Design- and Model-Based Approaches to Small-Area Estimation in a Low and Middle Income Country Context: Comparisons and Recommendations

John Paige, Geir-Arne Fuglstad, Andrea Riebler, Jon Wakefield*

Abstract

The need for rigorous and timely health and demographic summaries has provided the impetus for an explosion in geographic studies in low and middle income countries. Many of these studies present fine-scale pixel-level maps in an attempt to answer the needs of the current era of precision public health. However, even though household surveys with a two-stage cluster design stratified by region and urbanicity are a major source of data, cavalier approaches are taken to acknowledging the survey design.

We investigate the extent to which accounting for the sample design affects the predictive performance at the aggregate level of interest for health policy decisions. We consider various commonly-used models and introduce a new Bayesian cluster level model with a discrete spatial smoothing prior. The investigation is performed through a simulation study in which realistic sampling frames are created for Kenya, based on population and demographic information,

*John Paige was supported by The National Science Foundation Graduate Research Fellowship Program under award DGE-1256082, and Jon Wakefield was supported by the National Institutes of Health under award R01AI029168.

with a survey design that mimics a Demographic Health Survey (DHS). We find that including stratification and cluster level random effects can improve predictive performance. Spatially smoothed direct (weighted) estimates and area level models accounting for stratification were robust to the underlying population and survey design. Continuous spatial models showed some promise in the presence of fine scale variation; however, these models require the most "hand holding". Subsequently, we examine how the models perform on real data, estimating the prevalence of secondary education for women aged 20–29 and neonatal mortality rates, using data from the 2014 Kenya DHS.

KEY WORDS: Survey design; spatial statistics; small area estimation; integrated nested Laplace approximations; geostatistical models.

1 Introduction

Complex, multi-stage household surveys play an important role in producing a variety of estimates of health and demographic quantities of interest, especially in low and middle income countries (LMICs). Examples of surveys in this context include Demographic Health Surveys (DHS) (USAID, 2019), Multiple Indicator Cluster Surveys (MICS) (UNICEF - Statistics and Monitoring, 2012), AIDS Indicator Surveys (AIS) (DHS Program, 2019), and Living Standard Measurement Surveys (LSMS) (World Bank, 2019). The lack of high quality vital registration (VR) and administrative data often necessitates the use of these household surveys in LMICs (Li et al., 2019; Wagner et al., 2018; Sandefur and Glassman, 2015; Jerven, 2013; Devarajan, 2013), particularly in the context of precision public health, which is the practice of using data to guide public health interventions targeting specific subgroups or entire populations more effectively (Khoury et al., 2016; Desmond-Hellmann et al., 2016; Dowell et al., 2016). For

instance, it has been estimated that only 4% of neonatal deaths (deaths in the first 28 days of life) are recorded via high quality VR data (Lawn et al., 2014). In addition, in a comparison across 46 surveys and 21 African countries, growth in primary education enrollment from 1991 to 2011 was found to be higher in administrative statistics by a third on average relative to DHS data, with the largest discrepancies occurring in Kenya and Rwanda (Sandefur and Glassman, 2015).

The Sustainable Development Goals (SDGs) specify targets for a variety of health and demographic outcomes (United Nations, 2019). In particular, SDG 3 calls for an end to preventable deaths of newborns and children under 5 years of age and states that all countries should aim to reduce neonatal mortality to below 12 deaths per 1,000 live births by 2030. Additionally, SDG 4 calls for improved education for all, for all people to complete their secondary education, and for the elimination of inequalities in education due to gender or location. Hence, developing statistical models that can accurately account for the sampling design while also producing spatial estimates at subnational scales is of great importance.

Classical techniques for analyzing survey data that can account for the survey design often have difficulty producing estimates at the required spatial resolutions (interventions are often made at the Admin2 level, which is the second subnational area level commonly referred to as counties). For example, weighted (direct) estimates from DHS data are often practically unavailable at the Admin2 level, as there are areas with no data, or areas with data that produce estimates with unacceptably large variances. To improve estimation in such situations, a number of small area estimation (SAE) methods have been proposed (Rao and Molina, 2015) including those that extend upon the seminal Fay-Herriot model (Fay and Herriot, 1979). These methods not only acknowledge the survey design, but also "borrow strength" from data in nearby areas to produce reliable estimates with smaller uncertainty intervals (Marhuenda et al.,

2013; Chen et al., 2014; Mercer et al., 2015; Congdon and Lloyd, 2010; You and Zhou, 2011; Porter et al., 2014; Vandendijck et al., 2016; Watjou et al., 2017; Li et al., 2019). These approaches are all based on discrete spatial models, which employ arbitrary neighborhood structures that may be more or less realistic depending on the context and geography.

A number of papers have used continuous spatial models to analyze health and demographic outcomes using survey data (Wardrop et al., 2018; Gething et al., 2016; Golding et al., 2017; Utazi et al., 2018; Gething et al., 2015; Osgood-Zimmerman et al., 2018; Graetz et al., 2018; Diggle and Giorgi, 2016; Giorgi et al., 2018; Diggle and Giorgi, 2019). Included in this list are publications from WorldPop and the Institute for Health Metrics and Evaluation (IHME), both of which are large scale producers of health and demographic maps. In these references, all of the models ignore the stratification when analyzing survey data, with WorldPop also routinely ignoring the cluster sampling as well. In general, ignoring the design results in biased estimates and inaccurate uncertainty intervals. No study has been conducted to explore the effects of ignoring design stratification and cluster level correlation in the LMIC context, and here we aim to fill this gap in the literature, by comparing a variety of spatial modeling approaches, under different levels of stratification and clustering.

We explore the performance of different design- and model-based methods, when applied to simulated data. We also apply these methods to the analysis of two outcomes recorded in the 2014 Kenya DHS, the proportion of women between the ages of 20 and 29 who have completed their secondary education, and the neonatal mortality rate (NMR). Section 2 describes the data we will use in this analysis and Section 3 introduces the models that we compare and apply. Section 4 describes the simulation study, and in Section 5 we apply the models to the secondary education outcome and NMRs, reporting predictions, uncertainties, and using cross validation to assess the

out of sample performance of each of the models. We discuss the results and provide our conclusions in Section 6. Lastly, Appendix A gives details on how we aggregate model-based predictions to the county level.

2 Data

The DHS Program uses a consistent set of sampling approaches, with methods described in the 2012 DHS Sampling and Household Listing Manual (ICF International, 2012, Sec. 5.2, p. 80–85). The standard design is a stratified two-stage cluster sampling scheme with stratification by county crossed with urban/rural classification. The first sampling stage involves selecting enumeration areas (EAs) with probability proportional to size (PPS) sampling, where the probability of sampling each EA is proportional to the listed number of households in that EA, and the second stage involves simple random sampling of (typically) 25 households within each EA. Women within the household are then asked a number of demographic questions involving their education level, and mothers give information on their children's birth dates, and, if relevant, dates of death. The 2014 Kenya DHS (KDHS, 2014) follows the typical DHS scheme, though it is powered to the county level, which is the level at which policies are implemented in Kenya. Hence, sample sizes are chosen to provide sufficiently precise county level estimates. A total of 1,612 clusters were sampled out of the 96,251 total EAs that were in the 2009 Kenya Population and Housing Census (Kenya National Bureau Of Statistics, 2014). Of these clusters, 995 are urban while 617 are rural, with urban areas oversampled in the majority of the 47 counties. Mombasa and Nairobi are entirely urban and the remaining 45 counties have both urban and rural areas, so that there are 92 strata in total.

3 Methods

3.1 Models

We describe the notation for NMRs, but the models are applicable to arbitrary binary outcomes. In the case of NMR, the denominators are the number of children that were born in the relevant period, and the response is whether a death occurred in the first month after birth. Let $Y_{ck} = 0/1$ represent the binary response for child k in cluster c with the total number of deaths in each cluster being $Y_c = \sum_{k \in \mathcal{B}_c} Y_{ck}$ where $\mathcal{B}_c$ is the set of indices of the children in cluster c that are sampled. We let $\boldsymbol{x}_c$ be the spatial location of cluster c. Associated with location $\boldsymbol{x}_c$ is a county, which will be denoted $i[c]$, and the set of spatial locations that are urban is denoted U. We now describe the different models considered, focusing on inference at the county level since this is often relevant for public health policy decisions.

Naive: We fit a binomial model to the county-level data without accounting for the sampling design. In this case, we assume the probability of mortality for child k in cluster c is the same for all children in county i, and define the county specific logit probability as,

$$\log\left(\frac{p_{ck}}{1-p_{ck}}\right) = \beta_{i[c]},$$

where the models are fitted independently to the data from each county. The targets of inference are the county-level probabilities $\text{expit}(\beta_i)$, $i = 1, \ldots, 47$.

Direct: County-level direct estimates, $\widehat{p}_i^{\,\text{DIR}}$, are calculated using a weighted estimator that accounts for the survey design. The weights are proportional to the inverse probability of sampling each child. This estimator is reliable for large samples, but for small samples, will have high variance (Rao and Molina, 2015). Weighted estimators can yield estimates of exactly zero or one, and have variance instability in small samples.

These problems lead to yearly estimates at the Admin2 level that are typically not reliable when based on a DHS with around 400 clusters (a typical design).

In addition to these two classical approaches we also consider three hierarchical Bayesian space-time models.

Smoothed Direct: Following the approach of Mercer et al. (2015) we calculate $Z_i = \text{logit}(\widehat{p}_i^{\text{DIR}})$ along with its associated design-based variance estimate, $\widehat{V}_i$. We assume $Z_i|\eta_i \sim_{ind} N(\eta_i, \widehat{V}_i)$ with linear predictor,

$$\eta_i = \beta_0 + \frac{1}{\sqrt{\tau}}(\sqrt{\phi}S_i + \sqrt{1-\phi}\delta_i),$$

where β_0 is the intercept, and $S_1, \ldots, S_{47}$ and $\delta_1, \ldots, \delta_{47}$ are, respectively, mean zero county level intrinsic conditional autoregressive (ICAR) spatial terms and independent and identically distributed (iid) Gaussian random effects. The ICAR model, described in Besag et al. (1991), is a discrete spatial model that assumes the latent effect in each area is Gaussian whose mean is the average of the effects in neighboring areas. We apply a sum-to-zero constraint $\sum_{i=1}^{47} S_i = 0$ to the ICAR terms to make the intercept β_0 identifiable.

The parameterization adopted is a variation of the model introduced in Simpson et al. (2017), and is named the BYM2 model in Riebler et al. (2016) since it is a reparameterization of the model originally introduced by Besag et al. (1991). The total precision of the county level components of the model is τ, and the precision matrix of the ICAR random effects is scaled so that ϕ can be interpreted as the proportion of the BYM2 variance that is spatial. Under this approach, the posterior distribution is obtained for the county level probabilities: $p_i^{\text{SM-DIR}} = \text{expit}(\eta_i)$, $i = 1, \ldots, 47$. This model was used in the context of estimating under-5 mortality rates in Mercer et al. (2015) over space and time and in Li et al. (2019). Its predictions typically have lower

variance than the direct estimates, but it still struggles with zero or one estimates and undefined/unstable variances.

Model-based approaches: For the model-based spatial approaches, we assume that $Y_c|p(\boldsymbol{x}_c) \sim \text{Bin}(n_c, p(\boldsymbol{x}_c))$, where n_c is the total number of children sampled in cluster c. The underlying risk at location $\boldsymbol{x}_c$ for cluster c is modeled as

$$\log\left(\frac{p(\boldsymbol{x}_c)}{1-p(\boldsymbol{x}_c)}\right) = \beta_0 + u(\boldsymbol{x}_c) + \beta^{\text{URB}} I(\boldsymbol{x}_c \in U) + \epsilon_c, \tag{1}$$

where β_0 is the intercept, $u(\boldsymbol{x}_c)$ is a spatial random effect, β^{URB} is the association with the cluster being urban (as compared to rural), and ϵ_c is an iid Gaussian cluster random effect with variance σ^2_ϵ. This term is sometimes described as the "nugget" and can represent many things including unmodeled sampling variability and small-scale variation.

The first model-based approach is termed BYM2, and uses the spatial random effect $u(\boldsymbol{x}) = \frac{1}{\sqrt{\tau}}(\sqrt{\phi}S_{i[\boldsymbol{x}]} + \sqrt{1-\phi}\delta_{i[\boldsymbol{x}]})$, where $i[\boldsymbol{x}]$ denotes the county to which $\boldsymbol{x}$ belongs, and the structure of the model follows the description for the smoothed direct model. This binomial model naturally deals with responses of 0 or n_c.

The second model-based approach is termed SPDE and uses a Gaussian process (GP) for the spatial random effect, $u(\cdot) \sim \text{GP}(\boldsymbol{0}, \boldsymbol{\theta})$, with $\boldsymbol{\theta} = [\sigma^2_s, \rho]$. The marginal variance is σ^2_s and the spatial range at which the correlation is approximately 0.1 is ρ. The GP used is the solution to a stochastic partial differential equation (SPDE) approximated by a Gaussian Markov Random Field (GMRF) defined on a fine triangular mesh (Lindgren et al., 2011).

For the BYM2 and SPDE models, we consider four variations of (1) depending on whether or not the association with urban/rural classification and the cluster (nugget) effects are included. Models with and without urban effects are labeled 'U' and 'u',

respectively. Similarly, models with and without cluster effects are labeled 'C' and 'c', respectively.

For the continuous (SPDE) model, if we knew the complete list of EA locations in the sampling frame, we could predict at the county level using the posterior distribution of a weighted sum of the predicted probabilities $p(\boldsymbol{x})$, calculated from (1), at the EA locations. The majority of EA locations are unobserved, however. In the absence of such information, we can aggregate by integrating the spatial probability surface $p(\boldsymbol{x})$ with respect to population density. Let p_i denote the county level estimates for county i, then

$$p_i = \int_{A_i} p(\boldsymbol{x}) \times q(\boldsymbol{x}) \, d\boldsymbol{x} \approx \sum_{j=1}^{m_i} p(\boldsymbol{x}_j) \times q(\boldsymbol{x}_j), \tag{2}$$

where A_i is the geographical extent of area i, $q(\boldsymbol{x})$ is the target population density at location $\boldsymbol{x}$ normalized so that $\int_{A_i} q(\boldsymbol{x}) \, d\boldsymbol{x} = 1$ for each i, and m_i is the number of grid cells with centroids in area i that is used to approximate the integral. Accounting for cluster effects when making aggregated predictions is more complicated in continuous spatial models since $p(\cdot)$ varies within each stratum, and the locations of unsampled EAs are not necessarily known. Rather than leaving out unobserved cluster effects when producing pixel level and aggregated predictions, we integrate out the cluster effects at each location in order to achieve the correct expectation at the pixel level before aggregation. More information on how we integrate out cluster effects and account for stratification for the SPDE model can be found in Appendix A.1.

In order to generate maps of urbanicity and population density as given in the right panel of Figure 1, we use 1km $\times$ 1km gridded population density surfaces from WorldPop (Stevens et al., 2015; Tatem, 2017). The 2010 and 2015 population density is interpolated, assuming a constant rate of population growth, to produce the 2014 population density map used throughout this paper. The 2009 Kenya Population

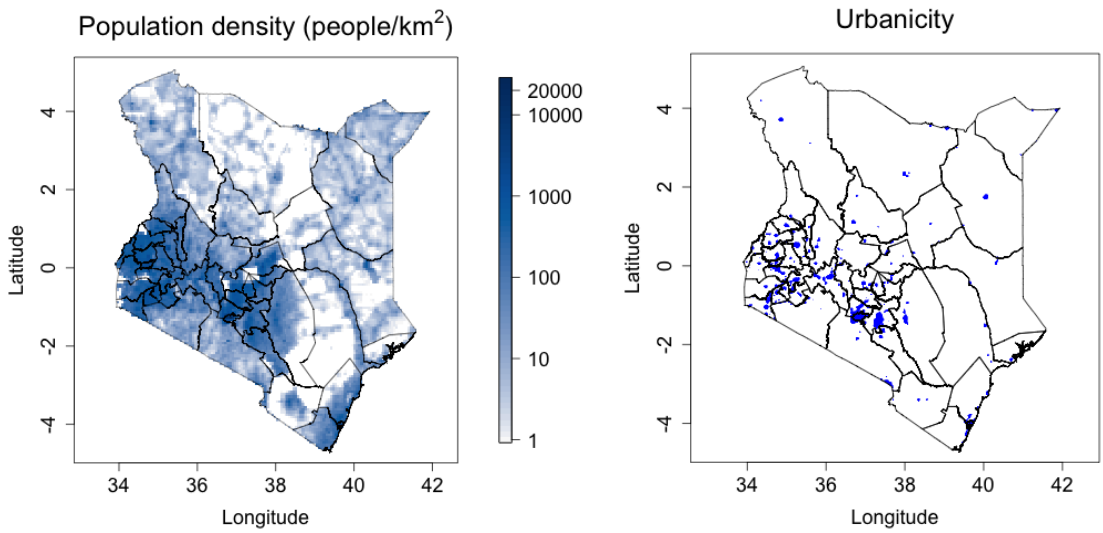

Figure 1: Left: WorldPop based population density estimates. Right: urban areas in Kenya used in this analysis are depicted in blue. Locations are determined to be urban versus rural based on thresholding population density.

and Housing Census provides information on the proportion of the population within each county that is urban and rural, and we generate urbanicity classifications by thresholding the population density maps within each county at the level required to achieve this proportion.

For the BYM2 model, a population density surface is not needed since the probabilities are modeled as constant within each area, and we can use,

$$\begin{aligned} p_i &= \text{expit}\left(\beta_0 + \frac{1}{\sqrt{\tau}}(\sqrt{\phi}S_i + \sqrt{1-\phi}\delta_i\right) Q_i \\ &+ \text{expit}\left(\beta_0 + \beta^{\text{URB}} + \frac{1}{\sqrt{\tau}}(\sqrt{\phi}S_i + \sqrt{1-\phi}\delta_i\right)(1-Q_i), \end{aligned}$$

where $Q_i = \int_{A_{iR}} q(\mathbf{x})\, d\mathbf{x}$, with A_{iR} representing the rural portion of A_i, is the proportion of the target population in county i that is rural. Further details of accounting for the cluster effects and performing the spatial aggregation for the BYM2 model are given in Appendix A.2.

WorldPop and IHME use a continuous GP model when analyzing household survey data. When aggregating to the county level, IHME simulates one cluster effect per pixel rather than integrating them out, while WorldPop does not include a nugget. The WorldPop and IHME models do not adjust for stratification. Hence, while WorldPop uses a 'uc' model, IHME uses a modified 'uC' model at the pixel level, but including slightly more variation. Both WorldPop and IHME, however, include pixel level covariates in their models that may fortuitously pick up part of any existing urban/rural association. Although the considered BYM2 models that include fixed effects for urbanicity account for stratification if the effect of urbanicity is the same in each county, this is an oversimplification. In addition, it is important to be clear that the model-based approaches we consider do not account for within stratum variations in the sampling weights, such as those due to nonresponse or PPS sampling in the case that nonresponse and EA size are associated with the latent probability surface p within strata.

3.2 Inference

Penalized complexity (PC) priors were introduced in Simpson et al. (2017), and penalize a model's "distance", on an appropriate scale, from a simple "base" model. For example, for iid random effects arising from a zero mean Gaussian distribution with variance σ^2, the base model corresponds to $\sigma = 0$. Following Fuglstad et al. (2019), we set a joint PC prior on the continuous spatial standard deviation and effective range parameters σ_s and ρ, respectively. We use the joint PC prior described by Riebler et al. (2016) on the BYM2 standard deviation $1/\sqrt{\tau}$ and the proportion of variation that is spatial, ϕ. We also set a marginal PC prior on the cluster effect standard deviation, σ_ϵ. The priors in the simulation study and in the application are set so that the

median of the prior for ρ is one fifth the diameter of the spatial domain, and so that $P(\sigma_s > 1) = P(1/\sqrt{\tau} > 1) = P(\sigma_\epsilon > 1) = 0.01$. This results in the continuous spatial effects, BYM2 effects, and cluster effects for the spatial smoothing models each having a roughly 95% prior chance of lying between 0.5 and 2 on an odds scale. The PC prior for the spatial proportion in the BYM2 model, ϕ, is given a 2/3 prior probability of being less than 1/2, implying that we slightly favor the iid county level effects when apportioning residual variation. We choose this prior on ϕ in order to promote less complex models with a smaller spatial contribution.

All design-based estimates were obtained using the `svyglm` function within the `survey` package (Lumley, 2004, 2018) in the `R` programming language. Each of the spatial models can be fitted using the integrated nested Laplace approximation (INLA) approach introduced in Rue et al. (2009), a method for fitting Bayesian models without the computational difficulties of Markov Chain Monte Carlo (MCMC) and implemented in the `INLA` package in `R`. The direct, smoothed direct and binomial BYM2 models are available in the SUMMER package (Martin et al., 2018). Code to reproduce the results can be found at https://github.com/paigejo/NMRmanuscript, and the 2014 Kenya DHS data can be requested from https://dhsprogram.com/.

4 Simulation Study

4.1 Comparison Measures

In this section, we describe an extensive simulation study in which we compare various models, in particular with respect to the inclusion of strata and cluster effects. We do this for multiple simulated populations and survey designs in order to test the models under a variety of circumstances. As in Section 3, the nominal response is a binary

indicator of whether or not death occurred within the first month of life.

We evaluate the model predictions at the county level using bias, mean squared error (MSE), the continuous rank probability score (CRPS), coverage of 80% intervals, and width of 80% intervals. All scoring rules are calculated on the probability scale. Note that CRPS is a strictly proper scoring rule (Gneiting and Raftery, 2007) that accounts for both predictive accuracy as well as accuracy of the uncertainty of the predictive distribution. Given the cumulative distribution function of the predictive distribution of a proportion in the finite population, F, and an empirical proportion response y/n, the CRPS is defined as:

$$\text{CRPS}(F, y) = \sum_{\tilde{y}=0}^{n} (F(\tilde{y}/n) - 1\{\tilde{y} \geq y\})^2.$$

Small values of CRPS are desirable.

The reported scoring rules are calculated using predictive distributions that have been calculated at the county level. The reported scores are the averages over counties and repeated surveys, and the full set of calculated scoring rules for all simulated populations and scenarios are given in the online supplementary material 3.3.

4.2 Simulation Setup

In order to generate a true, underlying population from which we can draw surveys, we first spatially partition Kenya into urban and rural zones by thresholding population density so that the proportion of population in each county that is urban matches the 2009 Kenya Population and Housing Census. We then simulate all 96,251 census EA locations such that the number in each of the 92 strata matches the true number, as given in the 2009 census. The EA locations in our simulated population are drawn proportional to population density within each strata. This information is all available

in the Kenya DHS final report (KDHS, 2014).

The number of households in each EA, as well as the number of mothers per household and children born per mother per year, are simulated based on the corresponding empirical distributions in the true population stratified by urban/rural. In order to estimate the empirical distribution for the number of households per EA, we take the maximum household ID sampled per cluster in the 2014 Kenya DHS as an estimate for the number of households in each EA.

We compared the models under 3 spatial parameter scenarios, 4 different populations for each scenario, and 2 different survey designs for each population. The four populations simulated per scenario have the following associated names: constant risk (Pop_{suc}), spatially-varying risk (Pop_{Suc}), spatially-varying risk with an urban association (Pop_{SUc}), and spatially-varying risk with an urban association and a cluster effect (Pop_{SUC}). Note that in the subscript labels for the populations, we again use U/u and C/c to indicate the presence of urban and cluster effects respectively, and we use S/s to indicate presence of a continuous spatial effect. In the case where we include spatial, urban, and cluster effects, we simulate NMRs at all 96,251 spatial locations using the SPDE model described above, with parameters depending on the population and scenario. In the first scenario, for the Pop_{SUC} population we simulate NMRs at all 96,251 spatial locations using the SPDE model described above with an effective correlation range of 150km, and with parameters $\beta_0 = -1.75$, $\sigma_s^2 = 0.15^2$, $\sigma_\epsilon^2 = 0.1^2$, and $\beta^{\text{URB}} = -1$. For "typical" rural/urban areas, with random effects of zero, the prevalences are 17%/6%. In the second scenario, the spatial range is decreased to 50km, and in the third scenario the spatial range is 150km, but the spatial variance is increased to 0.3^2. Only one Pop_{suc} population is simulated, since it includes no spatial effect. There are therefore 10 different simulated populations in total across all scenarios.

Within each simulated population we carry out "Unstratified" and "Stratified" sam-

pling, always taking 1,612 clusters to match the 2014 Kenya DHS. In the Unstratified design, we fix the total number of clusters in each county to be the same as in the Kenya DHS, and choose the average proportion of urban and rural clusters within each county to match the proportion of the urban and rural population in that county. The number of urban or rural clusters in any given stratum randomly varies by at most 1 from survey to survey if the proportion of urban population and urban clusters could not be matched exactly. This resulted in sampling 430 urban and 1182 rural clusters on average. In the Stratified design, we sample urban and rural clusters at different rates for each county so as to match the proportion of urban clusters in each county of the 2014 Kenya DHS, obtaining 995 and 617 urban and rural clusters respectively. Conditional on the total number of urban and rural clusters for each of the 92 strata, we use PPS sampling to determine which clusters are included in the surveys, sampling clusters with probability proportional to the number of households in each strata. Within each EA, 25 households are chosen at random to be included in the cluster sample. The simulated population and a single simulated survey based on the Stratified design are shown in Figure 2. We simulate 250 surveys for each design and each population, with naive and direct estimates being fit to all 250 of the surveys, and the other models being fit only to the first 100 due to computational constraints.

4.3 Simulation Results

In the following section, we discuss the first simulation scenario, where populations including the spatial effect were simulated with 150km spatial range and 0.15^2 spatial variance, unless otherwise stated. A more detailed analysis of the other scenarios is given in Section 3 of the supplementary material. The scoring rules summarizing the main results for the stratified design are plotted in Figure 3, and parameter summary

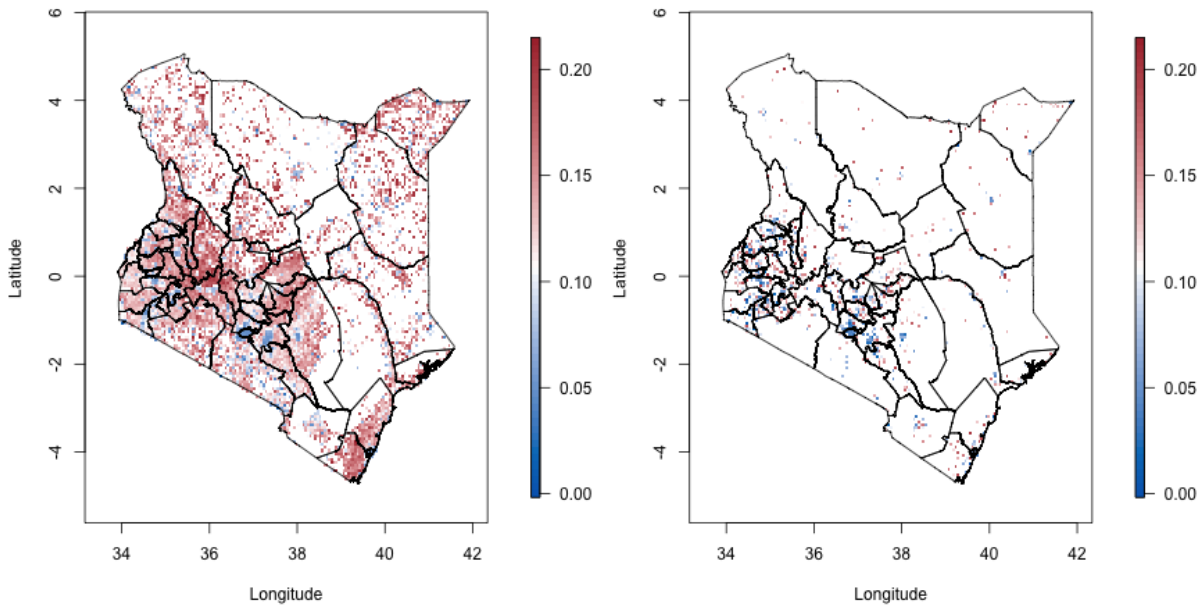

Figure 2: Simulated population of Kenya and associated NMRs at EAs (left), and at sampled clusters (right) for an example simulated dataset based on the "Stratified" design.

statistics are given in the supplementary material in Section 3.3. Scoring rules for additional model variations and designs are illustrated in the supplementary material in Section 3.2. When interpreting these scoring rules, it is important to keep in mind that SDG 3 calls for a reduction in NMRs to 12 deaths per 1,000 children, which corresponds to 120 $\times 10^{-4}$ children. When absolute bias is large relative to this number, it is an indicator of poor model performance. Since we are especially interested in the performance of the models in a feasible scenario in which spatial, urban and cluster effects must be accounted for, we will be discussing Pop_{SUC} under a Stratified design unless we state otherwise.

Of the direct, smoothed direct, BYM2_{UC}, and SPDE_{UC} models, the BYM2_{UC} model performed the best or very close to the best in terms of CRPS, MSE, coverage, and CI width across all populations and scenarios. Although the BYM2_{UC} model was slightly positively biased, the precision of its central predictions and the well-calibrated

predictive distribution and uncertainty intervals led to accurate coverages and good predictive performance.

Interestingly, although the SPDE_{UC} model matched the model used to simulate the data, it did not perform well in terms of MSE. Its MSE was 1.24×10^{-4} compared to 0.41×10^{-4} for the BYM2_{UC} model, 0.63×10^{-4} for the smoothed direct, and 0.72×10^{-4} for the estimates. Although SPDE_{UC} model estimates were somewhat positively biased, the high level of MSE was mainly due to lack of predictive precision. In spite of this, the SPDE_{UC} model had a CRPS of 4.7×10^{-3}, which was comparable to the value of 4.6×10^{-3} for the smoothed direct model, and was better than the direct estimates' value of 4.9×10^{-3}. Additionally, the coverage of the SPDE_{UC} model was 82%, second in accuracy only to the BYM2_{UC} model. Hence, although the predictions of the SPDE model were a little imprecise, the uncertainty of the posterior distribution was well-calibrated.

The direct, smoothed direct, and BYM2_{Uc} estimates had the smallest magnitude bias. An advantage of the smoothed direct model is that, regardless of the population and survey scheme considered, the model performed well from the standpoints of MSE, CRPS, bias, and coverage. Although the coverage for the Pop_{suc} (constant risk) population was over 90% for 80% nominal coverage intervals, that was also the case for the BYM2 models. Additionally, the constant risk setting is not realistic, and therefore should not be focused on too much. Overall, the smoothed direct model was robust in terms of its predictive accuracy and uncertainty.

Models that did not account for urbanicity either indirectly via sampling weights or directly as a covariate had relatively poor performance from MSE, bias, CRPS, and coverage standpoints. Even for populations without urban associations or under the unstratified design, there was little downside to including urban effects so long as the proportion of children in urban and rural areas was not poorly estimated (so that the

area averaging was poorly performed). Including urban effects led to MSE, bias, CRPS, credible interval width, and coverage that was on average either better or nearly equal to the corresponding models without urban effects throughout all simulated populations and designs. The benefit of including urbanicity as a covariate was increased under the Stratified design relative to the Unstratified design since urban and rural areas were not sampled proportionately in that case.

The scoring rules in Section 3.2 in the supplementary material shows the inclusion of a cluster effect led in general to better or equally good predictions for the BYM2 model in terms of MSE and CRPS. Although the MSE and bias of the BYM2_{uc} and BYM2_{uC} models were essentially the same, the inclusion of the cluster effect led to a dramatic improvement in coverage from 55% to 68%, indicating that cluster effects can lead to more accurate measures of uncertainty. Although the BYM2_{Uc} model arguably performs slightly better than the BYM2_{UC} model in terms of its MSE and CRPS, the coverage of the BYM2_{UC} model is better, and the uncertainty intervals are more conservative. The SPDE predictions were relatively more affected by the inclusion of the cluster effect, and its predictive performance was overall the most variable. Although there were some populations where the SPDE_{Uc} and SPDE_{UC} models had the best predictions in terms of MSE and CRPS, care must be taken when accounting for spatial and cluster level variation. To summarize, these simulations suggest that, amongst the BYM2 models, the BYM2_{UC} model is an effective and robust choice for the analysis of DHS household survey data, whereas more work must be done to ensure continuous spatial models share those same qualities.

Patterns in the model scores for the most part remained the same across the different simulation scenarios. The relative performances of the naive, direct, smoothed direct, and BYM2 models were mostly the same. The main differences were in the performance of the SPDE models. In spite of this, the SPDE models did not consistently perform

better than the BYM2 models in any of the scenarios, again indicating the relative robustness of the BYM2 and smoothed direct models, and the caution necessary when making aggregated predictions with such flexible spatial models using sparse household surveys.

5 Mapping Health and Demographic Indicators in Kenya

In this section, we use the 2014 Kenya DHS to estimate secondary education completion rates for women aged 20–29 in 2014, and to estimate NMRs for the five year period from 2010 to 2014. In the case of secondary education prevalence, we choose the 20–29 age group because most women that will complete their secondary education have already done so by that age, and also because the 2014 Kenya DHS indicates there are generational differences in secondary education levels. Although we find significant evidence of association with urbanicity in the case of secondary education completion, we did not find strong evidence of a marginal association between NMRs and urbanicity in Kenya. Since associations between urbanicity and the response lead to larger biases when stratification is not accounted for, we believe the secondary education completion dataset provides especially strong evidence for the importance of accounting for stratification in the design. Additional results for the women's secondary education and NMR applications are given in Sections 1 and 2 in the supplementary material respectively.

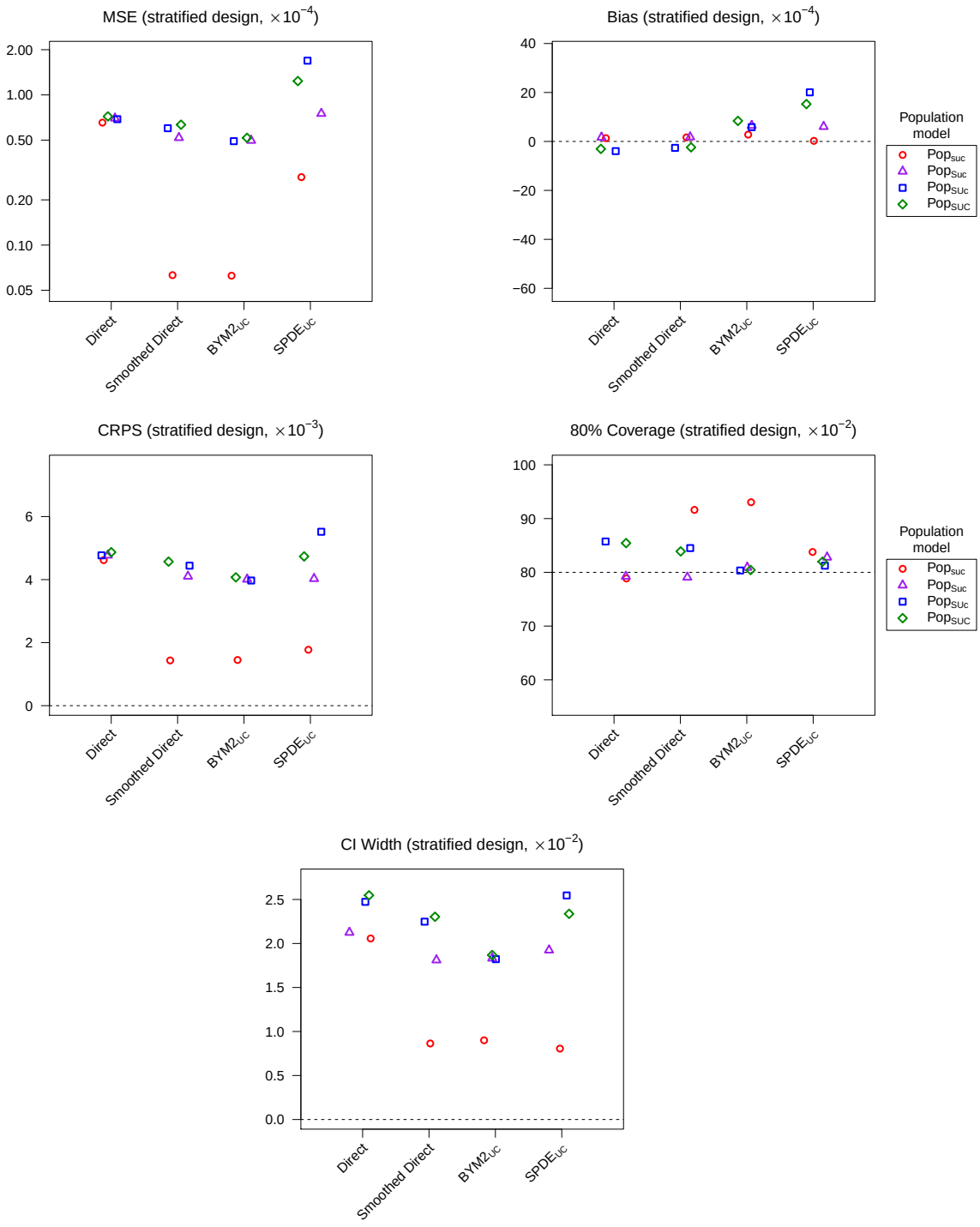

Figure 3: County level scoring rules plotted for each of the simulated populations and the main models considered for the Stratified design. The labels s/S, u/U, and c/C denote whether or not spatial, urban, and cluster effects are included in the models respectively. The "Population model" denotes the method by which the data were generated.

5.1 Prevalence of Women's Secondary Education

5.1.1 Prevalence Mapping

Central predictions as well as interval widths for the direct, naive, smoothed direct, and the full ('UC') spatial smoothing models are shown in Figure 4. The top row (point) estimates are quite similar, since there are a large number of clusters, but close examination shows there are differences. Prevalence tended to be higher in the central, southern, and western counties, and tended to be lower and with greater uncertainty in the more rural counties to the north and east. Section 1 in the supplementary material gives full numerical results and here we summarize. The odds (with associated 80% CIs) of young women in urban clusters completing their secondary education are larger, relative to rural clusters, by 210% (185%, 236%) or 170% (148%, 193%) as respectively calculated from the BYM2_{UC} and SPDE_{UC} model parameter estimates given in Table 1.

Table 1 in the supplementary material shows that the smoothed direct, SPDE_{UC}, and BYM2_{UC} models all estimate that the secondary education levels for young women in Kenya were highest in Nairobi, with point estimates (80% CIs) of 0.54 (0.49, 0.58), 0.55 (0.51, 0.58), and 0.53 (0.50, 0.57) respectively. On the other hand, Mandera was estimated to have the lowest secondary education levels for all models except for the SPDE_{UC} model (for which Turkana was estimated to have the lowest secondary education levels) with point estimates (80% CIs) of 0.088 (0.061, 0.13), 0.081 (0.058, 0.11), and 0.085 (0.060, 0.12) respectively. While Nairobi is designated as completely urban, approximately 18% of the population of Mandera is urban, which is very close to the median for counties in Kenya. This suggests there are other factors in Mandera that are reducing the secondary education prevalence for the women living there.

The credible interval widths were largest for the direct model and smallest for the SPDE_{UC} model. Of the displayed spatial smoothing models, the smoothed direct model

had the largest predictive variances. Both the smoothed direct and the BYM2 models had relatively high uncertainties for counties with fewer neighbors, whereas the SPDE model variances tended to be high near the edges and where the clusters were spatially distant from each other.

Figure 5 shows the continuous, 5km×5km pixel level predictions and credible interval widths for the SPDE_{uC} and SPDE_{UC} models. The urban effect is especially visible in the predictions of the SPDE_{uC}, since it oversmooths the effect of the urban areas into nearby rural regions. Interestingly, secondary education prevalences appear to be high not only in urban areas, but also in rural areas bordering urban centers. Figure 5 clearly shows that care that must be taken with stratification. In order to maintain confidentiality, the geographical locations of the DHS clusters are displaced (jittered): urban clusters by up to 2km, and rural clusters by up to 5km, with 1% of rural clusters jittered up to 10km. In general, we are nervous about presenting and over-interpreting pixel level maps due to jittering, cluster level overdispersion, data sparsity, and confounding pixel level covariates such as urbanicity. Although continuous spatial models in theory allow for high resolution predictions, it is very important when modeling demographic indicators from sparse household survey data to account for these complicating factors.

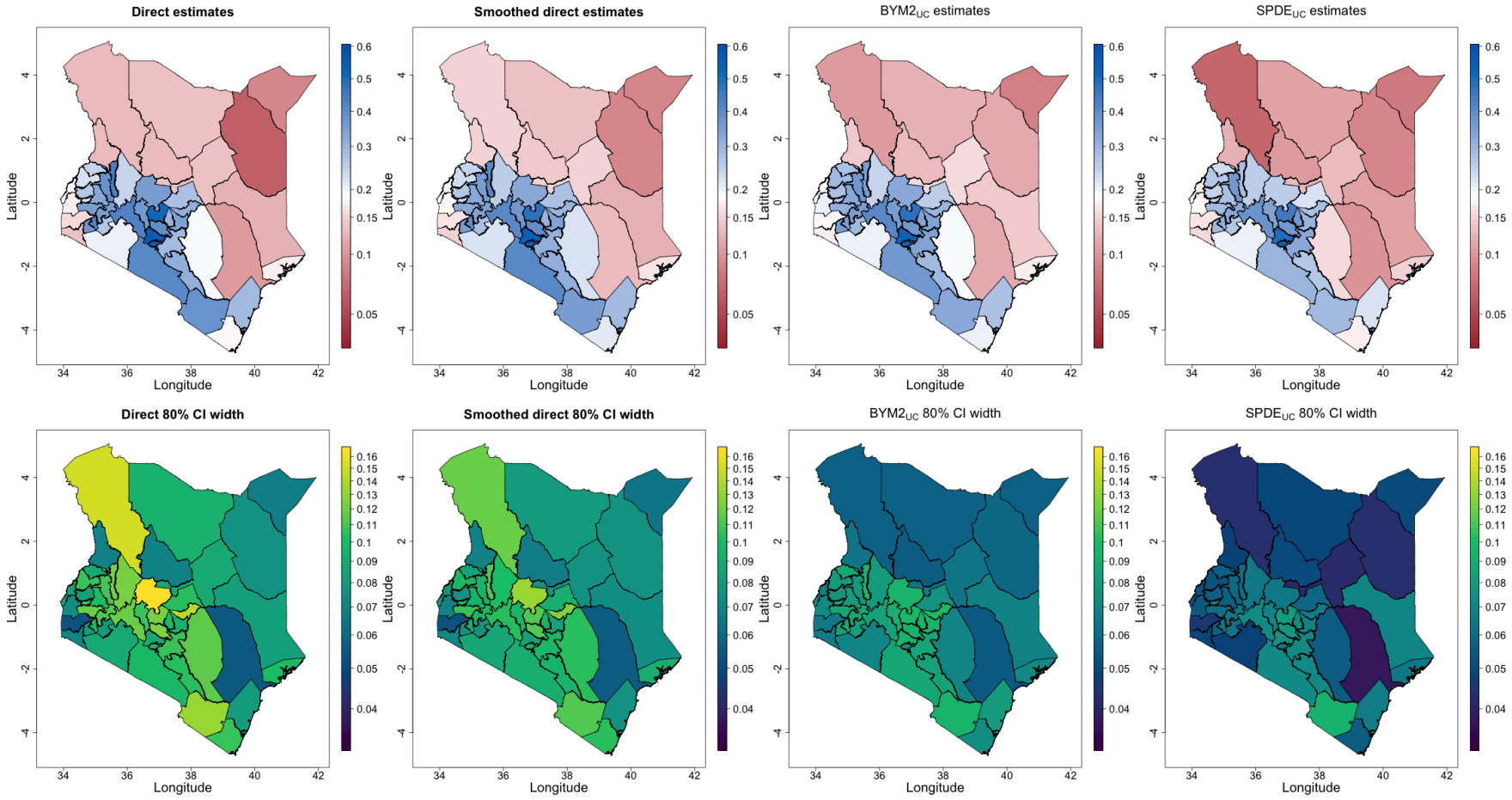

Figure 4: Kenya county level 2014 secondary education predictive estimates (top) and 80% uncertainty interval width (bottom) for women aged 20–29. The BYM2 and SPDE model predictions include both urban and cluster effects.

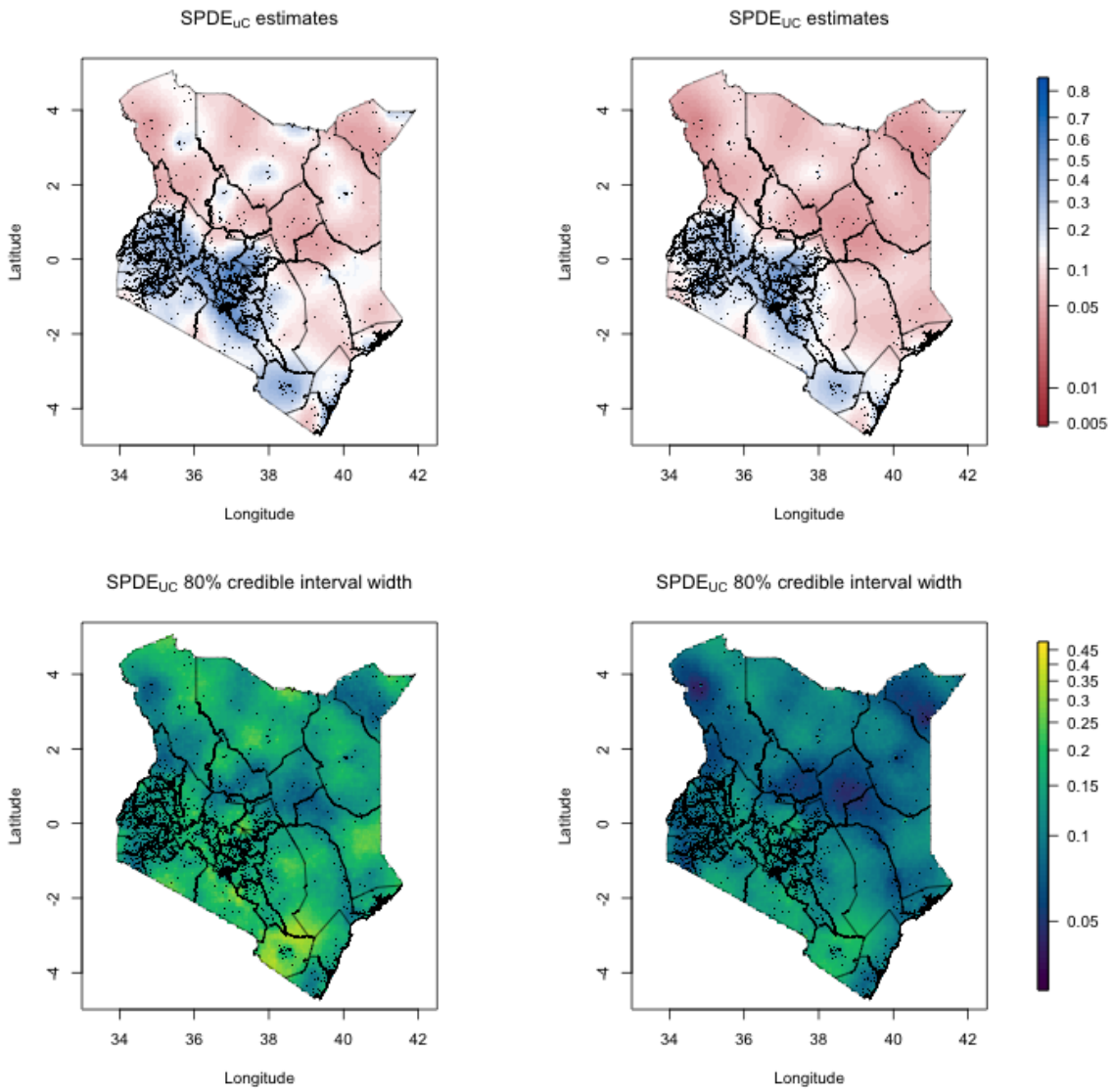

Figure 5: Kenya 5km × 5km pixel level 2014 secondary education predictive mean (top) and 80% uncertainty interval width (bottom) for women aged 20–29. Results are shown for the SPDE_{uC} (left) and SPDE_{UC} (right) models.

	Est	SD	Q10	Q50	Q90
Smoothed Direct					
Intercept	-1.04	0.05	-1.10	-1.04	-0.99
BYM2 Phi	0.80	0.16	0.57	0.84	0.97
BYM2 Tot. Var	0.32	0.08	0.22	0.31	0.42
BYM2 Spatial Var	0.26	0.09	0.15	0.25	0.37
BYM2 iid Var	0.06	0.05	0.01	0.05	0.13
BYM2 Tot. SD	0.56	0.07	0.47	0.55	0.65
BYM2 Spatial SD	0.50	0.09	0.38	0.50	0.61
BYM2 iid SD	0.22	0.10	0.10	0.22	0.36
$\boldsymbol{5}BYM2_{UC}$					
Intercept	-1.64	0.06	-1.72	-1.65	-1.57
Urban	1.13	0.07	1.05	1.13	1.21
Cluster Var	0.51	0.05	0.44	0.51	0.58
BYM2 Phi	0.84	0.14	0.64	0.88	0.98
BYM2 Tot. Var	0.33	0.08	0.24	0.32	0.44
BYM2 Spatial Var	0.28	0.09	0.17	0.27	0.40
BYM2 iid Var	0.05	0.05	0.01	0.04	0.11
Cluster SD	0.72	0.04	0.67	0.71	0.76
BYM2 Tot. SD	0.57	0.07	0.49	0.57	0.66
BYM2 Spatial SD	0.52	0.09	0.42	0.52	0.63
BYM2 iid SD	0.21	0.10	0.09	0.20	0.34
$\boldsymbol{SPDE_{UC}}$					
Intercept	-2.53	0.21	-2.80	-2.53	-2.26
Urban	0.99	0.07	0.91	1.00	1.08
Range	212	44	161	206	271
Spatial Var	0.84	0.24	0.57	0.81	1.16
Spatial SD	0.91	0.13	0.76	0.90	1.08
Nugget Var	0.43	0.05	0.36	0.43	0.50
Nugget SD	0.65	0.04	0.60	0.65	0.70

Table 1: Parameter and hyperparameter estimate summary statistics from the BYM2 and SPDE models including both urban and cluster effects, fit to the 2014 secondary education prevalence data for young women from the 2014 Kenya DHS.

5.1.2 Validation

We calculate a number of scoring rules at the cluster level to evaluate the spatial smoothing models. We compute the scoring rules by leaving out data from one county at a time and averaging the scoring rules over all 47 such experiments. We carry out the validation at the county level, since this is generally the target of inference. In addition to calculating MSE (broken down into variance and bias and in urban as well as rural areas) we also compute CRPS, the deviance information criterion (DIC), and the conditional predictive ordinate (CPO). The naive, direct, and smoothed direct models are fit at the county level, so we did not include their validation results, since they are not comparable with the cluster level data models.

The SPDE_{Uc} and SPDE_{UC} models had the two best average MSEs, and the SPDE_{Uc} model has the best CPO and CRPS. The SPDE_{UC} had better MSE than the Uc model, but it had worse CPO and CRPS. In terms of MSE and CRPS, the BYM2_{Uc} model also performed well, and had the smallest magnitude bias. The good performance of the SPDE models may be due to their ability to model continuous changes in secondary education near the borders of each county, whereas the BYM2 models are unable to distinguish between clusters at the border of a county versus clusters in the center. Section 1 in the online supplementary material shows that the spatial standard deviation (SD) of the SPDE_{UC} model is estimated to be 0.91, whereas the cluster effect is estimated to have a SD of 0.65. This is a higher proportion of variability going into the spatial term than in the BYM2_{UC} model, which estimates the total variance of county level random effects to be 0.57 and the cluster variance to be 0.71. This suggests the ability of the SPDE model to predict continuously through space gives it an advantage when making predictions at the cluster level.

	BYM2				***SPDE***			
	uc	uC	Uc	UC	uc	uC	Uc	UC
MSE ($\times 10^{-2}$)								
Average	5.9	*6.0*	5.1	5.2	5.6	5.4	5.0	**4.9**
Urban	*7.2*	*7.2*	6.2	6.2	7.1	6.8	6.2	**6.0**
Rural	5.0	*5.2*	4.5	4.5	4.6	4.5	**4.2**	**4.2**
Var ($\times 10^{-3}$)								
Average	59	*59*	51	52	56	53.8	50	**49**
Urban	61	62	62	62	*63*	**60**	62	60
Rural	45	45	45	45	*45*	42	42	**42**
Bias ($\times 10^{-3}$)								
Average	6.0	10.5	**1.0**	5.6	*-15.0*	-6.2	-3.6	-3.2
Urban	*-105*	-103	11.1	8.3	-93	-91	**1.0**	3.1
Rural	77	*83*	-5.4	**4.0**	35	48	-6.5	-7.2
CPO								
Average	0.22	*0.21*	0.25	0.24	0.26	0.25	**0.27**	0.26
CRPS								
Average	0.17	*0.21*	0.16	0.18	0.17	0.18	**0.15**	0.17

Table 2: Validation results calculated at the cluster level when leaving out one county at a time for the 2014 Kenya DHS secondary education data for women aged 20–29. The worst entries in each row are in *italics*, while the best entries in each row are in **bold**. In the table, the figures are rounded, but minimum and maximum were evaluated with more significant figures.

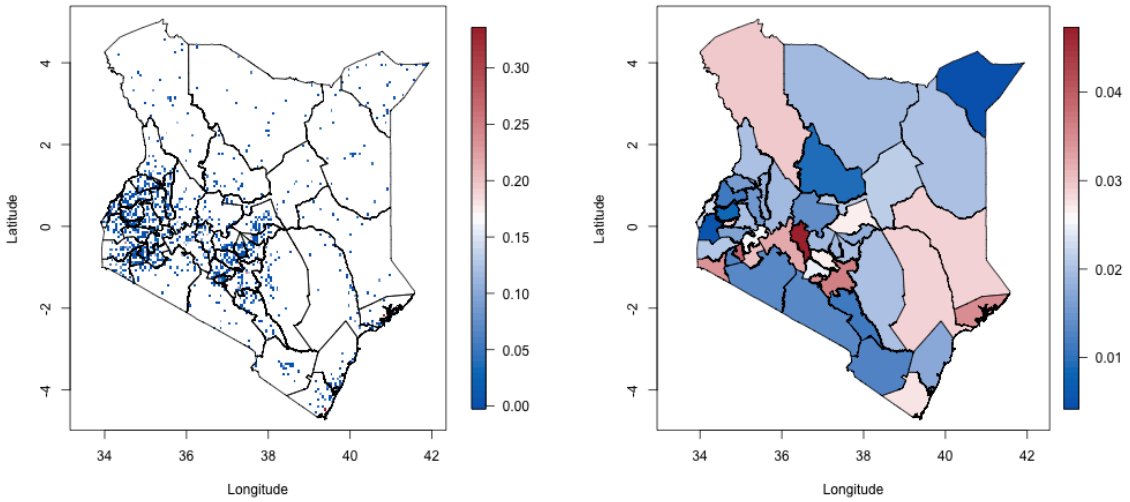

Figure 6: Empirical average of neonatal mortality rates in Kenya from 2010-2014 based on data from the 2014 Kenya DHS. Values are shown at both the cluster (left) and county levels (right).

5.2 Prevalence of Neonatal Mortality

5.2.1 Prevalence Mapping

We now estimate neonatal mortality rates (NMRs) for children in Kenya from 2010–2014 based on data from the 2014 Kenya DHS plotted in Figure 6. The figure shows that the vast majority of clusters have empirical NMRs very close to zero, though there are some clusters that have much higher NMRs with some even above 30%. Central NMR estimates, and 80% uncertainty interval widths in Figure 7, and individual county level predictions are given in Section 2 of the supplementary material along with upper and lower predictive quantiles at the county level. The largest NMRs were estimated to be in several counties just northwest of Nairobi as well as in central eastern Kenya, and the lowest NMRs were estimated to be in the counties near the central western and southwestern borders. Although we expected to find a significant urban effect, we found little evidence suggesting any difference in NMRs between rural and urban areas. Additionally, there is much less spatial variation relative to cluster level variance. For instance, the BYM2_{UC} and SPDE_{UC} models estimate the spatial effect variance to respectively be 0.060 and 0.069, whereas the estimated cluster effect variances were respectively 0.18 and 0.23. We found little difference in the quality of predictions of the models when we validated them using the same method as above, indicating that the low signal-to-noise ratio in this application makes improvement in spatial estimation difficult.

The fitted smoothed direct, BYM2_{UC}, and SPDE_{UC} models had smaller spatial effect variances relative to the predictive uncertainties when compared to the analysis of secondary education prevalence. For the SPDE_{UC} model, for instance, this is evidenced by the fact that the median 80% credible interval width for the NMR data is 0.0074, whereas the point estimates have a range of 0.0098. The equivalent values in the sec-

ondary education application, which are respectively 0.062 and 0.476, indicate much greater variability across space. The estimated variances of the county level random effects are also relatively small, and are estimated to be 0.059 and 0.060 for the smoothed direct and BYM2_{UC} models, respectively. The variance of the spatial component of the SPDE_{UC} model was estimated to be 0.069. The cluster effect variance, however, was estimated to be comparatively larger, with BYM2 and SPDE_{UC} estimates of 0.183 and 0.225 respectively, further indicating the large amount of noise relative to any spatial signal in the data. In spite of the lack of spatial signal, the spatial smoothing models helped to reduce the predictive uncertainties relative to the naive and direct models as evidenced from the plotted credible interval widths of the predictions in Figure 7.

Considering the difficulty of including cluster level variation when aggregating SPDE_{UC} predictions to the county level, combined with the fact that a substantial majority of the variation in the data occurs at the cluster level as opposed to spatial level variation, we believe the smoothed direct and BYM2 models might be better suited for this particular application. Not including variation due to cluster effects in the spatial aggregation, aside from integrating out cluster level variation, is probably what leads the SPDE_{UC} model predicted NMRs to have such narrow credible intervals relative to the other models.

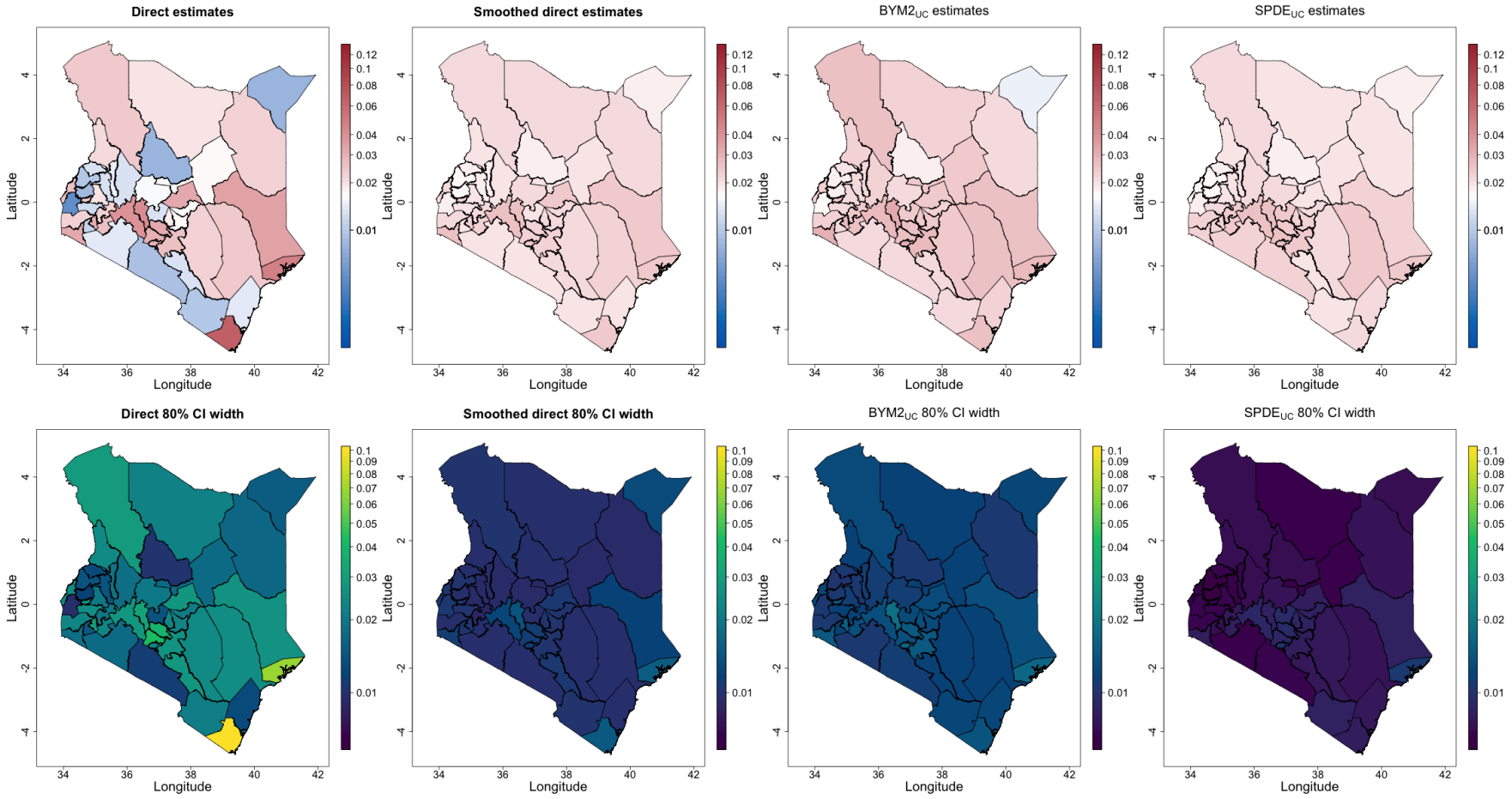

Figure 7: Kenya county level neonatal mortality rate predictive mean (top row) and 80% uncertainty interval widths (second row) from 2010-2014. The BYM2 and SPDE model predictions include both urban and cluster effects.

	Est	SD	Q10	Q50	Q90
Smoothed Direct					
Intercept	-3.85	0.07	-3.94	-3.85	-3.76
BYM2 Phi	0.30	0.25	0.04	0.23	0.70
BYM2 Tot. Var	0.06	0.04	0.02	0.05	0.11
BYM2 Spatial Var	0.02	0.02	0.001	0.01	0.05
BYM2 iid Var	0.04	0.03	0.009	0.03	0.08
BYM2 Tot. SD	0.23	0.08	0.13	0.22	0.33
BYM2 Spatial SD	0.11	0.07	0.04	0.10	0.22
BYM2 iid SD	0.19	0.08	0.09	0.18	0.29
$\boldsymbol{BYM2_{UC}}$					
Intercept	-3.99	0.09	-4.11	-3.99	-3.87
Urban	0.08	0.11	-0.07	0.08	0.22
Cluster Var	0.18	0.10	0.07	0.17	0.31
BYM2 Phi	0.42	0.31	0.044	0.36	0.88
BYM2 Tot. Var	0.06	0.04	0.02	0.05	0.11
BYM2 Spatial Var	0.02	0.02	0.002	0.02	0.05
BYM2 iid Var	0.04	0.03	0.004	0.03	0.08
Cluster SD	0.41	0.11	0.27	0.41	0.56
BYM2 Tot. SD	0.24	0.07	0.15	0.23	0.33
BYM2 Spatial SD	0.14	0.07	0.05	0.13	0.23
BYM2 iid SD	0.17	0.08	0.07	0.17	0.29
$\boldsymbol{SPDE_{UC}}$					
Intercept	-4.00	0.09	-4.12	-4.00	-3.89
Urban	0.08	0.11	-0.06	0.08	0.22
Range	241	195	79	178	451
Spatial Var	0.07	0.06	0.02	0.05	0.14
Spatial SD	0.24	0.10	0.13	0.23	0.38
Nugget Var	0.23	0.10	0.11	0.21	0.37
Nugget SD	0.46	0.11	0.33	0.46	0.61

Table 3: Parameter and hyperparameter estimate summary statistics from the BYM2 and SPDE models including both urban and cluster effects, fit to the 2010-2014 NMR data from the 2014 Kenya DHS.

5.2.2 Validation

We validate the spatial smoothing models that produce predictions at the cluster level by once again leaving out data from each county, one county at a time, and making predictions at the cluster level. Equivalent scoring rules to those used in Section 5.1.2 are calculated for the NMR dataset in Table 4, which shows that the SPDE_{UC} model performs just as well as any of the other spatial smoothing models when making cluster level predictions. In fact, all of the models perform very similarly, again suggesting that variation in the NMR data is primarily at either the cluster level or at spatial scales too fine to easily identify.

6 Discussion and Conclusions

Direct estimators remain the gold standard, provided there are sufficient data for an associated variance that is of acceptable size. The smoothed direct estimator can reduce the variance using the totality of data, albeit with an introduction of some bias due to the smoothing. This bias disappears, however, as sample size increases. When the direct estimates are unreliable, one is led to modeling at the cluster-level, and it is important to use a model that is consistent with the design. In this paper, we introduced a binomial sampling model with discrete spatial random effects, and it performed well in the simulations and real applications. The BYM2 random effects capture different levels in the regional strata. Hence the random effects should be nested within the regional strata to allow for different levels for the different sampling strata. Our method is designed for situations in which the sampling probabilities depend on stratification. Extending the methods to account for more complex sampling plans, non-response, and calibration will be the subject of future research.

We have also been experimenting with a beta-binomial model that allows for

	BYM2				***SPDE***			
	uc	uC	Uc	UC	uc	uC	Uc	UC
MSE ($\times 10^{-4}$)								
Avg	24.5	24.6	24.6	24.6	24.6	24.6	24.6	24.6
Urban	29.6	29.6	29.6	29.7	29.6	29.6	29.7	29.7
Rural	21.3	21.4	21.3	21.4	21.4	21.4	21.4	21.4
Var ($\times 10^{-4}$)								
Avg	24.5	24.5	24.6	24.6	24.6	24.6	24.6	24.6
Urban	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.7
Rural	21.3	21.3	21.3	21.3	21.4	21.4	21.4	21.4
Bias ($\times 10^{-4}$)								
Avg	8.2	26.0	8.8	27.1	4.1	3.3	5.1	7.5
Urban	-0.1	17.7	11.0	29.0	-3.0	-4.0	8.0	10.9
Rural	13.4	31.3	7.4	25.9	8.6	7.9	3.2	5.3
CPO								
Avg	0.657	0.649	0.657	0.649	0.659	0.662	0.659	0.660
CRPS								
Avg	0.024	0.025	0.024	0.025	0.024	0.025	0.024	0.025

Table 4: Validation results calculated at the cluster level when leaving out one county at a time for the 2014 Kenya DHS NMR data from 2010-2014.

overdispersion (within-cluster variation). Such approaches with discrete spatial models do not deal well with combining data at different geographical resolutions, and this is where the continuous spatial model is appealing. Unfortunately, aggregation with continuous models is the most difficult, since a population surface is required, and the model and population surface must, in general, be stratified by urban/rural.

In the simulations and application, we found that accounting for the design nearly always improved predictions in the case that strata were associated with the response, and did not reduce predictive performance otherwise. This was true whether the design was accounted for using sampling weights or by including stratification indicators as covariates. Although not included in our results, we have found that when the proportions urban required for aggregating strata predictions to the county level were biased, including stratification effects in the BYM2 model sometimes made the predictions

worse. This implies that not only must design stratification be accounted for, but in the case where it is included as a covariate, it is important to make an effort to obtain high quality estimates of the proportions of the studied population in each strata. In practice, this may be difficult.

Although we expected the SPDE models without urban effects to have better predictions than BYM2 models without urban effects since urbanicity is a spatial variable, we instead found that the spatial component of the SPDE models without urban effects had difficulty handling the sharp changes of urbanicity over short distances. As mentioned previously, WorldPop and IHME do not adjust for the urban/rural stratification, and WorldPop does not account for cluster level overdispersion, but they do include extensive covariates such as population density, which will, to some extent at least, adjust for urban/rural.

A remaining open avenue of research is to determine how best to include cluster effects in area-level aggregated predictions from spatial models. Since the SPDE model predictions are aggregated to the county level by numerically integrating predictions on a spatial grid, whereas cluster effects are modeled discretely at cluster and EA point locations, it is unclear how to accurately proceed when the EA locations are unknown. Simply leaving out cluster effects when aggregating predictions spatially may lead to undercoverage and also bias. Integrating out the cluster effect eliminates the bias, but not the undercoverage, whereas using Monte Carlo methods to sample possible EA locations and improve coverages is computationally expensive. Note that the the true data-generating mechanism for the cluster effects should ideally be a consideration when deciding how to handle cluster effects, and will affect the bias and coverage. Although we assume cluster effects are true latent differences in mortality rates, and fixed for each EA, it is possible they are due to measurement error, for instance, in which case they should be left out of predictions.

The simulation study and prevalence application indicated that the smoothed direct model had the least dependence on the specific implementation, performing well in nearly all circumstances, whereas the SPDE and BYM2 models that included stratification and cluster effects performed particularly well when there was a stratification effect in the population. This was especially the case if the proportion of the population of interest (i.e., children, or women aged 20–29) that is urban in each county is accurately known. The BYM2 model with urban and cluster effects performed the best or nearly the best in all scenarios and with all populations in the simulation study, in terms of MSE, CRPS, and credible interval width. The SPDE model including urban and cluster effects performed better in the cluster level validation, but care must be taken in selecting a prior due to its flexibility, and in generating spatially aggregated predictions when the estimated cluster effect accounts for a large proportion of the total variation.

In the simulation study, the DHS we emulated was powered to the Admin2 level, which coincided with the level of inference. More commonly, DHS data are powered to the Admin1 level and it is an open question as to what the recommendations are in this case if inference is still required at Admin2. In other work (Li et al., 2019) we could only perform Admin1 level inference for countries in Africa using the majority of the DHS surveys, because there were insufficient samples to apply the direct and smoothed direct methods at the Admin2 level.

References

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

Chen, C., J. Wakefield, and T. Lumley (2014). The use of sample weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-Temporal Epidemiology 11*, 33–43.

Congdon, P. and P. Lloyd (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science 8*, 235–252.

Desmond-Hellmann, S., K. L. Mueller, P. J. Hines, J. Travis, P. Szuromi, Y. Nusinovich, M. S. Lavine, S. Vignieri, B. Grocholski, W. Wong, et al. (2016). Progress lies in precision. *Science 353*.

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth 59*, S9–S15.

DHS Program (2019). The DHS program–AIDS indicator surveys (AIS). https://dhsprogram.com/What-We-Do/Survey-Types/AIS.cfm.

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

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

Dowell, S. F., D. Blazes, and S. Desmond-Hellmann (2016). Four steps to precision public health. *Nature News 540*, 189–191.

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

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

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

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

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

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

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

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

ICF International (2012). *Demographic and Health Survey Sampling and Household Listing Manual*. Calverton, Maryland, USA: ICF International.

Jerven, M. (2013). *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press.

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 (2009b). *The 2009 Kenya Population and Housing Census Volume IC: Population Distribution by Age, Sex, and Administrative Units*. Nairobi: Kenya National Bureau of Statistics.

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 (2015a). *Kenya Demographic and Health Survey 2014*. Rockville, Maryland, USA.

Khoury, M. J., M. F. Iademarco, and W. T. Riley (2016). Precision public health for the era of precision medicine. *American Journal of Preventive Medicine 50*.

Lawn, J. E., H. Blencowe, S. Oza, D. You, A. C. Lee, P. Waiswa, M. Lalli, Z. Bhutta, A. J. Barros, P. Christian, et al. (2014). Every newborn: progress, priorities, and potential beyond survival. *The Lancet 384*, 189–205.

Li, Z. R., Y. Hsiao, J. Godwin, B. D. Martin, J. Wakefield, and S. J. Clark (2019). Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLoS One 14*. Published January 22, 2019.

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

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

Lumley, T. (2018). survey: analysis of complex survey samples. R package version 3.35.

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

Martin, B. D., Z. R. Li, Y. Hsiao, J. Godwin, J. Wakefield, and S. J. Clark (2018). *SUMMER: Spatio-Temporal Under-Five Mortality Methods for Estimation*. R package version 0.2.1.

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

Osgood-Zimmerman, A., A. I. Millear, R. W. Stubbs, C. Shields, B. V. Pickering, L. Earl, N. Graetz, D. K. Kinyoki, S. E. Ray, S. Bhatt, A. Browne, R. Burstein, E. Cameron, D. Casey, A. Deshpande, N. Fullman, P. Gething, H. Gibson, N. Henry,

M. Herrero, L. Krause, I. Letourneau, A. Levine, P. Liu, J. Longbottom, B. Mayala, J. Mosser, A. Noor, D. Pigott, E. Piwoz, P. Rao, R. Rawat, R. Reiner, D. Smith, D. Weiss, K. Wiens, A. Mokdad, L. S.S., C. Murray, N. Kassebaum, and S. Hay (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature 555*.

Porter, A. T., S. H. Holan, C. K. Wikle, and N. Cressie (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics 10*, 27–42.

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

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

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

Sandefur, J. and A. Glassman (2015). The political economy of bad data: Evidence from African survey and administrative statistics. *The Journal of Development Studies 51*, 116–132.

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

Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015). Disaggregating

census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One 10*, e0107042.

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific data 4*.

UNICEF - Statistics and Monitoring (2012). Multiple Indicator Cluster Surveys (MICS). http://www.unicef.org/statistics/index_24302.html.

United Nations (2019). *Sustainable Development Goals*. http://sustainabledevelopment.un.org/owg.html.

USAID (2019). *Demographic and Health Surveys*. http://www.dhsprogram.com: United States Agency for International Development.

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

Vandendijck, Y., C. Faes, R. S. Kirby, A. Lawson, and N. Hens (2016). Model-based inference for small area estimation with sampling weights. *Spatial Statistics 18*, 455–473.

Wagner, Z., S. Heft-Neal, Z. A. Bhutta, R. E. Black, M. Burke, and E. Bendavid (2018). Armed conflict and child mortality in Africa: a geospatial analysis. *The Lancet 392*, 857–865.

Wardrop, N., W. Jochem, T. Bird, H. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. Tatem (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.

Watjou, K., C. Faes, A. Lawson, R. Kirby, M. Aregay, R. Carroll, and Y. Vandendijck (2017). Spatial small area smoothing models for handling survey data with nonresponse. *Statistics in Medicine 36*, 3708–3745.

World Bank (2019). Living standards measurement study (lsms) — surveyunit. http://surveys.worldbank.org/lsms.

You, Y. and Q. M. Zhou (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology 37*, 25–37.

Appendix A: Spatial Aggregation

Appendix A.1: BYM2 Model

Although spatial aggregation for the BYM2 models without cluster effects is relatively straightforward, it is less obvious how to produce county level estimates for the BYM2 models that include cluster effects. There may be census estimates of the proportion of population in each county that is urban versus rural, and the number of EAs that are urban and rural within each county may also be known; we will use this information when calculating county-level estimates.

It is possible to account for excess-variation due to cluster effects when producing estimates for each modeled stratum (county level estimates for BYM2_{uC} model and county $\times$ urban/rural for the BYM2_{UC} model) by averaging random variation over the known number of EAs for a given county:

$$\hat{p}_S^j = \frac{1}{C_S} \sum_{c \,:\, s[c]=S} \hat{p}_c^j, \tag{3}$$

where $\hat{p}_S^j$ is the jth drawn predicted probability from the posterior for a given stratum S, where stratum could be, for instance, county crossed with urban or rural designation, $s[c]$ is the stratum of EA c, C_S is the number of EAs in stratum S, and $\hat{p}_c^j$ is the jth drawn predicted probability from the posterior for EA c. Since the weights on each of the EAs are equal, this assumes that each EA within the stratum is approximately equal-sized, but if the number of EAs within any given stratum is large and their size is iid and independent of the response, then this method will also be a good approximation to the true county level posterior. Since the number of people in each EA is not known in practice, it is unclear how to better aggregate cluster level results to the modeled stratum level. For the BYM2_{UC} model, we can denote the two strata within each county as U and R for urban and rural with respective estimates $\hat{p}_U^j$ and $\hat{p}_R^j$. We can use this method to generate draws from the county level posterior for the BYM2_{uC} model and from the county $\times$ urban/rural level for the BYM2_{UC} model.

For the BYM2_{UC} model, we then use,

$$\hat{p}_i^j = \frac{T_{iU}}{T_{iU} + T_{iR}} \hat{p}_{iU}^j + \frac{T_{iR}}{T_{iU} + T_{iR}} \hat{p}_{iR}^j,$$

to sample from the posterior distribution of county i, where T_{iU} and T_{iR} are the total target population (i.e., children within the first month of life or women aged 20–29) in county i that is urban and rural respectively. Note that this requires knowledge of T_{iU} and T_{iR}, which might only be known approximately in practice. For Kenya, although we do not know the target population totals, we know the number of EAs in urban and rural strata for each county, say C_{iU} and C_{iR} respectively. We also use census data to estimate the distribution of the total target population per urban or rural EA, with expected values of E_U and E_R respectively. We then use $\widehat{T}_{iU} = E_U C_{iU}$ and $\widehat{T}_{iR} = E_R C_{iR}$ as plug-in estimates for the target population totals in each stratum.

Appendix A.2: SPDE Model

Recall,

$$p_i = \int_{A_i} p(\boldsymbol{x}) \times q(\boldsymbol{x})\ d\boldsymbol{x} \approx \sum_{j=1}^{m_i} p(\boldsymbol{x}_j) \times q(\boldsymbol{x}_j). \tag{4}$$

We would like to aggregate predictions over the 'target' population in county i. The target population might be children within the first month of birth or women aged 20–29.

Let $q(\cdot)$ be the population density throughout the county as a function of space, and normalized so that $\int_{A_i} q(\mathbf{x})\ d\mathbf{x} = 1$ for $i = 1, \ldots, 47$. Ideally, we would know the target population density rather than the overall population density, although that is not necessarily the case. If not, we can adjust the overall population density, q, as needed.

In particular, we would like the integral of the target population density to be proportional to T_{iU} and T_{iR} in the urban and rural parts of area i so that it is more representative of our target population. If A_i is the spatial domain of area i, we can partition it into urban and rural parts: $A_i = A_{iU} \cup A_{iR}$. We can then adjust the population density surface, creating a new surface more representative of the target population:

$$\tilde{q}(\boldsymbol{x}) = \begin{cases} \left[\int_{A_{iU}} q(\boldsymbol{x})\ d\boldsymbol{x}\right]^{-1} \frac{T_{iU}}{T_{iU}+T_{iR}} \times q(\boldsymbol{x}) & \boldsymbol{x} \in A_{iU}, \\ \left[\int_{A_{iR}} q(\boldsymbol{x})\ d\boldsymbol{x}\right]^{-1} \frac{T_{iR}}{T_{iU}+T_{iR}} \times q(\boldsymbol{x}), & \boldsymbol{x} \in A_{iR}. \end{cases} \tag{5}$$

If T_{iU} and T_{iR} are not known, but the number of enumeration areas within the strata in the area is known, as is the case with the 2014 Kenya DHS at the county level, we can again use $\widehat{T}_{iU}$ and $\widehat{T}_{iR}$ from Appendix A.1 as estimates of the target population totals in the urban and rural strata in area i. Plugging this adjusted density surface

into our area level estimator, we have:

$$p_i = \int_{A_i} p(\boldsymbol{x}) \times \tilde{q}(\boldsymbol{x})\ d\boldsymbol{x} \approx \sum_{j=1}^{m_i} p(\boldsymbol{x}_j) \times \tilde{q}(\boldsymbol{x}_j). \tag{6}$$

This is the foundation of the county level estimator that we use for predictions for the SPDE 'U' models.

In order to account for the effect of cluster level random variation on the expectation of $p(\boldsymbol{x})$ in SPDE 'C' models without knowing exact unobserved EA locations, we integrate out the cluster effect by modifying $p(\boldsymbol{x})$ in (1) and (6) with:

$$p(\boldsymbol{x}) = \int_{-\infty}^{\infty} \text{expit}\{\beta_0 + u(\boldsymbol{x}) + \beta^{\text{URB}} I(\boldsymbol{x} \in U) + \epsilon\} \cdot f(\epsilon)\ d\epsilon, \tag{7}$$

where ϵ is the cluster effect for an arbitrary unobserved cluster at location $\boldsymbol{x}$, and $f(\epsilon)$ is its probability density, which, conditional on the cluster variance σ_ϵ^2, is iid Gaussian with mean zero and variance σ_ϵ^2. Due to the nonlinearity of the expit function, integrating out the cluster effect will shift $p(\cdot)$ towards 0.5 relative to removing the cluster effect when generating predictions.