Small Area Estimation in a Global Health Context

Jon Wakefield[1,2]

[1]Department of Statistics, University of Washington
[2]Department of Biostatistics, University of Washington

IDM Symposium, Wednesday 18th April, 2018

Collaborators:

Geir-Arne Fuglstad, Andrea Riebler, Johnny Paige, Katie Wilson
Sam Clark, Tracy Dong, Jessica Godwin, Richard Li, Yuan Hsiao, Bryan Martin
Danzhen You, Patrick Gerland, Lucia Hug, David Sharrow, Jon Pederson, Ken Hill, Leontine Alkema

Outline

Pixel Surfaces

It is now common to construct spatial surfaces of demographic and health indicators at the "pixel" level:

- Population (Wardrop et al., 2018).
- Malaria (Gething et al., 2016).
- U5MR (Golding et al., 2017).
- Vaccination (Utazi et al., 2018)
- HIV testing in women; stunting in children; anemia in children; household access to improved sanitation (Gething et al., 2015).
- Child growth failure (Osgood-Zimmerman et al., 2018).
- Educational attainment (Graetz et al., 2018).
- . . .

These maps are based, in large part, on data from surveys, often DHS which typically use stratified cluster sampling with the strata usually corresponding to region crossed with urban/rural and households sampled within enumeration areas which are sampled within strata.

Disease mapping $\rightarrow$ SAE

In spatial epidemiology there is a long history of mapping disease rates/risk (particularly cancer) at the areal level:

- Data are based on complete enumeration of cases (and population).
- Smoothing via discrete spatial models is the norm (e.g., Besag et al., 1991; Leroux, 2000); alleviates problems with small numbers of cases for a rare disease.
- Hypothetical risk is usually of primary interest, rather than the true fraction of population that are cases.

Disease mapping $\rightarrow$ SAE

In traditional SAE the aim is to estimate true counts or population averages (e.g., fraction with disease) over a group of domains (areas).

Data arise from surveys, often with a complex design.

Areas historically correspond to administrative regions (in which people live) rather than pixel regions (in many of which, nobody lives).

Traditional SAE (Rao and Molina, 2015) does not emphasize spatial smoothing, so no accepted approach as yet (at least not amongst the statistical community...).

Design-Based Inference

Suppose θ_i is the target of inference in area i (e.g., Admin-1 regions).

Direct Estimation:

- Weighted estimator $\widehat{\theta}_i$ with asymptotic distribution N(θ_i, $\widehat{V}_i$), where $\widehat{V}_i$ is the variance, which acknowledges the design.
- Design is accounted for in estimation by weighting, and in variance calculation.
- Population information is implicit in the weighting, and is not needed for construction of estimate or variance. For simple random sampling:

$$\widehat{\theta}_i = \frac{\sum_{k=1}^{n_i} w_{ik} y_{ik}}{\sum_{k=1}^{n_i} w_{ik}},$$

  with $w_{ik} = N_i/n_i$ where N_i is the population and n_i is the sample size in area i.
- With small samples in an area, instability in estimates/low precision.

Smoothed Direct Estimation (Mercer et al., 2015):

- Smooth direct estimator using disease mapping discrete spatial models.
- Alleviates small sample size problems.

Scaling Up the Smoothed Direct Model (Li et al., 2018)

The smoothed direct model has been used for 35 African countries to estimate U5MR in Admin-1 regions by year.

Includes space-time interactions that cross random walk models in time with ICAR models in space (Knorr-Held, 2000).

Data:

- 121 DHS in 35 countries
- 1.2 million children
- 192 million child-months

UN have supported this research and these estimates.

Takes around 2.5 hours to obtain estimates for all countries – separate models for each country.

Spatial and space-time smoothed direct estimates models are available in `R`, via the `SUMMER` package.

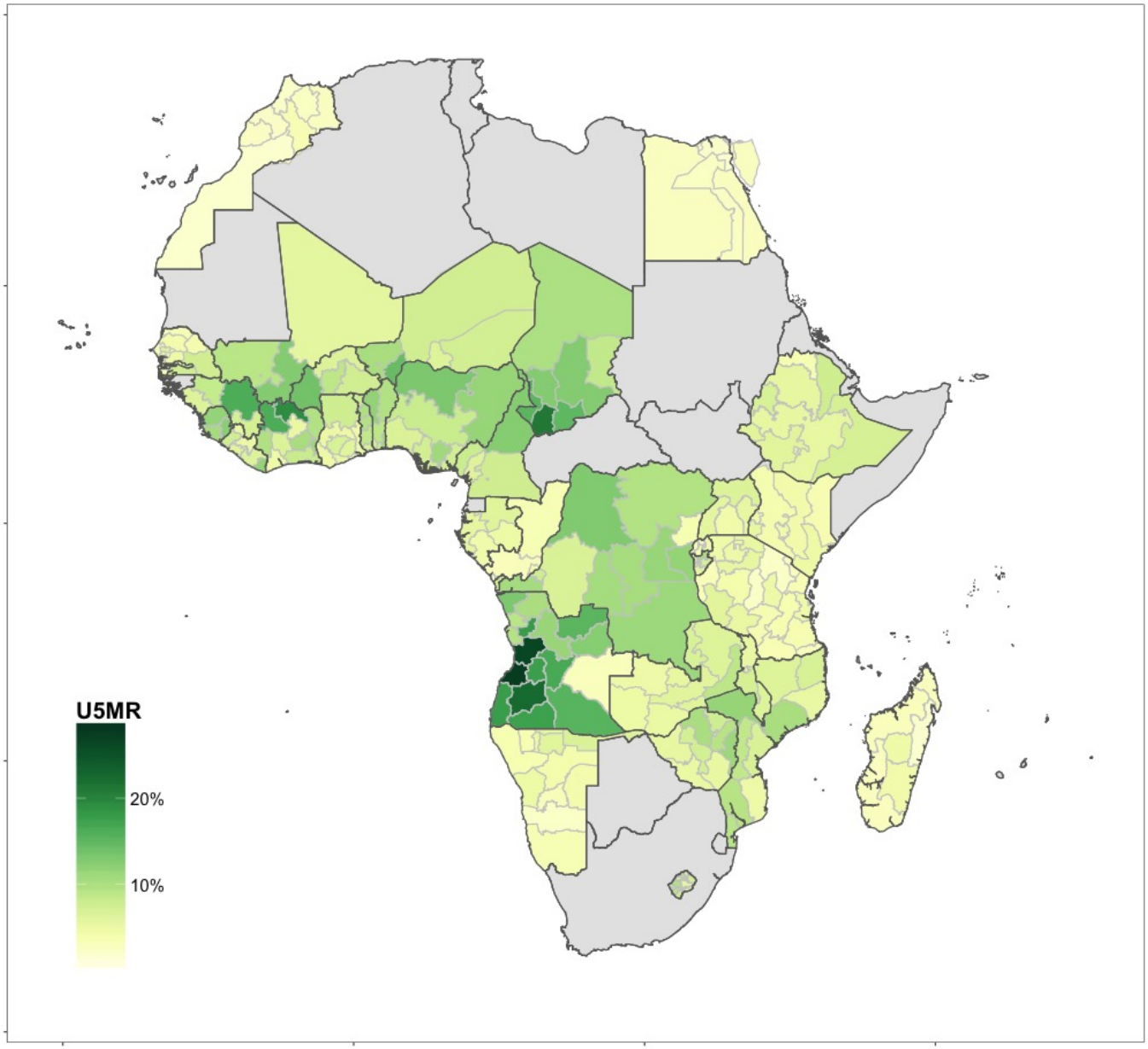

Figure 1: Predictions of U5MR for 2015, in 35 countries of Africa.

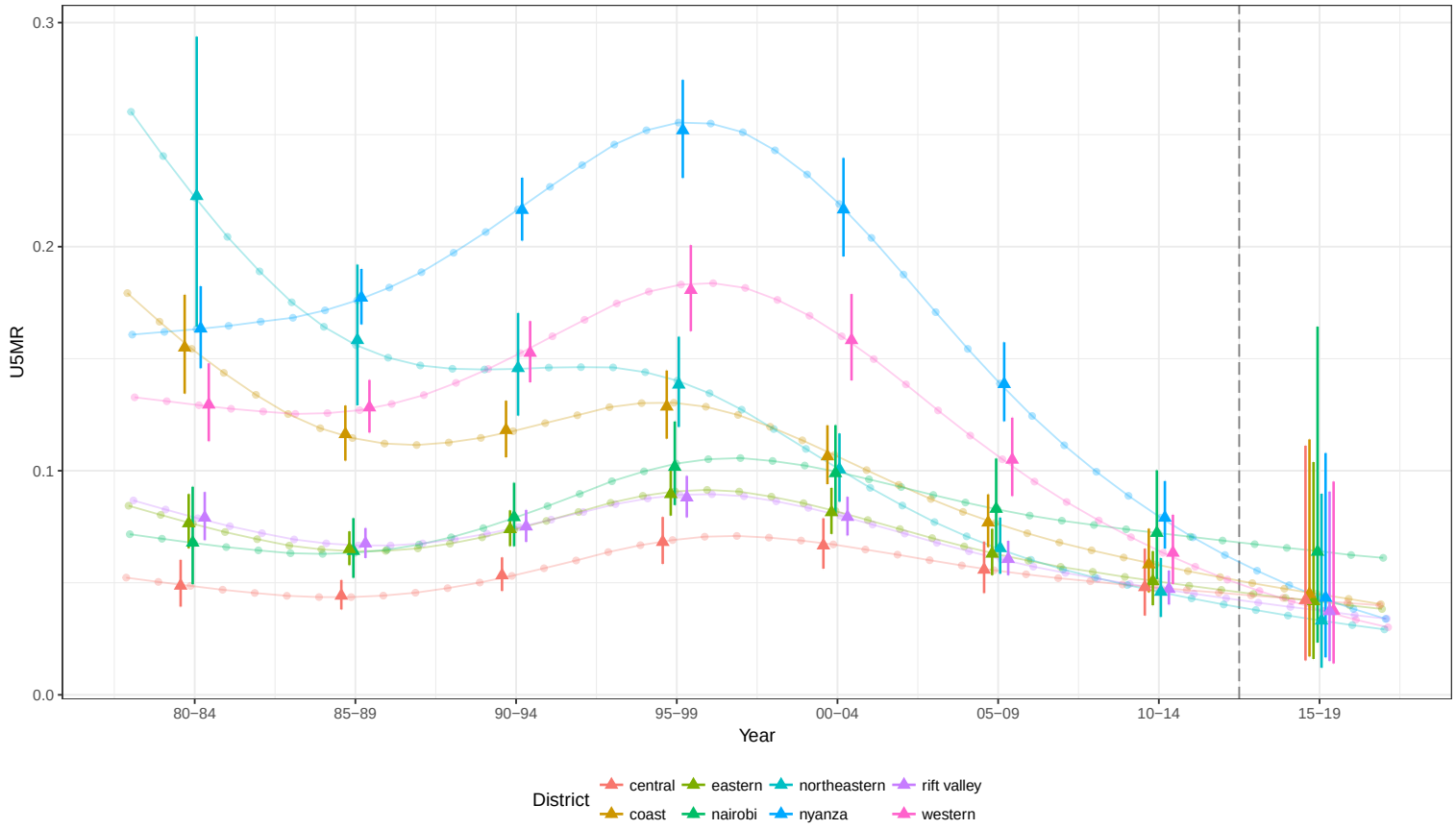

Figure 2: Posterior median estimates for Kenya districts.

Model-Based Inference

For simplicity consider a binary outcome and let Y_{ik} be the number of individuals out of n_{ik} with the characteristic of interest in cluster k of area i.

It has become the norm to ignore stratification and assume the geostatistics model:

$$\begin{aligned} Y_{ik}|\theta_{ik} &\sim \text{Binomial}(n_{ik}, \theta_{ik}) \\ \log\left(\frac{\theta_{ik}}{1-\theta_{ik}}\right) &= \beta_0 + \beta x_{ik} + \epsilon_{ik} + S_{ik}^{\text{CONT}} \end{aligned}$$

where

- $\theta_{ik} = \theta(\boldsymbol{s}_{ik})$ is the risk at location $\boldsymbol{s}_{ik}$,
- x_{ik} are covariates,
- $\epsilon_{ik} \sim \text{N}(0, \sigma_\epsilon^2)$ is the nugget,
- S_{ik}^{CONT} are spatial random effects, assumed to arise from a Gaussian process.

Gething and Burgert-Brucker (2017) reported mixed accuracy for different outcomes using this model (poor for vaccination surfaces, for example).

Model-Based Inference

Alternatively a discrete spatial model can be used:

$$\log\left(\frac{\theta_{ik}}{1-\theta_{ik}}\right) = \alpha + \beta x_{ik} + \epsilon_{ik} + S_i^{\text{DISC}}$$

where

- S_i^{DISC} are discrete spatial random effects that follow an ICAR (Markov Random Field) model (Besag et al., 1991).

For either model, area estimates are obtained by averaging point estimates with respect to the population from:

$$\theta_i = \frac{\int_s \theta(\boldsymbol{s}) d(\boldsymbol{s})\, d\boldsymbol{s}}{\int_s d(\boldsymbol{s})\, d\boldsymbol{s}}$$

where $d(\boldsymbol{s})$ is population density at location $\boldsymbol{s}$.

In practice, the continuous spatial model is always approximated by some form of discretization, so the integral is approximated by summing over a grid.

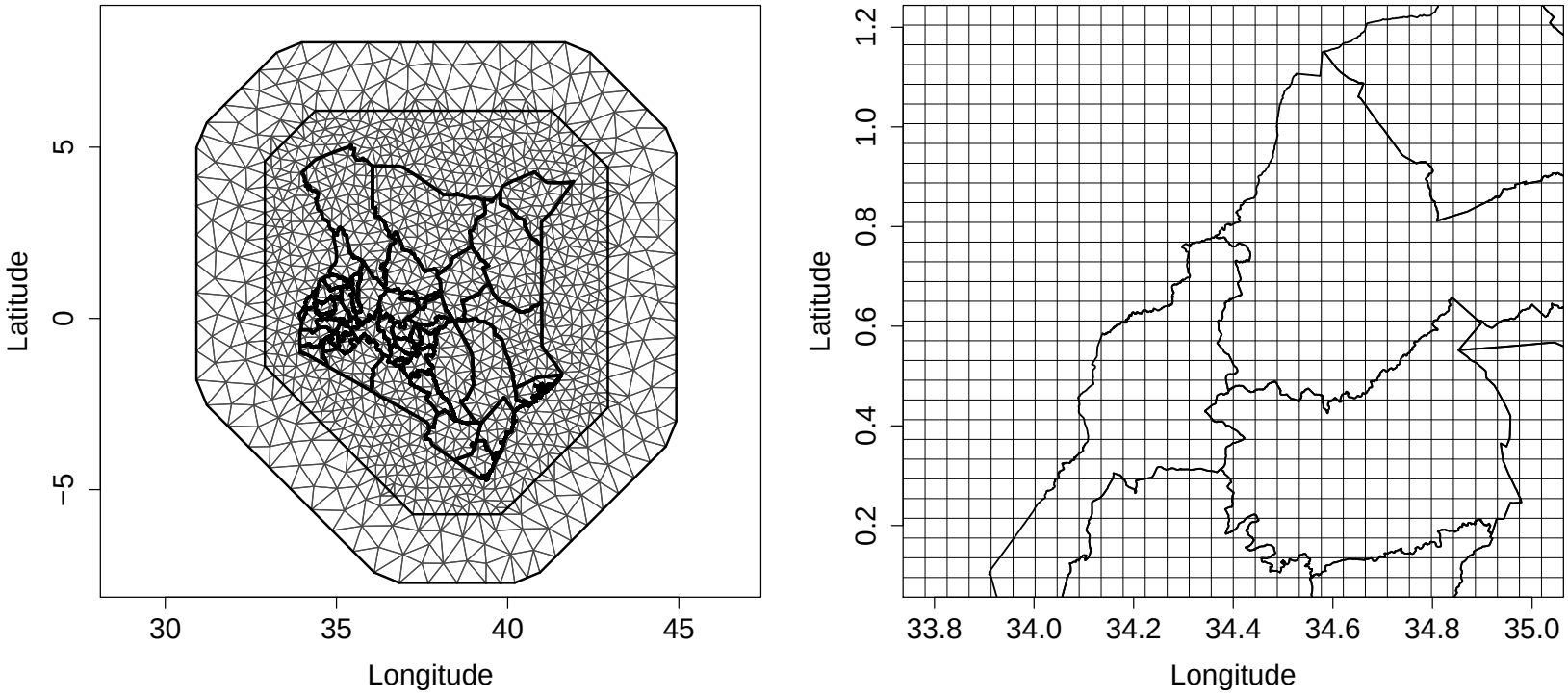

Figure 3: Mesh on which SPDE calculations are carried out (top left), zoomed in grid on which predictions are performed (right).

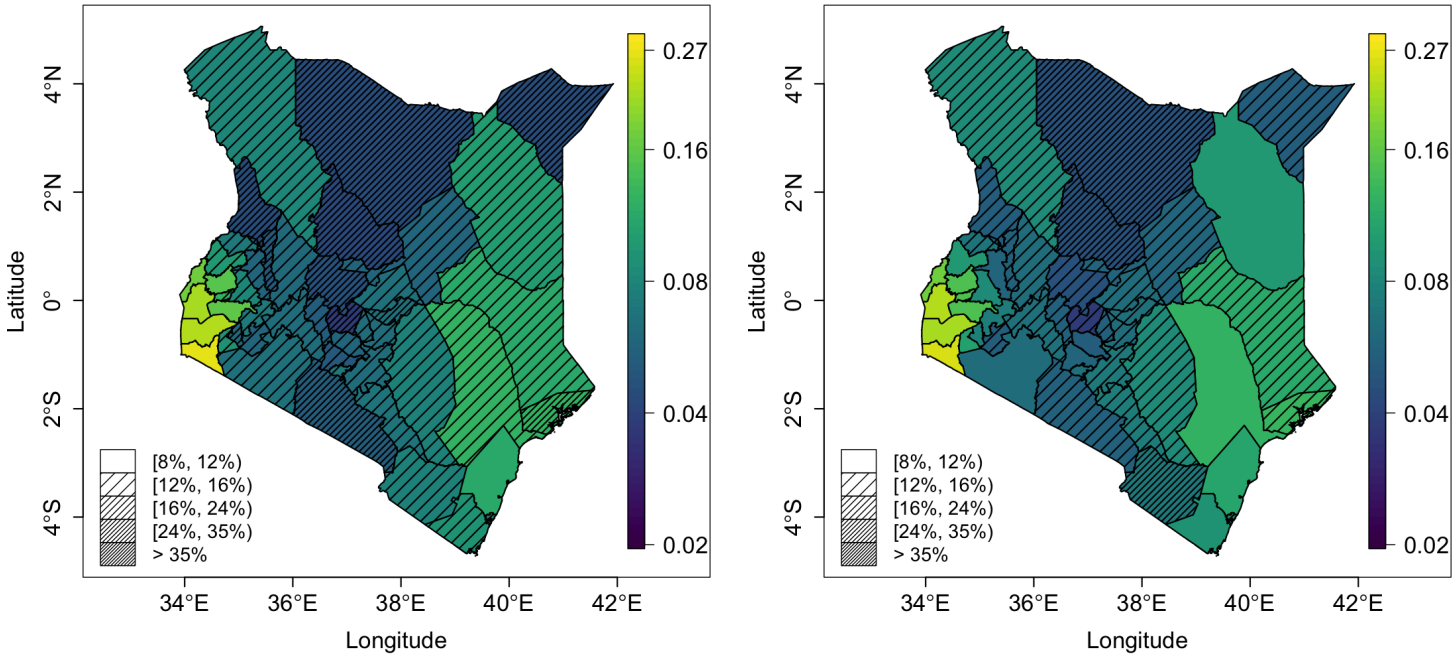

Figure 4: Kenya U5MR estimates in 2000 using discrete spatial model (left), and continuous spatial model (right).

Point estimates are very similar, but more uncertainty associated with the discrete spatial model estimates.

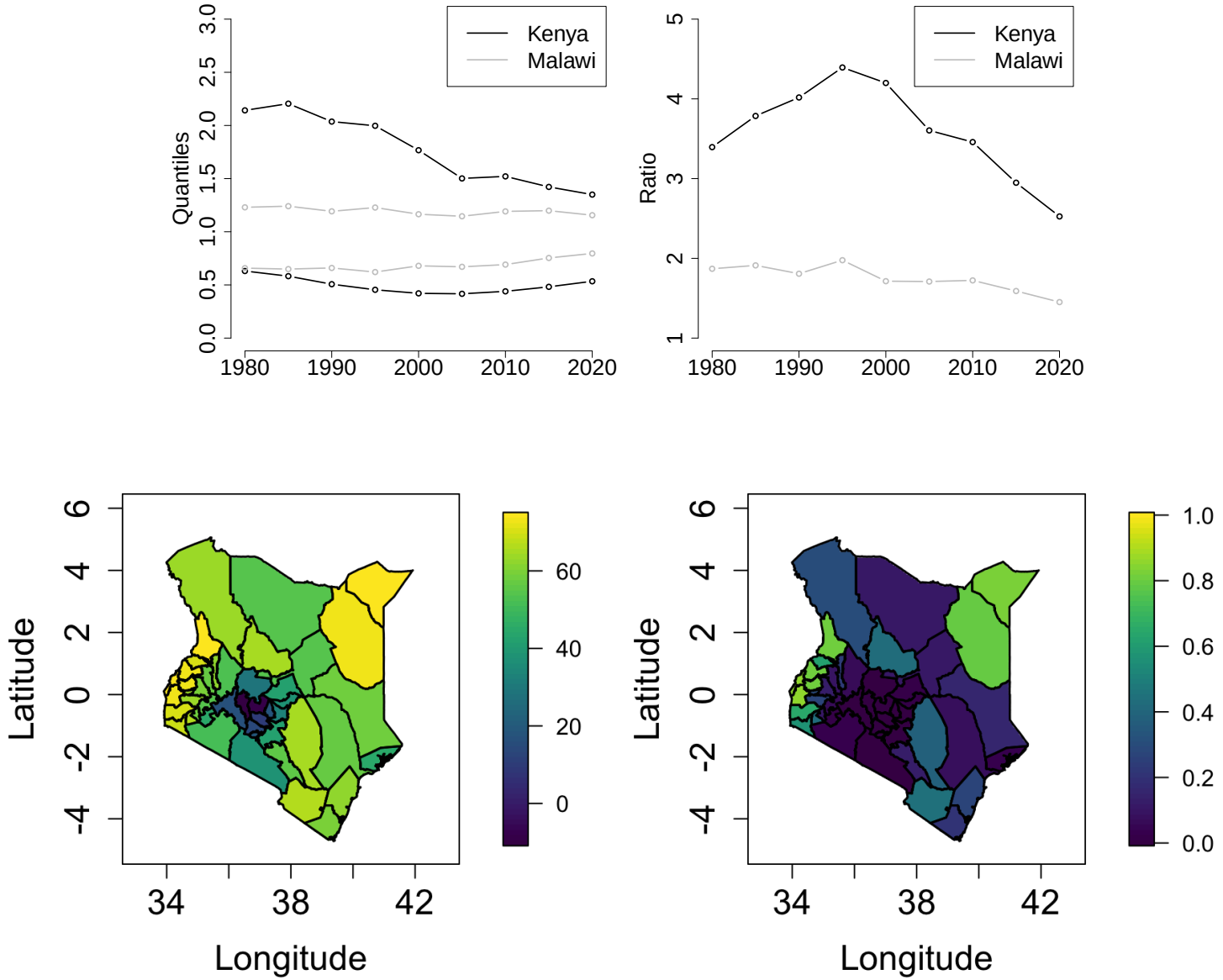

Figure 5: Top row: Kenya and Malawi within-country variability in U5MR (5% and 95% quantiles of pixel distribution). Bottom row: percentage drop from 1990–2015 (left), posterior probability of attaining MDG goal (right).

Comparison of Discrete and Continuous Spatial Models

MSE comparison based on 400 (out of 1600) clusters from 2014 Kenya DHS.

Let:

- $Y_{ip}^{(1)}$ denote the weighted estimator.
- $Y_{ip}^{(2)}$ the smoothed estimator from continuous space model.
- $Y_{ip}^{(3)}$ the smoothed estimator from discrete space model: ICAR $\times$ AR(1), with the latter having yearly resolution,

in county i and period p.

We compare these estimates with the weighted estimates from (approximately) 1200 (left-out) clusters from 2014, y_{ip} (the "truth").

In particular, we calculate,

$$\text{MSE}_p^{(j)} = \frac{1}{47} \sum_{i=1}^{47} \left(Y_{ip}^{(j)} - y_{ip} \right)^2, \tag{1}$$

for $p = \{$1990–1994,1995–1999, 2000–2004, 2005–2009, 2010–2014 $\}$ and $j = 1, 2, 3$.

MSE Comparison

Period	Weighted	Continuous Space	Discrete Space
1990–1994	49	29	29
1995–1999	46	21	21
2000–2004	40	22	22
2005–2009	41	20	20
2009–2014	37	15	15

Table 1: Mean-squared errors ($\times 10^2$) comparing weighted and spatially and temporally smoothed estimates.

Conclusions:

- Spatial models have very similar predictive ability, with the continuous model being slightly more accurate.
- Both show a dramatic improvement over the weighted estimates.

Statistical Issues with Complex Sampling

Ignoring the design leads to the possibility of:

- Bias (if stratification variables are associated with the outcome).
- An inappropriate measure of variance (cluster sampling breaks independence of outcomes).

We report on a limited simulation exercise that investigates the impact of ignoring the design.

As a simple example, suppose the strata are urban/rural.

If we ignore this aspect then

- area-level estimates will be biased unless:
  - the outcome does not depend on strata membership, or
  - sampling of strata is in the same proportion as the population frequencies (so not stratified!).
- pixel-level estimates will be biased unless:
  - the outcome does not depend on strata membership.

Note: If population density and/or travel time are in the covariate model, may get partial correction.

Accounting for Complex Sampling

We consider the simplified situation in which we have:

- A single survey.
- A binary outcome.

Using Kenya geography, we simulate a single complete population:

- Clusters: 96,251 enumeration areas (EAs), 32% are urban.
- Strata used in DHS in 2014 are 47 counties and urban/rural (92 in total, Nairobi and Mombasa are entirely urban).
- From the Kenya 2014 DHS report we know the numbers of urban/rural EAs by district and we match these numbers by thresholding on a population density surface.
- Within each EA, assume 25 households, with one mother in each household and one birth per mother.

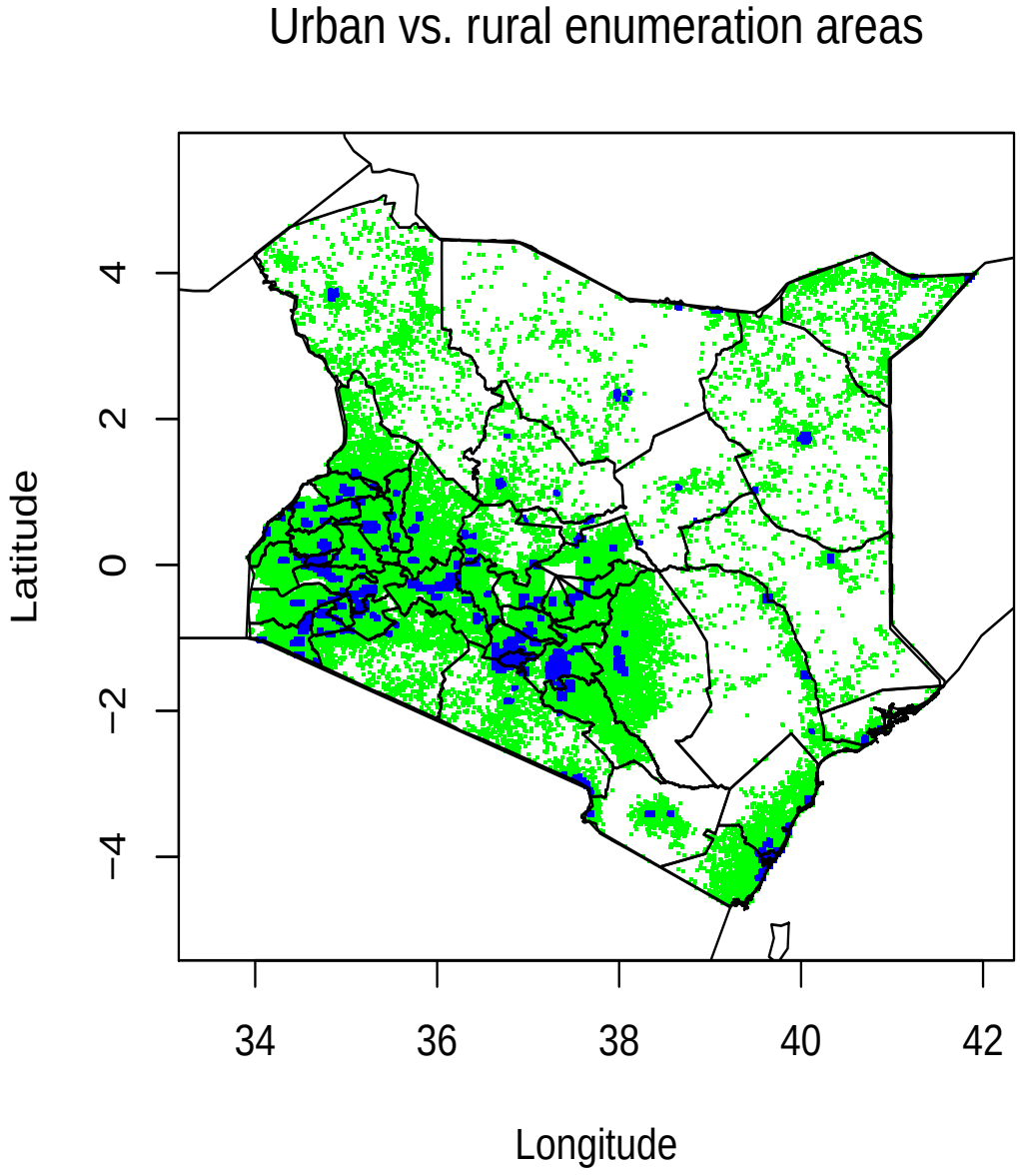

Figure 6: Sampling frame for Kenya simulation.

Accounting for Complex Sampling

We have $n_j = 25$ births at each EA (cluster) location $\boldsymbol{s}_j$, $j = 1, \ldots, n$, and we generate neonatal deaths Y_j according to

$$\begin{aligned} Y_j|\theta(\boldsymbol{s}_j) &\sim \text{Binomial}\,(n_j, \theta(\boldsymbol{s}_j)) \\ \log\left(\frac{\theta(\boldsymbol{s}_j)}{1-\theta(\boldsymbol{s}_j)}\right) &= \beta_0 + \gamma I(\boldsymbol{s}_j \in \text{ urban }) + \epsilon_j + S(\boldsymbol{s}_j), \end{aligned}$$

where

- $\epsilon_j \sim_{iid} \text{N}(0, \tau^2)$ (the nugget),
- $S(\boldsymbol{s})$ is a Gaussian Process (GP) with Matérn covariance function and (effective) range ϕ and variance σ^2.

The nugget term induces within-cluster dependence.

Accounting for Complex Sampling

Assume inference is at the county level.

Methods to be compared:

- Naive: Assume binomial (unweighted) counts in each county. This gives an estimate θ_i^{BIN} and a variance from which an asymptotic CI can be calculated.
- Direct estimates: This gives an estimate θ_i^{DIR} and a variance from which an asymptotic CI can be calculated.

Accounting for Complex Sampling

- Smoothed Direct: Take logit of direct estimates θ_i^{DIR} with appropriate design-based estimator and model as Mercer et al. (2015),

$$\begin{aligned} \text{logit}(\theta_i^{\text{DIR}}) &\sim \text{N}(\eta_i, \widehat{V}_{\text{DES},i}) \\ \eta_i &= \beta_0 + \underbrace{\epsilon_i}_{\text{Independent}} + \underbrace{S_i}_{\text{ICAR}} \end{aligned}$$

  County smoothed direct estimate

$$\widehat{\theta}_i^{\text{SDIR}} = \text{expit}(\widehat{\beta}_0 + \widehat{\epsilon}_i + \widehat{S}_i).$$

Accounting for Complex Sampling

- Smoothed Adjusted Discrete Spatial Model at the cluster level:

$$\begin{aligned} Y_j|\theta_j &\sim \text{Binomial}(n_j, \theta_j) \\ \text{logit}(\theta_j) &= \beta_0 + \gamma I(\boldsymbol{s}_j \in \text{ urban }) + \underbrace{\epsilon_{i[j]}}_{\text{Independent}} + \underbrace{S_i}_{\text{ICAR}} + \underbrace{\delta_j}_{\text{Independent}} . \end{aligned}$$

  Obtain 2 estimates for each county i:

$$\begin{aligned} \widehat{\theta}_{i1} &= \text{expit}(\widehat{\beta}_0 + \widehat{\epsilon}_i + \widehat{S}_i) \\ \widehat{\theta}_{i2} &= \text{expit}(\widehat{\beta}_0 + \widehat{\gamma} + \widehat{\epsilon}_i + \widehat{S}_i) \end{aligned}$$

  Then

$$\widehat{\theta}_i = q_i \widehat{\theta}_{i1} + (1 - q_i)\widehat{\theta}_{i2}$$

  where q_i is the proportion of the births that occur in rural clusters.

- Smoothed Adjusted Continuous Spatial Model at the cluster level:

$$\begin{aligned} Y_j|\theta_j &\sim \text{Binomial}(n_j, \theta_j) \\ \text{logit}(\theta_j) &= \beta_0 + \gamma I(\boldsymbol{s}_j \in \text{ urban }) + \underbrace{\epsilon_j}_{\text{Independent}} + \underbrace{S_j}_{\text{GP}} \end{aligned}$$

Accounting for Complex Sampling

Methods comparison: bias, MSE, Average of Variance, 80% CI coverage.

Parameters (in all simulations):

- $\beta_0 = -2$, $\gamma = -0.5$ (so urban lower)
- $\sigma^2 = 0.15^2$, effective range $\phi = 300\,\text{km}$, $\tau^2 = 0.1^2$.

Two simulations:

1. Unstratified sampling.
2. Stratified sampling in which we oversample urban clusters. Specifically, in each county sample twice as many urban as rural clusters.

Results[1]

- Unstratified sampling:

Method	Bias	MSE	Ave. Var.	80% coverage
Naive	-0.020	0.060	0.051	0.78
Direct estimates	-0.020	0.060	0.053	0.75
Smoothed Direct	0.012	0.018	0.018	0.78
Discrete Spatial	-0.014	0.011	0.015	0.84
Continuous Spatial	-0.005	0.012	0.010	0.72

- Stratified sampling:

Method	Bias	MSE	Ave. Var.	80% coverage
Naive	-0.082	0.069	0.053	0.75
Direct estimates	-0.029	0.066	0.058	0.73
Smoothed Direct	0.005	0.021	0.020	0.78
Discrete Spatial	-0.015	0.011	0.016	0.86
Continuous Spatial	-0.005	0.012	0.010	0.72

To be continued...

[1] Bias is logit $\widehat{\theta}_i$ − logit θ_i where θ_i is truth

Model Validation

No consensus on how to validate model, cross-validation is the most common approach, but details on how splits were made often sketchy, as are exact ways in which predictions obtained (supplementary materials hide many sins...).

When bias is reported, what is the "truth"?

By construction, spatial models smooth the covariate mean in areas with no data.

Wakefield et al. (2018) compared predictions for U5MR in Kenya from discrete and continuous spatial models:

- "Truth" (direct estimates with small variance) is only available at Admin-1, 5-year scale.
- Discrete and continuous models performed equally well, but below Admin-1, who knows?

Now investigating the use of proper scoring rules (Gneiting and Raftery, 2007).

Model Validation

When interpreting surfaces based on DHS data, should also bear in mind:

- Jittering (Gething et al., 2015).
- Boundary changes.
- Migration.
- Recall bias.
- Non-response.
- Linear systematic sampling (explicit stratification).
- Every country has its own idiosyncrasies.

Covariate Modeling

Distinguish between:

- Individual-level modeling, for example, for U5MR, Balk et al. (2004).
- Surface modeling, in which we require covariates to be available at all prediction points.

Some approaches:

- Often some kind of backward elimination (e.g., Utazi et al., 2018) or all subsets (e.g., Gething et al., 2015).
- Stacked generalization/super learner (Bhatt et al., 2017; Golding et al., 2017).

In general, inference/uncertainty estimates do not correctly account for the selection of the final covariate model.

Discussion: Comparison of Models

	Direct Estimation	Smoothed Direct	Discrete Spatial	Continuous Spatial
Robustness	✓✓✓✓	✓✓✓	✓✓	✓
Transparency	✓✓✓✓	✓✓✓	✓✓	✓
Sparse Data	✓	✓✓	✓✓✓✓	✓✓✓✓
Spatial Scale	✓	✓	✓✓✓✓	✓✓✓✓
Data Required	✓✓✓✓	✓✓✓✓	✓✓✓	✓✓
Flexibility	✓	✓✓	✓✓✓	✓✓✓✓

Table 2: Comparison of approaches to SAE.

General strategy: See if estimates from different models are consistent with each other.

There is some skepticism of even national estimates (e.g., Boerma et al., 2018), let alone SAE or pixel level estimation.

Discussion

Substantive:

- Follow-up to Admin-1 in sub-Saharan Africa paper: Admin-2 including summary birth history data.
- Asia at Admin-1.
- Examination of biases in DHS data.
- Measles: modeling vaccination coverage and spatio-temporal disease count data.

Methodological:

- Consensus on pixel modeling.
- Modeling summary birth history.
- Examination of implications of ignoring the design.
- Points/polygons problem.
- Examination of model validation techniques.
- Covariate modeling (how to use information on conflicts?).
- Spatial APC models with survey data.

References I

Balk, D., T. Pullum, A. Storeygard, F. Greenwell, and M. Neuman (2004). A spatial analysis of childhood mortality in West Africa. *Population, Space and Place 10*, 175–216.

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

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

Boerma, T., C. Victora, and C. Abouzahr (2018). Monitoring country progress and achievements by making global predictions: is the tail wagging the dog? *The Lancet*.

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. and C. R. Burgert-Brucker (2017). The DHS program modeled map surfaces: understanding the utility of spatial interpolation for generating indicators at subnational administrative levels.

References II

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, et al. (2016). Mapping plasmodium falciparum mortality in africa between 1990 and 2015. *New England Journal of Medicine 375*(25), 2435–2445.

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 neonatal 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, et al. (2018). Mapping local variation in educational attainment across africa. *Nature 555*, 48.

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

References III

Leroux, B. (2000). Modeling spatial disease rates using maximum likelihood. *Statistics in Medicine 19*, 2321–2332.

Li, R., Y. Hsiao, J. Godwin, B. B. Martin, J. Wakefield, and S. Clark (2018). 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. *Submitted*.

Mercer, L., J. Wakefield, A. Pantazis, A. Lutambi, H. Mosanja, and S. Clark (2015). Small area estimation of childhood 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, et al. (2018). Mapping child growth failure in africa between 2000 and 2015. *Nature 555*(7694), 41.

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

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.

References IV

Wakefield, J., G.-A. Fuglstad, A. Riebler, J. Godwin, K. Wilson, and S. J. Clark (2018). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*. To Appear.

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*(14), 3529–3537.