generate predictions for all 17 Admin2 areas (constituencies) in Nairobi, Kenya, calculating 95% credible interval (CI) widths and associated empirical coverages. For prediction, we fix the response model parameters to the true values given in Fig. 4, where β_0 and β^{URB} are the intercept and urban effect respectively. We also assume for simplicity that there are exactly 25 households and neonatals per EA. Predictions are generated for the gridded, smooth latent, latent, and empirical aggregation models, which share



Response model

$$y_c \mid r_c, n_c \sim \text{Binomial}(n_c, r_c)$$
$$\text{logit}(r_c) = \boldsymbol{d}(\boldsymbol{s}_c)^T \boldsymbol{\beta} + u(\boldsymbol{s}_c) + \epsilon_c$$

Simulation studies

Grid resolution test (Section 5)

$$\boldsymbol{\beta} = (\beta_0 \ \beta^{\text{URB}})^T = (-2.9 \ -1)^T$$
$$\sigma_S^2 \in \{1/27, \ 1/9, \ 1/3\}$$
$$\sigma_\epsilon^2 = 2/5$$
$$\rho = 400\text{km}$$

Main simulation study (Section 6)

$$\boldsymbol{\beta} = (\beta_0 \ \beta^{\text{URB}})^T \in \{(-4 \ -1)^T, (0 \ -1)^T\}$$
$$\sigma_S^2 \in \{\sigma_\epsilon^2, \ \sigma_\epsilon^2/3, \ \sigma_\epsilon^2/9\}$$
$$\sigma_\epsilon^2 = 0.45$$
$$\rho = 450\text{km}$$

Data analysis

Application (Section 7) and sensitivity anaalysis (Section 8)

	Est	SD	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$
β_0	-4.13	0.25	-4.40	-4.12	-3.86
β^{URB}	-0.00	0.11	-0.15	-0.00	0.14
$\sigma_S^2 + \sigma_\epsilon^2$	0.64	0.16	0.45	0.61	0.96
σ_S^2	0.20	0.16	0.06	0.16	0.56
σ_ϵ^2	0.44	0.09	0.29	0.45	0.56
ρ (km)	411	183	205	356	737

(Parameters estimated from KDHS2014)

Fig. 4. The response model used throughout the analysis along with the parameters used for simulation or estimated. Each response model is linked with a sampling frame model to form an aggregation model with aggregate predictions. The 2.5th, 50th, and 97.5th percentiles of the posterior are denoted by $Q_{0.025}$, $Q_{0.5}$, and $Q_{0.975}$ respectively. In Section 5 parameters are fixed to the truth when predicting, in Section 6 they are estimated for each simulated dataset, and in Sections 7 and 8 they are estimated from 2014 KDHS NMR data.

Table 1
95% credible interval (CI) widths in neonatals per thousand, and empirical coverages in percent for considered sampling frame models as a function of aggregation grid resolution, with $\sigma_S^2 = 1/27$ and standard errors given in parentheses.

	Model	Units	200 m	1 km	5 km	25 km
95% CI width	Empirical	(per 1000)	9.8 (0.1)	9.8 (0.1)	9.8 (0.1)	9.9 (0.1)
	Latent		8.2 (0.1)	8.2 (0.1)	8.2 (0.1)	8.3 (0.1)
	Smooth Latent		7.7 (0.1)	7.7 (0.1)	7.7 (0.1)	7.7 (0.1)
	Gridded		8.5 (0.1)	19 (0.3)	56.5 (1.0)	60.1 (1.0)
95% CI coverage	Empirical	(Percent)	96 (1.0)	96 (1.0)	96 (1.0)	96 (1.1)
	Latent		91 (1.5)	91 (1.6)	91 (1.5)	91 (1.6)
	Smooth Latent		89 (1.7)	89 (1.7)	89 (1.6)	89 (1.6)
	Gridded		92 (1.4)	100 (0.1)	100 (0.0)	100 (0.0)

the same response model, and vary only in their sampling frame model. This process is repeated for three different values of σ_S^2: 1/27, 1/9, and 1/3.

Table 1 (and Table D.6 in the supplement) show how 95% (and 50%) CI width and empirical coverage estimates and their standard errors change as a function of integration grid resolution for each model when $\sigma_S^2 = 1/27$. Tables D.5–D.10 in Section S.4 in the supplement provide these results for all values of σ_S^2 considered, although coverages do not substantially depend on σ_S^2, so in this section we discuss only the case of $\sigma_S^2 = 1/27$, which has the highest precision estimates of coverage and CI widths.

Despite the spatial response model being identical in all aggregation models, the differences in the CI widths and coverages as a function of grid resolution show how using the gridded sampling frame model can lead to problematic results: a 1 km analysis will yield very different CI widths and coverages from a 5 km analysis. At 200 m resolution the gridded risk 50% CIs achieved 45% coverage, but as resolution increased, so did the overcoverage, which rose to 99% at 25 km resolution. This is made more problematic by the fact that there is no standard resolution at which to perform spatial aggregation, and the lack of guarantees that a given resolution that may work well predicting some areas will also achieve good coverage in other areas.

Unlike the gridded aggregation model, the empirical, latent, and smooth latent models are highly robust to the grid resolution, achieving consistent CI widths and coverages near the nominal level for all resolutions. The empirical sampling frame model achieved coverages closest to the nominal rate at 95% significance, and at both 95% and 50% significance the empirical sampling frame model was never significantly different from the nominal rate. The coverage of the empirical model in all cases was higher than that of the latent model, which was in turn higher than the smooth latent model. The latent and smooth latent aggregation models consistently exhibited undercoverage at the 95% level. This difference in coverage is due to the different levels of aggregation error accounted for by the models. The main difference between the latent and empirical models is that the empirical model accounts for uncertainty in the outcome, which is part of finite population variation. Differences in coverages between these models in turn suggests that finite population variation can be an important source of aggregation error.

It is worth noting that, as the grid resolution increases, the gridded sampling frame model performs more similar to the smooth latent model. In fact, we show in Section S.3.3 in the supplement that the gridded aggregation model converges to the smooth latent as their resolutions both increase. Since the smooth latent model accounts for fine scale variation in a more principled way, and since it is more robust to the grid resolution, we advocate for using the smooth latent in lieu of the gridded aggregation model, and using the empirical sampling frame model in lieu of the smooth latent.

Overall, the at times considerable differences in the empirical coverages of the aggregation models can be attributed entirely to the differences in the sampling frame models. This suggests that the assumptions made when aggregating from point level predictions to the areal level matter should be considered carefully.

6. What factors influence aggregation error?

6.1. Simulation setup

To test the performance of the aggregation models under various conditions, and identify what factors influence aggregation error, we simulate from 54 different population × survey scenarios, each with a different set of simulation parameters. For each population × survey scenarios, we use the empirical aggregation model to simulate 100 populations, and then generate 100 associated surveys under a sample design intended to mimic the KDHS2014. We calculate prevalence, burden, and the relative prevalence in urban versus rural parts of areas from the simulated populations (i.e., the prevalence in the urban part of an area divided by the prevalence in the rural part). These quantities are calculated at the Admin1 and Admin2 levels as well as for the urban and rural parts of each Admin2 area when defined, and relative prevalence is only calculated at the Admin1 and Admin2 levels since it can only be calculated for areas with both urban and rural parts.

Some parameters used for the simulation study are held fixed, since they are not the focus of the simulation study. All parameter choices are given in Fig. 4. We fix some response model parameters in order to approximately match parameters estimated for an equivalent response model as fit in Section 7. We also set the urban effect to be $\beta^{\text{URB}} = -1$ in order to add stratification by urbanicity to ensure that predictions are able to account for stratification. We choose $\varphi \equiv \sigma_S^2/\sigma_\epsilon^2 \in \{1, 1/3, 1/9\}$ to be the signal to noise ratio of the response model.

For the sampling frame model we choose $r_{\text{pop}} \in \{1/5, 1, 5\}$ to be the number of EAs (and the number of members of the population) included in the simulated population per EA in Kenya compared to the number in the KDHS2014 survey frame, and $r_{\text{samp}} \in \{1/3, 1, 3\}$ to be the number of clusters included in the simulated survey per cluster included in the KDHS2014. For example, if r_{pop} is 1/5, then we assume the simulated population and the simulated number of EAs are only a fifth that in the KDHS2014 sampling frame. Similarly, if r_{samp} is 1/3, then the simulated survey has only a third of the observed clusters as in the KDHS2014. The possible values of r_{pop} (1/5, 1, and 5) are chosen to roughly match the variation in the number of EA in low and middle income countries, where, for example, Nigeria has 664,999 EAs and Malawi has 12,569 EAs (NDHS, 2019; MDHS, 2017). The possible values of r_{samp} (1/3, 1, and 3) are chosen since DHS surveys sometimes have fewer than 400 clusters, such as in the 2010 Burundi DHS, which contains only 376 EAs (BDHS, 2012), and supplementing DHS data with other sources of data is possible and is an area of active research (Godwin and Wakefield, 2021; Burstein et al., 2018). Overall, there are $2 \times 3 \times 3 \times 3 = 54$ different combinations of parameters, and therefore 54 different scenarios.

Throughout this work, we set a joint penalized complexity (PC) prior (Simpson et al., 2017) on the spatial parameters (ρ, σ_S^2) based on Fuglstad et al. (2019) so that the median prior range is approximately a fifth of the spatial domain diameter, and so that $P(\sigma_S > 1) = 0.05$. We also set a PC prior on the cluster variance satisfying $P(\sigma_\epsilon > 1) = 0.05$. We use INLA's default priors on the fixed effects, placing an improper $N(0, \infty)$ prior on the intercept, and an uninformative $N(0, 1000)$ prior on the urban effect.

The response models are fit to the simulated surveys, and then used in conjunction with one of three aggregation models (smooth latent, latent, and empirical) to produce estimates of prevalence, burden, and relative prevalence. Posteriors of the quantities are produced at Admin1, Admin2, and Admin2 × urban/rural levels when defined, and compared with the true population using a number of scoring rules and metrics.

6.2. Areal prediction performance measures

We evaluate the aggregation model performance via continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), 95% interval score (Gneiting and Raftery, 2007), 95% fuzzy empirical CI coverage (Paige et al., 2022a), 95% fuzzy CI width (Paige et al., 2022a), and total computation time in minutes including fitting the response model to a single survey as well as generating all associated predictions using the aggregation model. CRPS is a strictly proper scoring rule, whereas interval scores are proper scoring rules (Gneiting and Raftery, 2007). For (strictly) proper scoring rules, the expected score of incorrect models are (strictly) larger than the expected score of the correct one, and lower scores are better. 'Fuzzy' CI's (Geyer and Meeden, 2005), are used to make empirical coverage more precise under discrete outcomes. When predicting prevalence, scores are produced on a probability scale rather than being for unnormalized population counts. Given n_a areas at a given areal level, we calculate a given score as the average of the individual area scores,

$$\begin{aligned}
\text{CRPS}(\boldsymbol{y}^*, F_1(y^*), \ldots, F_{n_a}(y^*)) &= \frac{1}{n_a} \sum_{i=1}^{n_a} \int_{-\infty}^{\infty} \Big(F(y) - I(y \geq y_i^*)\Big)^2 \, \mathrm{d}y \\
\text{Int}_\alpha(\boldsymbol{y}^*, l_1, u_1, \ldots, l_{n_a}, u_{n_a}) &= \frac{1}{n_a} \sum_{i=1}^{n_a} \Bigg[(u_i - l_i) + \frac{2}{\alpha}(l_i - y_i^*) I(y_i^* < l_i) \\
&\quad + \frac{2}{\alpha}(y_i^* - u_i) I(y_i^* > u_i)\Bigg], \qquad (6)
\end{aligned}$$

where $\boldsymbol{y}^* = (y_1^*, \ldots, y_{n_a}^*)$ is a vector of quantities we wish to estimate (a possibly normalized version of the response), $\hat{\boldsymbol{y}}^* = (\hat{y}_1^*, \ldots, \hat{y}_{n_a}^*)$ is the vector of associated model estimates, $F_1(y^*), \ldots, F_{n_a}(y^*)$ are the associated predictive distribution cumulative distribution functions, and $l_1, \ldots, l_{n_a}$ and $u_1, \ldots, u_{n_a}$ are the lower and upper ends respectively of the predictive distribution α-significance level CIs. We will take $\boldsymbol{y}^*$ to be the areal prevalences, burdens, or relative prevalences at the given areal level.

We also calculate relative scores. Given a proposed model, a reference model, and their respective scores S_P and S_R, we calculate the relative score $S_{\text{rel}} = 100\% \times (S_P - S_R)/S_R$. In particular, we will calculate the CRPS and 95% interval scores of the proposed empirical aggregation model relative to the smooth latent aggregation model when predicting Admin2 prevalence. We also provide other relative scores in the supplement in Section S.5.

Table 2
Mean percent change in CRPS of the empirical aggregation model relative to the smooth latent aggregation model, where the response is Admin2 ×stratum prevalence. Yellow–green values are better, while indigo values are worse. Results most representative of the application in Section 7 are outlined in red.

r_{pop}		1/5			1			5		
r_{samp}	β_0	1/9	1/3	1	1/9	1/3	1	1/9	1/3	1
3	0	-14.6	-12.1	-9.4	-7.6	-5.9	-4.3	-2.7	-1.9	-1.4
	-4	-17.1	-14.5	-11.5	-8.5	-6.7	-5.2	-2.9	-2.0	-1.4
1	0	-11.6	-9.0	-6.8	-5.3	-3.7	-2.4	-1.5	-0.8	-0.6
	-4	-12.6	-11.1	-9.0	-5.7	-4.4	-3.5	-1.5	-1.3	-0.8
1/3	0	-7.6	-5.7	-4.6	-2.7	-2.1	-1.4	-0.8	-0.4	-0.3
	-4	-8.9	-7.7	-5.8	-3.0	-2.5	-1.9	-0.6	-0.5	-0.3

Table 3
Mean percent change in 95% interval score of the empirical aggregation model relative to the smooth latent aggregation model, where the response is Admin2 ×stratum prevalence. Yellow–green values are better, while indigo values are worse. Results most representative of the application in Section 7 are outlined in red.

r_{pop}		1/5			1			5		
r_{samp}	β_0	1/9	1/3	1	1/9	1/3	1	1/9	1/3	1
3	0	-68.5	-62.9	-55.8	-49.9	-43.0	-34.3	-24.4	-18.4	-13.5
	-4	-68.4	-63.8	-56.8	-50.6	-43.5	-36.5	-24.8	-18.9	-13.1
1	0	-61.3	-54.5	-46.6	-39.5	-30.8	-22.5	-14.5	-9.7	-6.2
	-4	-58.1	-55.1	-47.9	-37.3	-31.4	-25.5	-13.6	-11.8	-7.8
1/3	0	-48.1	-41.3	-35.3	-23.9	-19.5	-13.6	-7.3	-4.5	-3.7
	-4	-44.5	-40.3	-33.4	-22.0	-18.2	-13.1	-5.0	-4.6	-2.7

6.3. Simulation study results

Tables 2 and 3 give the Admin2 × stratum CRPS and 95% interval relative scores respectively for prevalence, and relative scores for other target quantities and/or at other areal levels are given in the supplement in Section S.5. Overall, the empirical aggregation model performs as good as or better than the smooth latent model for nearly all considered scoring rules and parameters at all areal levels and for all target quantities. In particular, the empirical model performed best when making predictions in areas with small populations, as evidenced by its improvement in CRPS and 95% interval score being largest when $r_{\text{pop}} = 1/5$.

In addition, the relative performance of the empirical aggregation model improved as the predictive spatial variance decreased relative to the nugget variance, such as when the sample size, r_{sample}, increased, and when the signal to noise ratio, φ, decreased from 1 to 1/9. This is expected, since under these circumstances, the sources of aggregation error will have large variance relative to the predictive variance of the smooth latent model, particularly fine scale and finite population variability.

Differences between the smooth latent and empirical aggregation models were nearly always larger for the interval score than for CRPS, which is expected since their central predictions are identical for prevalence and burden. The largest improvement in scores from smooth latent to empirical sampling frame models consistently occurred when $r_{\text{samp}} = 3$, $r_{\text{pop}} = 1/5$, and $\varphi = 1/9$. This is because predictive spatial variability is minimized when $r_{\text{samp}} = 3$ (due to having more data in the survey), populations are smallest when $r_{\text{pop}} = 1/5$ (resulting in less fine scale and finite population being averaged out), and fine scale variability is highest compared with smooth spatial variability when $\varphi = 1/9$ (resulting in larger fine scale variability). For example, the smallest (most negative) relative 95% interval score, −79.2%, occurred under these conditions, and when predicting Admin2 relative prevalence for $\beta_0 = -4$. Hence, the empirical model 95% interval score was nearly

80% better than that of the smooth latent due to its ability to account for aggregation uncertainty. The equivalent relative CRPS was −27.1%, indicating a substantial but smaller improvement from the smooth latent to the empirical aggregation model.

Relative scores for parameters most similar to the application indicate smaller but clear performance gain, particularly at the Admin2 and Admin2 × urban/rural levels. For example, we observe relative 95% interval (and relative CRPS) scores of −31.4% (−4.4%), −4.9% (−0.5%), and −0.7% (−0.1%) at the Admin2 × stratum, Admin2, and Admin1 levels respectively when predicting prevalence. Relative scores for relative prevalence are typically more negative than for prevalence and burden, indicating a larger improvement. That is particularly the case for CRPS, since the central prediction for relative prevalence of the three considered aggregation models differ.

7. Application to 2014 KDHS neonatal mortality

We apply the empirical aggregation model, in conjunction with the response model introduced in Section 5, to 2014 KDHS 2010–2014 NMR data. Unlike for the simulation study, we use an integration grid with 5 km rather than 25 km resolution for generating areal predictions, since we only need to apply the model to 1 survey rather than 5,400. In addition to generating estimates and 95% CIs for the prevalence, burden, and relative prevalence at the Admin1, Admin2, and Admin2 × urban/rural levels when defined, we also generate smooth risk estimates and CIs at the 5 km pixel level.

It is important to acknowledge that pixel level smooth risk and smooth burden uncertainties are underreported since they do not account for sources of aggregation error that are especially important at such small spatial scales. While pixel level prediction of the latent and empirical aggregation models is technically possible, their interpretation breaks down at the pixel level due to the fact that EA boundaries can span multiple pixels. In addition, burden estimates are inherently difficult due to the unknown level of uncertainty in the population totals, and the challenge of validation at the areal level when using DHS data since reported DHS survey weights are normalized.

The response model parameter central estimates and other summary statistics are given in Fig. 4. Central estimates of prevalence along with CIs for EAs are given at the 5 km pixel, Admin2, and Admin1 level in Fig. 5, whereas predictions and CIs for burden and relative prevalence are given in Figure F.3 and F.4 respectively in the supplement in Section S.6.

Fig. 4 gives little evidence of an urban effect, and the estimated nugget variance is 0.44, which is over twice the spatial variance of 0.20. The estimated spatial range of 411 km is somewhat high compared to the diameter of the spatial domain, which is approximately 1,169 km. The long spatial range indicates a large degree of spatial smoothing and a decrease in the spatial variance in the predictive distribution due to increased 'borrowing of strength' from observations across long distances.

We also find evidence that the sampling frame models make a difference when producing NMR estimates. Fig. 6 shows how the relative posterior standard deviation (SD) of the empirical to the smooth latent model varies with the number of EAs in an area, calculated as 100% × $(\widehat{\mathrm{SD}}(p_{\mathrm{emp}}(A)) - \widehat{\mathrm{SD}}(r_{\mathrm{smooth}}(A)))/\ \widehat{\mathrm{SD}}(r_{\mathrm{smooth}}(A))$ for counties, constituencies, and the urban/rural portion of constituencies. The relative standard deviations in Fig. 6 are averaged within each area level and each quantity of interest in Table 4, which shows that the empirical model results in 8% larger uncertainties in prevalence at the Admin2 level, and 31% larger uncertainties at the Admin2 × stratum level. The associated relative uncertainties tend to be larger for burden and for relative prevalence.

As in the simulation study, we find that the relative SD increases as the size of the area, in terms of the number of EAs, decreases. We find the relationship to be approximately linear on a log–log scale for both prevalence and burden, although the relationship is less clear for relative prevalence. Areas with fewer than 300 EAs in this application are especially impacted by the sampling frame model, with posterior SDs that are typically 10 percent higher or more for the empirical aggregation model as opposed to the smooth latent aggregation model. We therefore recommend using the empirical aggregation model for areas finer than the Admin2 level, and for areas with fewer than 300 EAs for this application when estimating prevalence or burden, and we recommend using the

Table 4
Mean percent increase in posterior standard deviation (SD) of the empirical versus smooth latent aggregation models for prevalence, burden, and relative prevalence and for each considered areal level.

Area level	*Mean percent increase SD*		
	Prevalence	Burden	Relative prevalence
Constituency ×stratum	31	44	–
Constituency	8	13	132
County	2	2	21

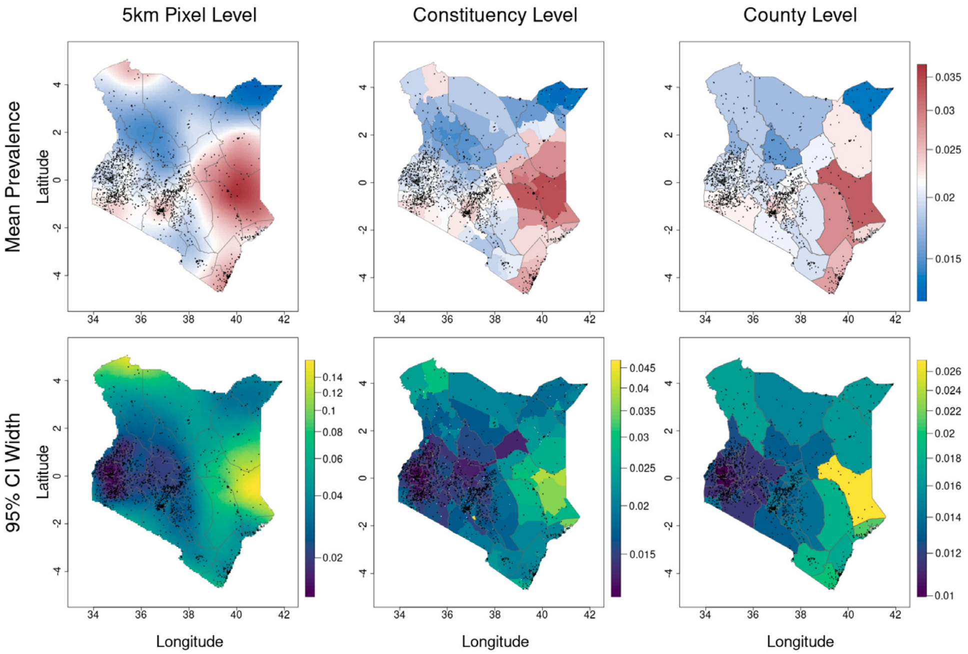


Fig. 5. Central predictions (top row) and 95% credible interval widths (bottom row) of neonatal mortality rates in Kenya in 2010–2014. Observation locations are plotted as black dots, provinces as thick black lines, and counties as thin gray lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

empirical aggregation model when estimating relative prevalence even at the Admin1 level for the considered application. The sampling frame model used matters less, however, when estimating prevalence and burden at the Admin1 level in this case.

The urban and rural parts of constituencies with a very small number of average EAs tend to have very high relative SD. For example several Admin2 × stratum areas have relative SDs of over 500% when predicting prevalence, and one such area has a relative SD of over 1000% when predicting burden. This highlights just how much of a difference the sampling frame model can make in some areas, regardless of their average behavior. Moreover, this also indicates that the smooth latent model would likely produce estimates that drastically underestimate uncertainty in very small areas, such as those with fewer than approximately 30 EAs in this application.

8. Sensitivity to population information

The analysis in Section 7 depends on a combination of population density and census information. WorldPop population density estimates in Kenya have no estimated uncertainty, we will see

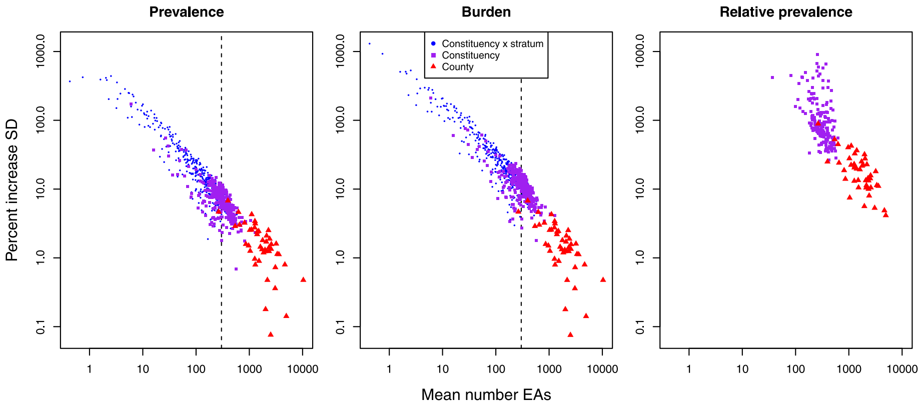


Fig. 6. Percent increase in posterior standard deviation (SD) of the empirical versus smooth latent models for prevalence (left), burden (middle), and relative prevalence (right). Dashed black lines at 300 mean EAs give a simple threshold for when aggregation error is significant.

that population density estimates from Meta differ considerably from those of WorldPop. Censuses typically occur no more often than once every decade, and they may therefore be outdated for the time period being analyzed. If one wishes to estimate finite population quantities, however, such potentially imprecise population information is necessary, and finite population quantities are often desired. It is therefore important to assess how robust model predictions and uncertainties are to such population information.

In order to assess how the considered aggregation models' posterior means and SDs are influenced by population information, we perform a sensitivity analysis by observing how they change under a number of scenarios. First, we test whether results are influenced by changes in population information at fine spatial scales using population density estimates by Meta (Tiecke et al., 2017) in lieu of the WorldPop population density data. Second, we test whether results are robust to systematic changes in population information using Kenya's latest census, conducted in 2019 (Kenya National Bureau of Statistics, 2019), instead of the 2009 census. Lastly, we test how results are impacted by asystematic errors in census EA and household totals by considering a scenario where there is 5% error in the 2009 census EAs and households per stratum. We included this last scenario in order to assess how robust the empirical and latent models are to their assumptions on the distributions of EAs and households.

8.1. Setup

Using the original 2009 census information, we adjust Meta's population density estimates by a scalar factor in each to match the population totals in each stratum in what we call the 'Meta' scenario. Since Meta's population density estimates are uniformly zero in 3 Admin2 areas with GADM names Turkana East in Turkana County, Unknown 1 in Mandera County, and Unknown 7 in Machakos County, we set the population density in those to be uniform in those areas before adjusting them to match the census stratum totals. The resulting adjusted version of Meta's population density estimates is given in Fig. 7 along with its ratio to those of WorldPop. After the adjustments, we rerun the analysis from Section 7 with the resulting population density estimates.

To assess how sensitive our results in Section 7 are to the census data, we use 2019 census information instead of the 2009 census in what we call the '2019 census' scenario. However, the number of EAs per stratum was not available, and we instead divided the 129,123 total EAs in the 2019 census among the strata proportionally to the 2009 census, which sampled from only 96,251 EAs in total. After adjusting WorldPop population density estimates based on the 2019 Census

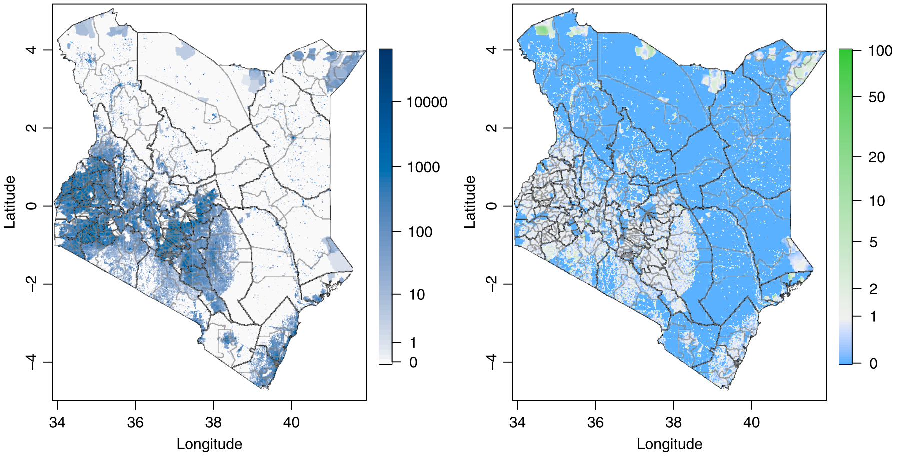


Fig. 7. Meta population density estimate (left) and its ratio over WorldPop (right) after adjusting the population densities to match the estimated population totals.

population totals per stratum, we reran the analysis in Section 7. The total Kenyan population changed from approximately 43 to 48 million people after the adjustment.

It is important to note that, if population information changes systematically, such as from the 2009 to 2019 census, predictions and uncertainties *should* be different. Hence, the 2019 census scenario is specifically intended to check how census information can influence predictions. In order to assess how sensitive the empirical and latent model distributional assumptions on EAs and households are, we consider one additional scenario where census EA and household totals per stratum are randomly varied by and uniformly distributed percentage between −5% and 5%, while holding the population per stratum the same. This ensures that predictions and smooth latent model uncertainties stay the same, while influencing empirical and latent SDs. We then combine the resulting census information with WorldPop population density information, and rerun the analysis of Section 7. We call this the 'jittered census' scenario.

For each scenario, areal level, model, and outcome, we calculate the mean absolute percent difference (MAPD) relative to the 'standard' scenario (i.e. the original data), calculated as:

$$\mathrm{MAPD}(\boldsymbol{x}_{\mathrm{scen}}) = 100\% \cdot \frac{1}{N_A} \sum_{i=1}^{N_A} \left| \frac{(x_{\mathrm{scen}})_i - (x_{\mathrm{stand}})_i}{(x_{\mathrm{stand}})_i} \right|,$$

where $\boldsymbol{x}_{\mathrm{scen}}$ is a vector of length N_A, the number of areas at the specified level, of either posterior means or SDs for the given scenario, aggregation model, and outcome, and $\boldsymbol{x}_{\mathrm{stand}} = ((x_{\mathrm{stand}})_1 \; \dots \; (x_{\mathrm{stand}})_{N_A})^T$ is the equivalent vector for the standard scenario.

While the response model does not change among the scenarios, the new population information does change the sampling frame models, and the aggregation models formed by linking them with the response model. Their aggregate predictions and uncertainties may change as a result.

8.2. Results

Table 5 shows MAPD in the prevalence estimate and SD for each scenario, area level and model outcome, while Tables G.25 and G.26 in the supplement show equivalent results for burden and relative prevalence respectively.

We find that, for Kenya 2010–2014 NMR estimates, the proposed model's prevalence and even relative prevalence estimates are fairly robust to changes in population information, whereas we find burden estimates to be robust at the Admin1 level under non-systematic changes in

Table 5
Mean absolute percent difference of prevalence posterior mean and SD in sensitivity analysis scenarios compared to with the original data.

Scenario	Area level	*Mean*	*Standard deviation*		
		All models	Smooth latent	Latent	Empirical
Meta	Constituency ×stratum	0.8	2.9	7.1	10.7
	Constituency	0.6	3.0	3.5	4.1
	County	0.9	2.7	2.6	2.4
2019 census	Constituency ×stratum	0.6	2.9	5.3	8.8
	Constituency	0.5	2.9	2.9	2.9
	County	0.4	2.7	2.7	2.7
Jittered census	Constituency ×stratum	0.4	2.3	2.5	2.3
	Constituency	0.4	2.4	2.4	2.2
	County	0.4	2.4	2.5	2.4

population information. In cases where burden estimates and their SDs are particularly influenced by population information, they are influenced for all considered sampling frame models, including the smooth latent model. Such changes are mainly due to large percentage changes in small population totals as a result of large systematic or fine scale changes in populations.

Changes in population information will invariably affect areal population, household, and EA totals—this is unavoidable in a geostatistical framework. Yet we find that estimates and estimate SDs are not substantially influenced for any models or outcome, including burden, under moderate sized asystematic errors in population information as in the jittered census scenario. For example, MAPD of central estimates was at or below 0.4% for al outcomes and areal levels, while the MAPDs of the SDs ranged between 2.2% and 2.5% over all models and area levels for prevalence and burden, and was between 3.8% and 6.3% for relative prevalence, with, perhaps surprisingly, the largest and smallest relative prevalence MAPDs occurring for the smooth latent and empirical models respectively. The results therefore suggests that all considered sampling frame models are fairly robust to asystematic errors in household and EA totals, despite their explicit use in the latent and empirical sampling frame models.

Changes in SDs were typically largest, as measured by MAPD, for the smallest areas. This is expected, since areas with small population and EA totals have the least precise estimates, and so will be more affected by changes in the input population information.

9. Discussion and conclusions

Typical geostatistical workflows for spatial aggregation with respect to a population distribution involve the ad-hoc aggregation of point-referenced predictions to the areal level. We propose including a *sampling frame model* to such workflows. The combination of a *response model* for the data and a sampling frame model that incorporates uncertainty in the distribution of the population results in what we call a *spatial aggregation model*. By explicitly incorporating a sampling frame model that includes uncertainty about the population, aggregation models can account for three major sources of aggregation error: aggregation weights, fine scale variation, and finite population variation. Including a sampling frame model also makes more transparent which types of aggregation error are accounted for.

The three main considered sampling frame models are the smooth latent model, the latent model, and the empirical model, which is the model that we propose. The smooth latent model integrates out the spatial nugget effect from the point level risk, creating a smoother risk surface that is integrated with respect to population density. The latent and empirical aggregation models, in contrast, explicitly model how the population is distributed among the enumeration areas (EAs) along with EA locations. This allows them to account for fine scale variability due to EA level effects. The empirical model also accounts for finite population variation by modeling individual outcomes. The empirical model is therefore the only considered sampling frame model that fully accounts for all three major sources of aggregation error.

We have shown that aggregation uncertainty can substantially influence prediction uncertainty in some cases, and so the aggregation procedure should be considered carefully. When the nugget effect in the sampling frame model does not correspond well to that of the response model such as in the 'gridded risk' sampling frame model, where a single nugget is included at each spatial aggregation grid point, we have shown that aggregation error can cause predictions to lack robustness to the aggregation grid. If all parameters in the sampling frame model correspond well with those of the spatial response model, the predictions are more robust to the choice of the integration grid, and the grid resolution can be reduced, improving computational performance.

Despite accounting for more sources of uncertainty, the empirical and latent aggregation models required an average of 12 min, while the smooth latent model required 44 min on average to perform the aggregation in the application and the three scenarios considered in the sensitivity analysis. This speedup is due almost entirely to the empirical and latent models not needing to numerically integrate the nugget effect out of the risk, unlike the smooth risk model.

The differences in the considered spatial aggregation models highlights the difference between risk and prevalence, where risk is an expected prevalence, and therefore has less uncertainty. In small area estimation we are ultimately interested in information about an existing population, so we advocate for inference on prevalence. Inference on prevalence results in more conservative uncertainty estimates due to the additional variation of prevalence when compared to risk.

A potential benefit of inference on prevalence is that prevalence is observable while risk is not, making models for prevalence potentially able to be validated more directly. Validating estimates of population averages based on point referenced data is an unsolved problem, but aggregation error may be an essential part of the solution, particularly when it comes to predicting averages of small, left out portions of the data.

While we only considered the effects of aggregation uncertainty in the context of geospatial models, it is possible fine scale and finite population variation may be equally or more important for models with areal spatial effects, since reduced flexibility in the spatial effect could lead to relatively more variance in the nugget. Extending the empirical aggregation model to space–time is also possible, although incorporating multiple age bands in space–time could add complication due to needing to incorporate time and age information to the simulated population in the sampling frame model. Fixing the number of target individuals per EA in each age band and time period to an expected value rather than simulating counts through time could simplify such an extension.

Extending the model to include more geospatial covariates would also be useful if done carefully. In particular, geospatial covariates that change significantly in a stratum could influence prediction uncertainty in the context of EA location uncertainty. Although we do not focus on EA location uncertainty here since we do not consider any spatial variables changing within strata, it is possible it could be a major source of aggregation error when accounting for more geospatial covariates. However, including more covariates is an open question, and is complicated by the fact that positional uncertainty in the survey data makes interpretation of their effect sizes problematic. While urbanicity can be interpreted as a spatial variable and is also available at the sampled clusters, with population and EA totals available in each urban/rural × Admin1 area stratum, other covariates used in this context lack this luxury.

It is worth mentioning that accounting for survey designs and design weights in geospatial models and their estimates of population quantities is a difficult problem that we do not address. In fact, even in classical design-weighted estimators, estimating burden is nontrivial in some instances, such as the considered application, due to the renormalization of the reported design weights. While our proposed model does not preclude weights from being incorporated into the fitting of the response model, it does model the population and sampling frame explicitly rather than incorporating design weights directly into the aggregation.

Ultimately, we find that aggregation uncertainty increases as areas decrease in population size. We observe that aggregation uncertainty is sometimes larger for burden than prevalence due to uncertainty in population totals, and it is often even larger when estimating the relative prevalence between two strata within an area. As a result, we recommend using the empirical or latent aggregation models for small areas, particularly areas smaller than the Admin2 level, and the smooth latent aggregation model for larger areas, such as at the Admin1 level, due to it being easier to implement. When estimating relative prevalence in the urban versus rural part of an area, we recommend using the empirical aggregation model even at the Admin1 level, or the latent model when estimating risk.

In a sensitivity analysis we show that prevalence estimates and uncertainties are much more robust to changes in population information than burden for all considered aggregation models. Burden estimates were sensitive to changes in population densities and systematic changes in population totals, while relative prevalence estimates were less so, and prevalence estimates were especially robust. We therefore urge caution when interpreting burden estimates. Moderately sized asystematic errors in stratum EA and household totals did not substantially effect the estimates and uncertainties of prevalence, burden, or relative prevalence for areas at least the population of the urban or rural parts of Admin2 areas.

CRediT authorship contribution statement

John Paige: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft, Writing - review and editing. **Geir-Arne Fuglstad:** Conceptualization, Writing - review and editing, Supervision. **Andrea Riebler:** Conceptualization, Writing - review and editing, Supervision. **Jon Wakefield:** Conceptualization, Writing - review and editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

DHS survey data can be freely obtained at https://dhsprogram.com/, and WorldPop and Meta population density data can be freely obtained at https://www.worldpop.org/ and https://dataforgood.facebook.com/dfg/. The postprocessed population density data as well as all code directly used to produce all results in this analysis is available on JP's Github at https://github.com/paigejo/continuousNugget and at https://github.com/paigejo/continuousNugget respectively. The code also uses functions that will soon be incorporated into the SUMMER R package (Li et al., 2021), available at https://github.com/paigejo/SUMMER. Kenya administrative area shapefiles are available at GADM's website at https://github.com/paigejo/SUMMER.

Acknowledgments

This work would not be possible without the financial support of the Research Council of Norway or the National Institutes of Health. They played no role in the formulation of this manuscript, however, and the views expressed in this manuscript are solely those of the authors.

Funding

GF and AR were supported by project number 240873 from the Research Council of Norway. JW was supported by the National Institutes of Health, USA under award R01 AI029168-30.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2022.100714.

References

Ballmann, A.E., Torkelson, M.R., Bohuski, E.A., Russell, R.E., Blehert, D.S., 2017. Dispersal hazards of pseudogymnoascus destructans by bats and human activity at hibernacula in summer. J. Wildl. Dis. 53, 725–735.

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

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 209–226.

Diggle, P., Giorgi, E., 2016. Model-based geostatistics for prevalence mapping in low-resource settings. J. Amer. Statist. Assoc. 111, 1096–1120.

Dong, T.Q., Wakefield, J., 2021. Space-time smoothing models for subnational measles routine immunization coverage estimation with complex survey data. Ann. Appl. Stat. 15, 1959–1979.

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., et al., 2019. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. Nature 570, 189–193.

Fuglstad, G.-A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. J. Amer. Statist. Assoc. 114, 445–452.

Funwi-Gabga, N., Mateu, J., 2012. Understanding the nesting spatial behaviour of gorillas in the Kagwene Sanctuary, Cameroon. Stoch. Environ. Res. Risk Assess. 26, 793–811.

Gehlke, C.E., Biehl, K., 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material. J. Amer. Statist. Assoc. 29 (185A), 169–170.

Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M. (Eds.), 2010. Handbook of Spatial Statistics. CRC Press, Boca Raton, FL.

Geyer, C.J., Meeden, G.D., 2005. Fuzzy and randomized confidence intervals and p-values. Statist. Sci. 20, 358–366.

Giorgi, E., Diggle, P.J., Snow, R.W., Noor, A.M., 2018. Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys. Internat. Statist. Rev. 86, 571–597.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, 359–378.

Godwin, J., Wakefield, J., 2021. Space-time modeling of child mortality at the admin-2 level in a low and middle income countries context. Stat. Med. 40, 1593–1638.

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

Institut de Statistiques et d'Études Économiques du Burundi (ISTEEBU), Ministère de la Santé Publique et de la Lutte contre le Sida [Burundi] (MSPLS), et ICF International, 2012. Enquête démographique et de santé burundi 2010. Technical Report, ISTEEBU, MSPLS, et ICF International, Bujumbura, Burundi.

Keller, J.P., Peng, R.D., 2019. Error in estimating area-level air pollution exposures for epidemiology. Environmetrics 30, e2573.

Kenya National Bureau of Statistics, 2019. The 2019 Kenya Population and Housing Census Volume II: Distribution of Population By Administrative Units. Kenya National Bureau of Statistics, Nairobi.

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, and National Council For Population And Development/Kenya, 2009. The 2009 Kenya Population and Housing Census Volume IC: Population Distribution by Age, Sex, and Administrative Units. Kenya National Bureau of Statistics, Nairobi.

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, and National Council For Population And Development/Kenya, 2015, 2014. Kenya demographic and health survey 2014. Rockville, Maryland, USA. http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

Laber, E.B., Meyer, N.J., Reich, B.J., Pacifici, K., Collazo, J.A., Drake, J.M., 2018. Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. J. R. Stat. Soc. Ser. C Appl. Stat. 67, 743–789.

Lessler, J., Moore, S.M., Luquero, F.J., McKay, H.S., Grais, R., Henkens, M., Mengel, M., Dunoyer, J., M'bangombe, M., Lee, E.C., et al., 2018. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. Lancet 391, 1908–1915.

Leyk, S., Gaughan, A.E., Adamo, S.B., de Sherbinin, A., Balk, D., Freire, S., Rose, A., Stevens, F.R., Blankespoor, B., Frye, C., et al., 2019. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth Syst. Sci. Data 11, 1385–1409.

Li, Z.R., Martin, B.D., Hsiao, Y., Godwin, J., Paige, J., Wakefield, J., Clark, S.J., Fuglstad, G.-A., Riebler, A., 2021. SUMMER: Small-area-estimation unit/area models and methods for estimation in R. R package version 1.2.0. URL https://github.com/richardli/SUMMER.

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). J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 423–498.

Local Burden of Disease Child Growth Failure Collaborators, et al., 2020. Mapping child growth failure across low-and middle-income countries. Nature 577, 231–234.

Local Burden of Disease Vaccine Coverage Collaborators, et al., 2021. Mapping routine measles vaccination in low-and middle-income countries. Nature 589, 415–419.

Manley, D., 2014. Scale, aggregation, and the modifiable areal unit problem. In: Fischer MM, N.P. (Ed.), The Handbook of Regional Science. Springer Berlin, Germany, pp. 1157–1171.

Matérn, B., 1986. Spatial Variation, Second ed. Springer-Verlag, Berlin..

National Population COmission (NPC) [Nigeria] and ICF, 2019. Nigeria demographic and health survey 2018. Technical Report, NPC and ICF, Abuja, Nigeria and Rockville, Maryland, USA.

National Statistical Office (NSO) [Malawi] and ICF, 2017. Malawi demographic and health survey 2015-16. Technical Report, NSO and ICF, Zomba, Malawi, and Rockville, Maryland, USA.

Nguyen, H., Cressie, N., Braverman, A., 2012. Spatial statistical data fusion for remote sensing applications. J. Amer. Statist. Assoc. 107, 1004–1018.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Statist. 24, 579–599.

Openshaw, S., Taylor, P.J., 1979. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley, N. (Ed.), Statistical Applications and the Spatial Sciences, twenty first ed. Pion, pp. 127–144.

Osgood-Zimmerman, A., Millear, A.I., Stubbs, R.W., Shields, C., Pickering, B.V., Earl, L., Graetz, N., Kinyoki, D.K., Ray, S.E., Bhatt, S., Browne, A., Burstein, R., Cameron, E., Casey, D., Deshpande, A., Fullman, N., Gething, P., Gibson, H., Henry, N., Herrero, M., Krause, L., Letourneau, I., Levine, A., Liu, P., Longbottom, J., Mayala, B., Mosser, J., Noor, A., Pigott, D., Piwoz, E., Rao, P., Rawat, R., Reiner, R., Smith, D., Weiss, D., Wiens, K., Mokdad, A., S.S., L., Murray, C., Kassebaum, N., Hay, S., 2018. Mapping child growth failure in Africa between 2000 and 2015. Nature 555, 41–47.

Paige, J., Fuglstad, G.-A., Riebler, A., Wakefield, J., 2022a. Bayesian multiresolution modeling of georeferenced data: An extension of 'LatticeKrig'. Comput. Statist. Data Anal. 173, 107503.

Paige, J., Fuglstad, G.-A., Riebler, A., Wakefield, J., 2022b. Design- and model-based approaches to small-area estimation in a low and middle income country context: Comparisons and recommendations. J. Surv. Stat. Methodol. 10, 50–80.

Riebler, A., Sørbye, S., Simpson, D., Rue, H., 2016. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat. Methods Med. Res. 25, 1145–1165.

Simpson, D., Rue, H., Riebler, A., Martins, T., Sørbye, S., 2017. Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). Statist. Sci. 32, 1–28.

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.

Tatem, A.J., 2017. WorldPop, open data for spatial demography. Sci. Data 4, 1–4.

Tiecke, T.G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E.B., et al., 2017. Mapping the world population one building at a time. arXiv Preprint arXiv:1712.05839.

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.

Zammit-Mangion, A., Cressie, N., 2017. FRK: An R Package for Spatial and Spatio-Temporal Prediction with Large Datasets. Technical Report, National Institute for Applied Statistics Research Australia.