2022 SISMID Module 9
Lecture 10: Prevalence Mapping

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Prevalence Mapping

- Prevalence is defined as the proportion of a population who have a specific characteristic in a given time period.
- Public health targets are often expressed as prevalences. For example, the Sustainable Development Goals (SDGs) have a number of such targets including Goal 3.2 which states:

  > *"By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births".*

- As data availability escalates, there has been a corresponding increase in the production of maps displaying the prevalence of a large range of health and demographic outcomes — an endeavor that has been labeled prevalence mapping.

Prevalence Mapping

- In this lecture we distinguish between modeling at the area level, which is often dubbed small area estimation (SAE) (Rao and Molina, 2015; Pfeffermann, 2013), and at the point level, which is referred to as model-based geostatistics (MBG) (Diggle and Giorgi, 2019).
- Often, maps are produced for low- and middle-income countries based on household surveys which have complex designs.
- Bayesian smoothing models are convenient for both SAE and MBG, and computation is no longer a major problem for most prevalence mapping endeavors.
- We focus on MBG here, but briefly discuss SAE techniques.

References: Wakefield (2020); Wakefield *et al.* (2020) – these papers may be found at
`http://faculty.washington.edu/jonno/space-station.html`

Prevalence Mapping

- Prevalence mapping allows fundamental questions such as: “how many people in my area have condition X or need treatment Y.”
- Disease mapping is traditionally based on a complete enumeration of disease cases, and may differ from prevalence mapping which may be based on a subset of individuals, selected via a survey, which may have a complex design.

Motivating Example

- We take as a motivating example SAE of HIV prevalence among females aged 15–29, in districts of Malawi, using data from the 2015–16 Malawi DHS.
- We will refer to the Malawi districts as admin-2 areas; there are 3 admin-1 areas, 28 admin-2 areas and 243 admin-3 areas.
- A two-stage stratified cluster sample was implemented, with the sampling clusters (enumeration areas) being stratified by district and urban/rural.
- The Malawi Population and Housing Census (MPHC), conducted in Malawi in 2008 provided the sampling frame for the survey (Malawi DHS, 2016)..
- The sample for the 2015–16 Malawi DHS was designed to provide estimates of key indicators for the country as a whole, for urban and rural areas separately, and for each of the 28 districts.

Motivating Example

- The sampling frame contained 12,558 clusters and our analyses use data from 827 sampled clusters (the supplementary materials give more details). In the 2015–16 DHS survey for Malawi, 8,497 women in the age range 15–49 were eligible for HIV testing, and 93% of them were tested.
- HIV prevalence data was obtained from voluntarily taken blood samples from survey respondents.

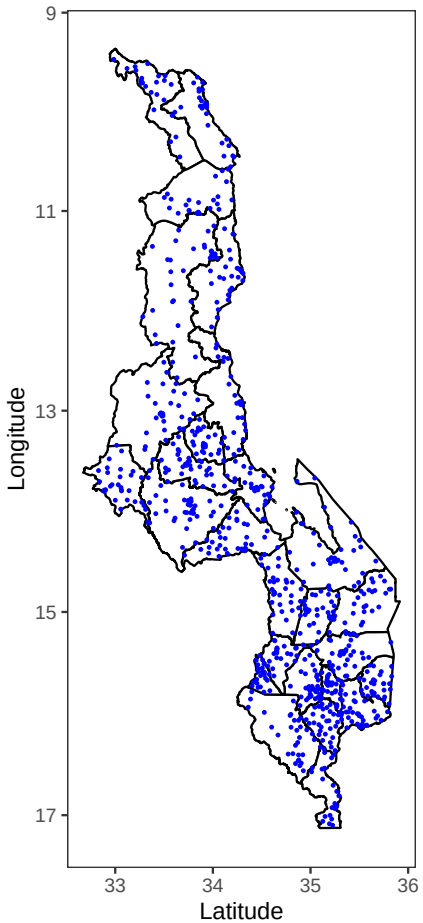


Figure 1: Cluster locations in 2015–16 Malawi DHS.

Area-Level Models

Weighted Estimate

- In a potentially complex survey situation, let π_{ik} be the probability of selection for individual k in area i.
- Let $d_{ik} = 1/\pi_{ik}$ be the design weight associated with individual k in area i, whose response is y_{ik}.
- Within area i, the design-based weighted (direct) estimator (Horvitz and Thompson, 1952; Hájek, 1971) is

$$\widehat{m}_i^{\mathrm{HT}} = \frac{\sum_{k \in s_i} d_{ik} Y_{ik}}{\sum_{k \in s_i} d_{ik}}. \tag{1}$$

- The variance of the estimator, $V_i^{\star}$, may be calculated using standard methods.
- For simple random sampling (SRS), this estimator simplifies to the sample mean.
- In the SAE literature, this is known as a direct estimator.

Area-Level Models

- In a major advance, Fay and Herriot (1979) introduced a very clever approach that models a transform of the weighted estimate, in order to gain precision by using a random effects model.
- For binary outcomes (for example, HIV positive/negative), one choice of transform is $Z_i = \text{logit}\left(\widehat{m}_i^{\text{HT}}\right)$.
- We denote the associated design-based variance of Z_i by V_i.

Area-Level Models

- An area-level model is,

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

where θ_i is the logit of the true proportion in area i, and V_i is the variance of the logit estimator (obtained from $V_i^\star$ via the delta method).

- We model θ_i via a BYM2 specification:

$$\theta_i = \beta_0 + \beta_1 x_i + b_i,$$

where

$$b_i = e_i + S_i$$

with $e_i \sim_{iid} \mathrm{N}(0, \sigma_e^2)$ and $[S_1, \ldots, S_n] \sim \mathrm{ICAR}(\sigma_s^2)$.

- This model produces what is termed a smoothed direct estimator.
- In the HIV prevalence example, we use the HIV prevalence from antenatal care (ANC) clinics, as covariate.

The Data Aggregated to Districts

Region	HIV +ve	No. Tested	Sampled Clusters		Sampling Frame	
			Urban	Rural	Urban	Rural
Balaka	13	176	6	24	17	275
Blantyre	19	185	19	16	412	381
Chikwawa	4	136	4	27	16	380
Chiradzulu	10	132	2	27	2	334
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Rumphi	8	130	6	20	12	156
Salima	5	168	6	23	22	416
Thyolo	8	177	4	30	12	674
Zomba	19	194	9	26	79	584
Total	278	4427	168	659	1409	11149

Table 1: Summary statistics of Malawi 2015–16 DHS data, by district. These summaries are for females aged 15–29.

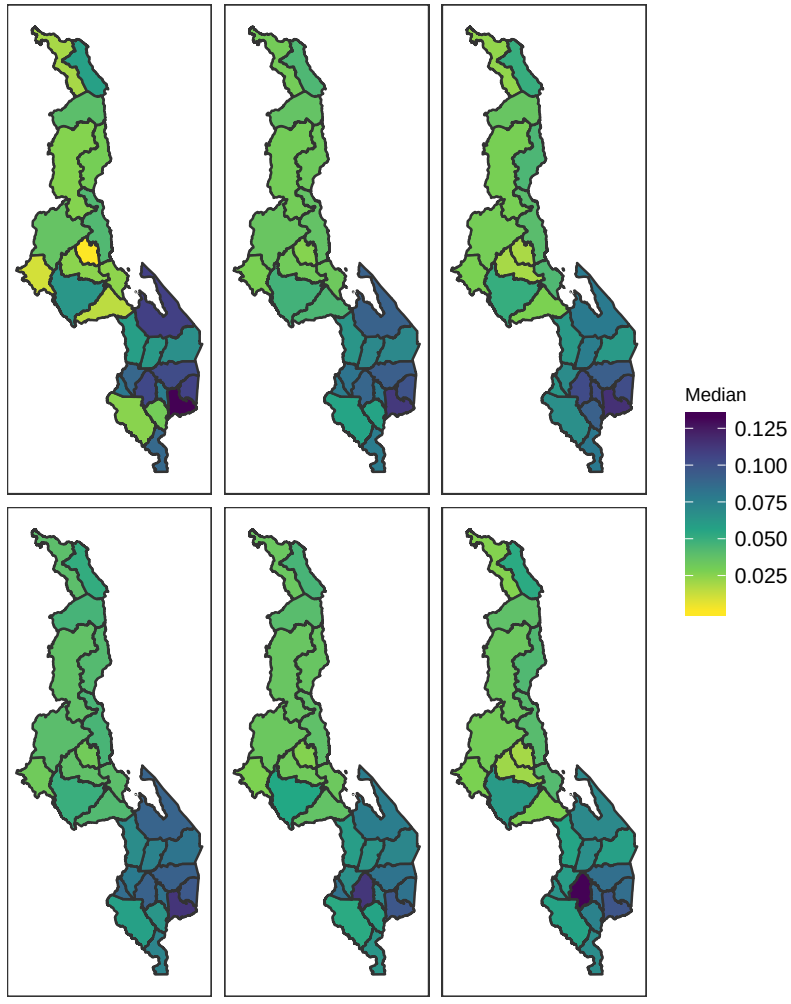


Figure 2: Estimates of HIV prevalence among females aged 15–29 in districts of Malawi in 2015–16. Top row estimates are from area-level models: direct estimates; smoothed direct estimates; smooth direct estimates with antenatal care (ANC) HIV prevalence covariate. Bottom row estimates are from unit-level models: no urban/rural adjustment and no covariate; urban/rural adjustment only; urban/rural adjustment and ANC HIV prevalence covariate.

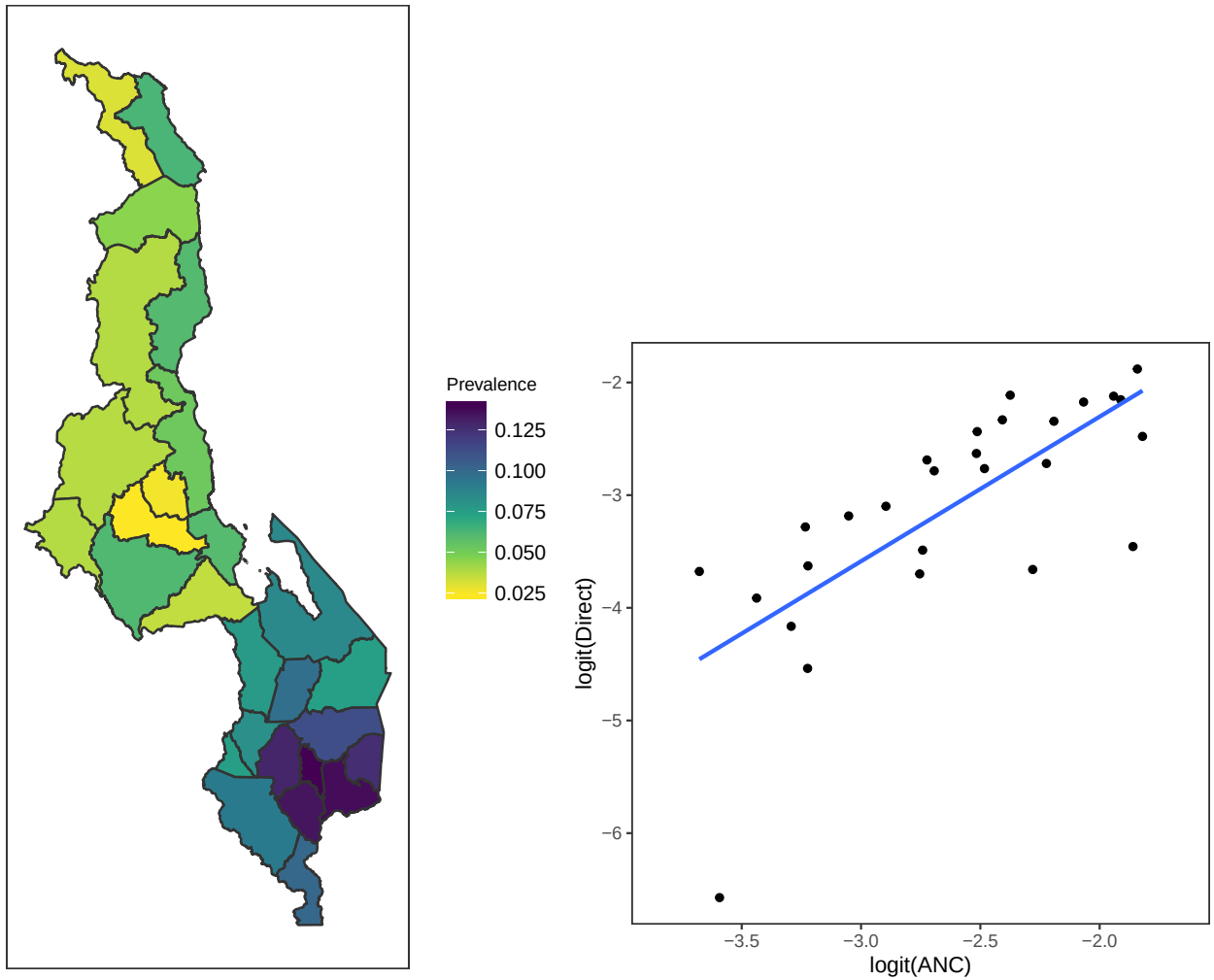


Figure 3: Left: Map of ANC prevalence. Right: logit of direct prevalence estimates versus logit of ANC prevalence estimates.

Model		2.5%	Median	97.5%
No Covariates				
	BYM2 total variance	0.07	0.19	0.48
	Proportion spatial	0.14	0.57	0.94
logit(ANC)				
	BYM2 total variance	0.00	0.04	0.19
	Proportion spatial	0.01	0.17	0.85
	logit(ANC): odds ratio	1.59	2.72	4.03

Table 2: Posterior quantiles for the area-level smoothed direct models. The BYM2 total variance is σ_b^2, the proportion spatial is ϕ, and the logit ANC (odds ratio) is $\exp(\beta_1)$.

The linear predictor is:

$$\theta_i = \beta_0 + x_i^\top \beta_1 + \underbrace{e_i + S_i}_{b_i},$$

with total residual variation σ_b^2 and proportion spatial ϕ.

Unit-Level Models

Unit-Level Model

- We let $\boldsymbol{s}_{ic}$ represent the geographical location of cluster c in area i, and explicitly index the counts and sample sizes as Y_{ic} and n_{ic}, respectively, for $c = 1, \ldots, C_i$, $i = 1, \ldots, m$.
- A crucial assumption here (Rao and Molina, 2015, Section 4.3) is that the probability of selection, given covariates, does not depend on the values of the response.
- This implies that if stratified random sampling is used, stratification variables must be included in the model.
- One would expect cluster sampling to lead to correlated responses within clusters, and cluster-level random effects are introduced to accommodate this aspect (Scott and Smith, 1969).

Unit-Level Model

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

$$Y_{ic}|p_{ic} \sim \text{Binomial}(n_{ic}, p_{ic}). \tag{2}$$

- One candidate model to accompany (2) is,

$$p_{ic} = \text{expit}\left(\beta_0 + \boldsymbol{x}(\boldsymbol{s}_{ic})^{\mathsf{T}}\boldsymbol{\beta}_1 + z(\boldsymbol{s}_{ic})\gamma + S(\boldsymbol{s}_{ic}) + \epsilon_{ic}\right) \tag{3}$$

  where
  - $z(\boldsymbol{s}_{ic})$ represents the strata within which cluster c lies,
  - $\exp(\gamma)$ is the associated odds ratio, and $\boldsymbol{x}(\boldsymbol{s}_{ic})$ are covariates available at location $\boldsymbol{s}_{ic}$, with odds ratios $\exp(\boldsymbol{\beta}_1)$.
  - The spatial random effect $S(\boldsymbol{s}_{ic})$ is associated with cluster location $\boldsymbol{s}_{ic}$, and may be continuous or discrete.
  - The cluster-level error $\epsilon_{ic} \sim \text{N}(0, \sigma^2_\epsilon)$ is the so-called *nugget*, which is traditionally vaguely specified as representing short scale variation and/or "measurement error".

Unit-Level Model

- A model-based geostatistics (MBG) model takes $S(\boldsymbol{s}_{ic})$, as a realization of a zero-mean Gaussian process (GP).
- GP models are common choices for continuously-indexed spatial models and imply that any collection of spatial random effects have a multivariate normal distribution.
- A popular choice for the variance-covariance is the Matérn covariance function (Stein, 1999), for which the covariance is,

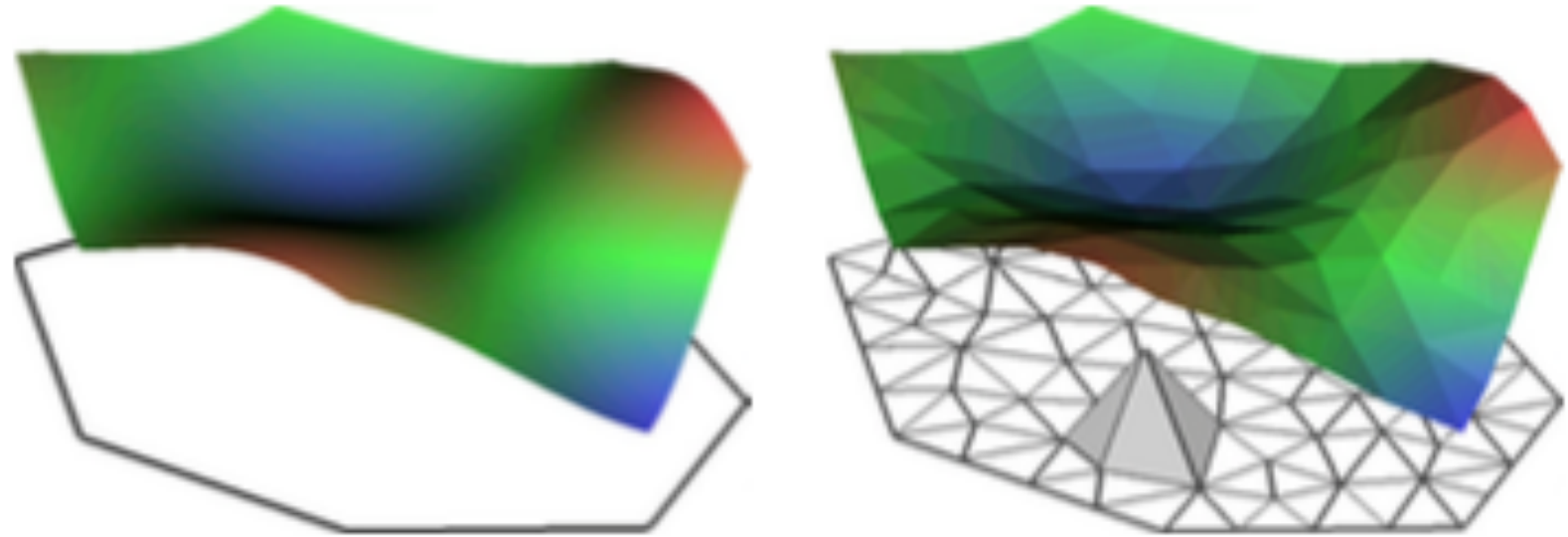
$$\text{cov}(S(\boldsymbol{s}_1), S(\boldsymbol{s}_2)) = \sigma_s^2 \frac{2^{1-\nu_s}}{\Gamma(\nu_s)} \left(\sqrt{8\nu_s} \frac{||\boldsymbol{s}_2 - \boldsymbol{s}_1||}{\rho_s} \right) \mathrm{K}_{\nu_s} \left(\sqrt{8\nu_s} \frac{||\boldsymbol{s}_2 - \boldsymbol{s}_1||}{\rho_s} \right),$$

where

  - ρ_s is the spatial range corresponding to the distance at which the correlation is approximately 0.1,
  - σ_s is the spatial standard deviation,
  - ν_s is the smoothness (which is usually fixed, since it is difficult to estimate), and
  - K_{ν_s} is a modified Bessel function of the second kind, of order ν_s.

Computation for unit-level model

- When the number of clusters $C = \sum_{i=1}^{m} C_i$ is large, computation is an issue, because we need to manipulate $C \times C$ matrices which involves $O(C^3)$ operations (Rue and Held, 2005).
- Various approximations have been proposed to overcome this problem, for example, the stochastic partial differential equations (SPDE) approach pioneered by Lindgren *et al.* (2011) – this is the approach we use for the HIV prevalence example.
- Other approaches are described by Heaton *et al.* (2018).



(a) A continuous function. (b) A piecewise linear approximation.

Fig. 2. Piecewise linear approximation of a function over a triangulated mesh.

Figure 4: GMRF representation of a Markovian GRF, via triangulation, from Simpson *et al.* (2012). Used in the SPDE approach.

Unit-Level Model

- Aggregation to the area-level is carried out via,

$$p_i = \int_{A_i} p(\boldsymbol{s}) \times q(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{l=1}^{M_i} p(\boldsymbol{s}_l) \times q(\boldsymbol{s}_l) \tag{4}$$

where

- the point level risk is,

$$p(\boldsymbol{s}) = \text{expit}(\beta_0 + \boldsymbol{x}(\boldsymbol{s})^{\mathsf{T}}\boldsymbol{\beta}_1 + z(\boldsymbol{s}_{ic})\gamma + S(\boldsymbol{s}))$$

  is the risk at location $\boldsymbol{s}$ (the nugget is, for better or worse, frequently left out, since it is viewed as measurement error) and
- $q(\boldsymbol{s})$ is the population density at $\boldsymbol{s}$, which is needed at all locations on the approximating mesh, $\boldsymbol{s}_l$, $l = 1, \ldots, M_i$.

Unit-Level Model

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

$$Y_{ic} \mid p_{ic}, \lambda \sim \text{BetaBinomial}(n_{ic}, p_{ic}, \lambda) \quad (5)$$

$$p_{ic} = \text{expit}\left(\beta_0 + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_1 + z_{ic}\gamma + e_i + S_i\right) \quad (6)$$

where

- λ is the overdispersion parameter and
- we have taken the spatial random effect to be decomposed as $S(\boldsymbol{s}_{ic}) = e_i + S_i$, with e_i and S_i iid and ICAR, respectively, i.e., a BYM2 model.

Malawi HIV Prevalence Example

Model		2.5%	Median	97.5%
No Covariates				
	Overdispersion	0.01	0.02	0.05
	BYM2 total variance	0.05	0.14	0.35
	Proportion spatial	0.15	0.62	0.96
U/R In				
	Overdispersion	0.01	0.02	0.04
	BYM2 total variance	0.05	0.13	0.33
	Proportion spatial	0.20	0.71	0.98
	Urban: odds ratio	1.73	2.29	3.00
U/R In, logit(ANC)				
	Overdispersion	0.01	0.02	0.04
	BYM2 total variance	0.00	0.02	0.12
	Proportion spatial	0.01	0.22	0.91
	Urban: odds ratio	1.70	2.24	2.94
	logit(ANC): odds ratio	1.59	2.32	3.35

Table 3: Posterior quantiles for the unit-level betabinomial models. The overdispersion parameter is λ, BYM2 total variance is σ_b^2, the proportion spatial is ϕ, the odds ratio associated with an urban cluster is $\exp(\gamma)$, and the logit ANC odds ratio is $\exp(\beta_1)$.

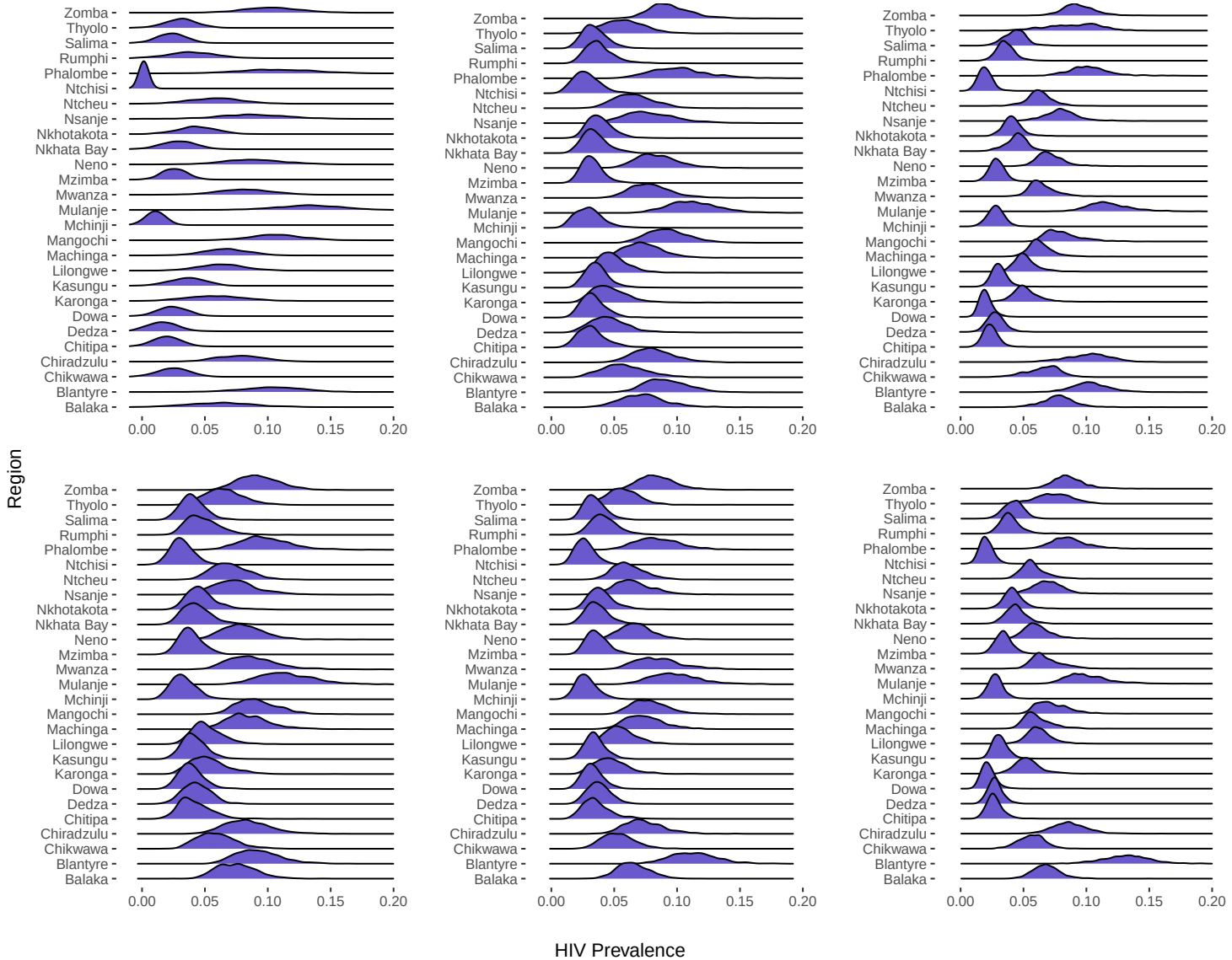


Figure 5: Posterior distributions for HIV prevalence. Top row area-level models: direct; smoothed direct; smoothed direct with ANC covariate. Bottom row unit-level (betabinomial) models: no urban/rural, no covariate; urban/rural only; urban/rural and ANC covariate.

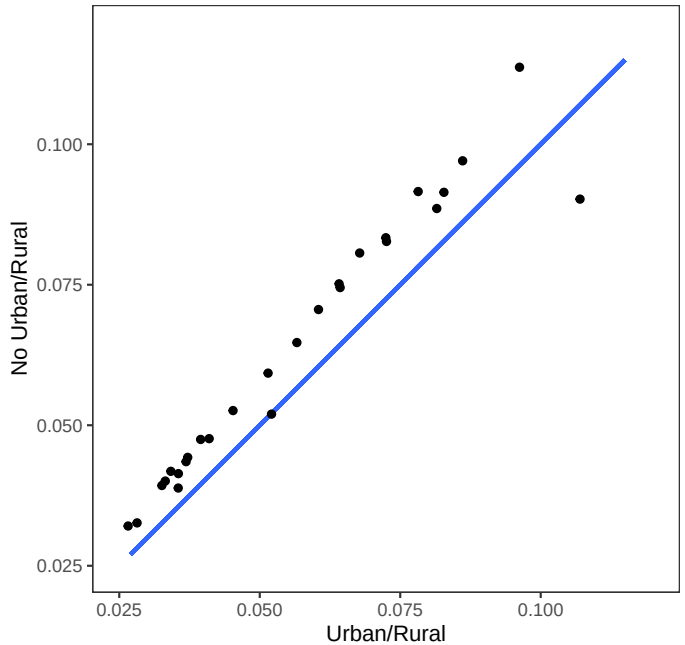


Figure 6: District prevalence estimates from two unit-level models. On the y-axis, the prevalence estimates are from a model with no urban/rural adjustment, while on the x-axis the model has an adjustment.

The estimates from the no adjustment model are too high because of the oversampling of urban areas, which have higher HIV prevalence.

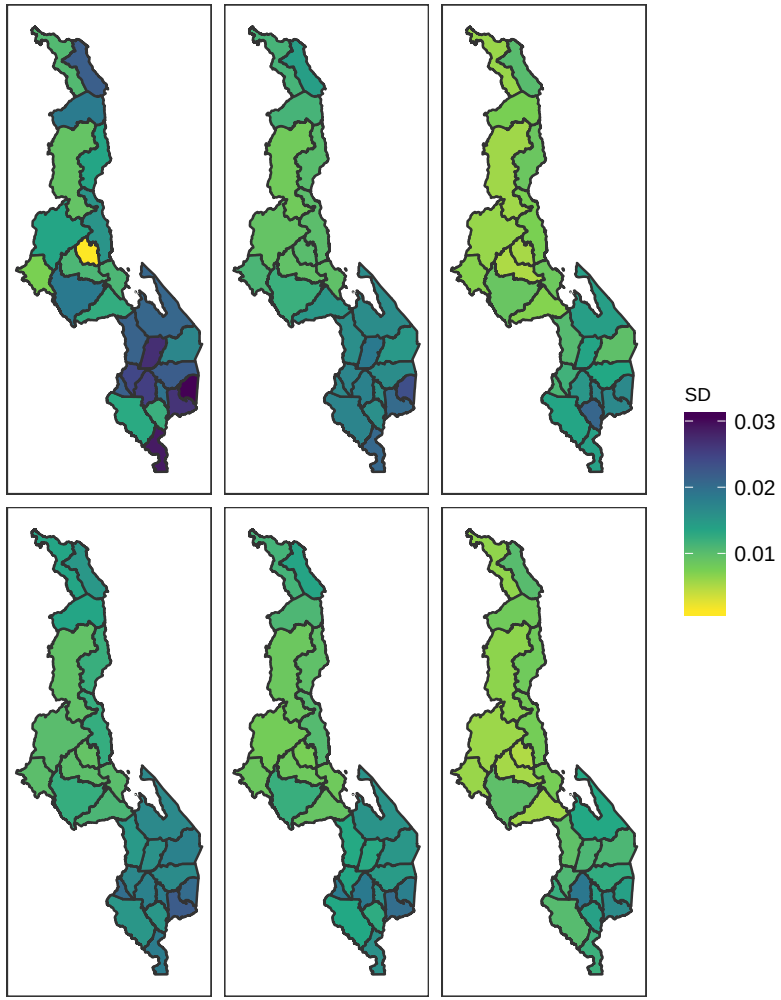


Figure 7: Uncertainty estimates (standard errors for direct estimates, posterior standard deviations for the remainder) of HIV prevalence. Top: area-level models: direct estimates; smoothed direct estimates with no ANC covariate; smooth direct estimates with ANC covariate. Bottom: unit-level models: no urban/rural adjustment, no ANC covariate; urban/rural adjusted, no ANC covariate; urban/rural covariate and ANC covariate.

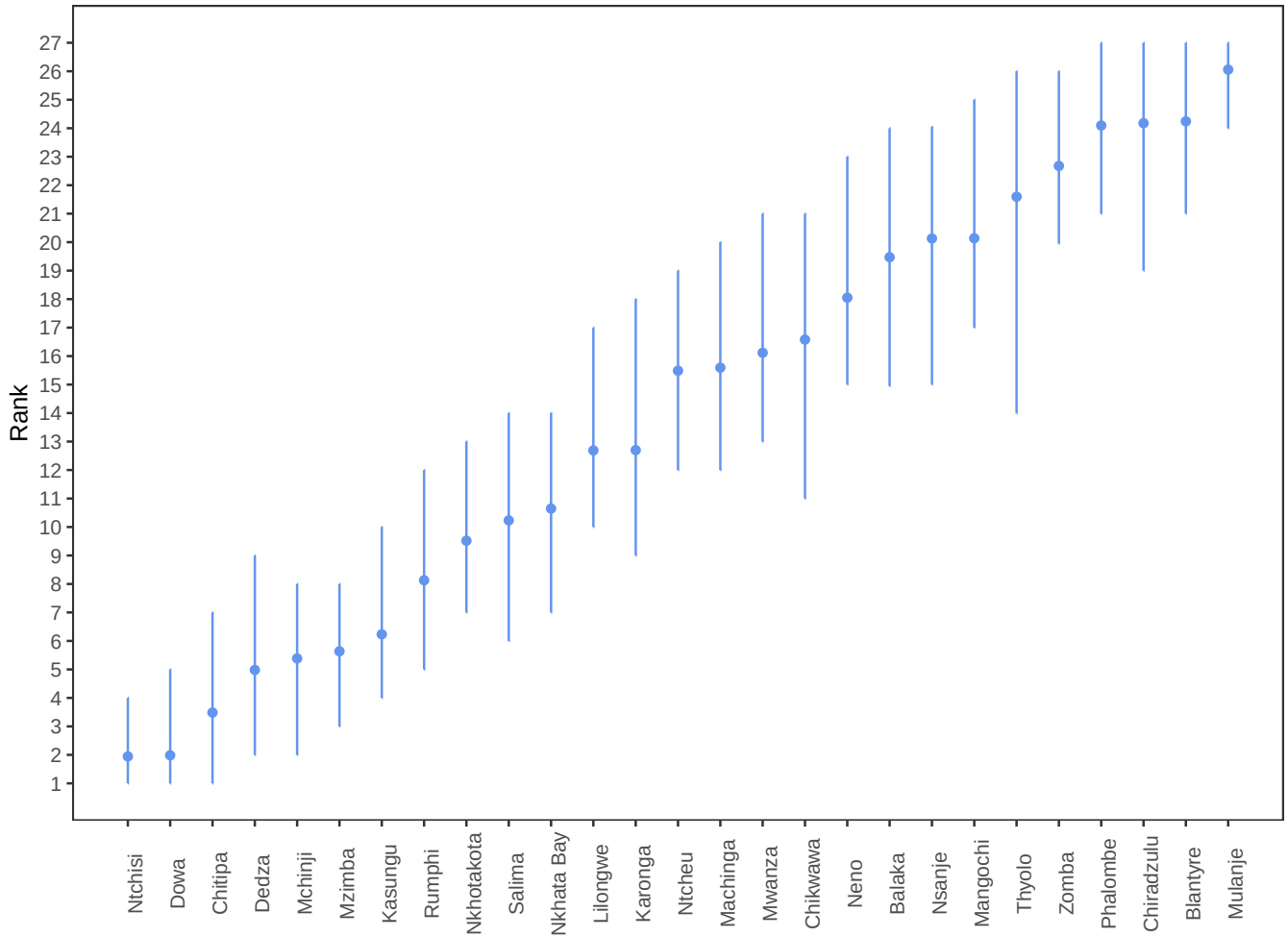


Figure 8: Distributions on the rankings for the smoothed direct estimates with the ANC covariate. The lines represent 90% intervals based on samples from the posterior, with rank = 1 on the y-axis corresponding to the lowest HIV prevalence and rank = 27 corresponding to the highest HIV prevalence.

Model Assessment

- One of the hardest parts of model-based approaches to SAE is assessment of model assumptions.
- A cross-validation strategy is to systematically remove one area at a time, and then obtain a prediction of the missing area's (logit of the) direct prevalence estimate, based on the remaining areas.
- The asymptotic distribution of this direct estimate is

$$\text{logit}(\widehat{m}_i^{\text{HT}}) \sim \text{N}(\text{logit}(m_i), V_i).$$

- We simulate samples from the approximation to the posterior of $\text{logit}(p_i)$ that is provided by INLA, and add iid $\text{N}(0, V_i)$ errors to each sample.
- The result is the predictive distribution of what the model thinks the direct estimate will be in the area for which the data were removed.
- We then plot representations of these 27 predictive distributions, and compare with the observed points $\text{logit}(\widehat{m}_i^{\text{HT}})$.

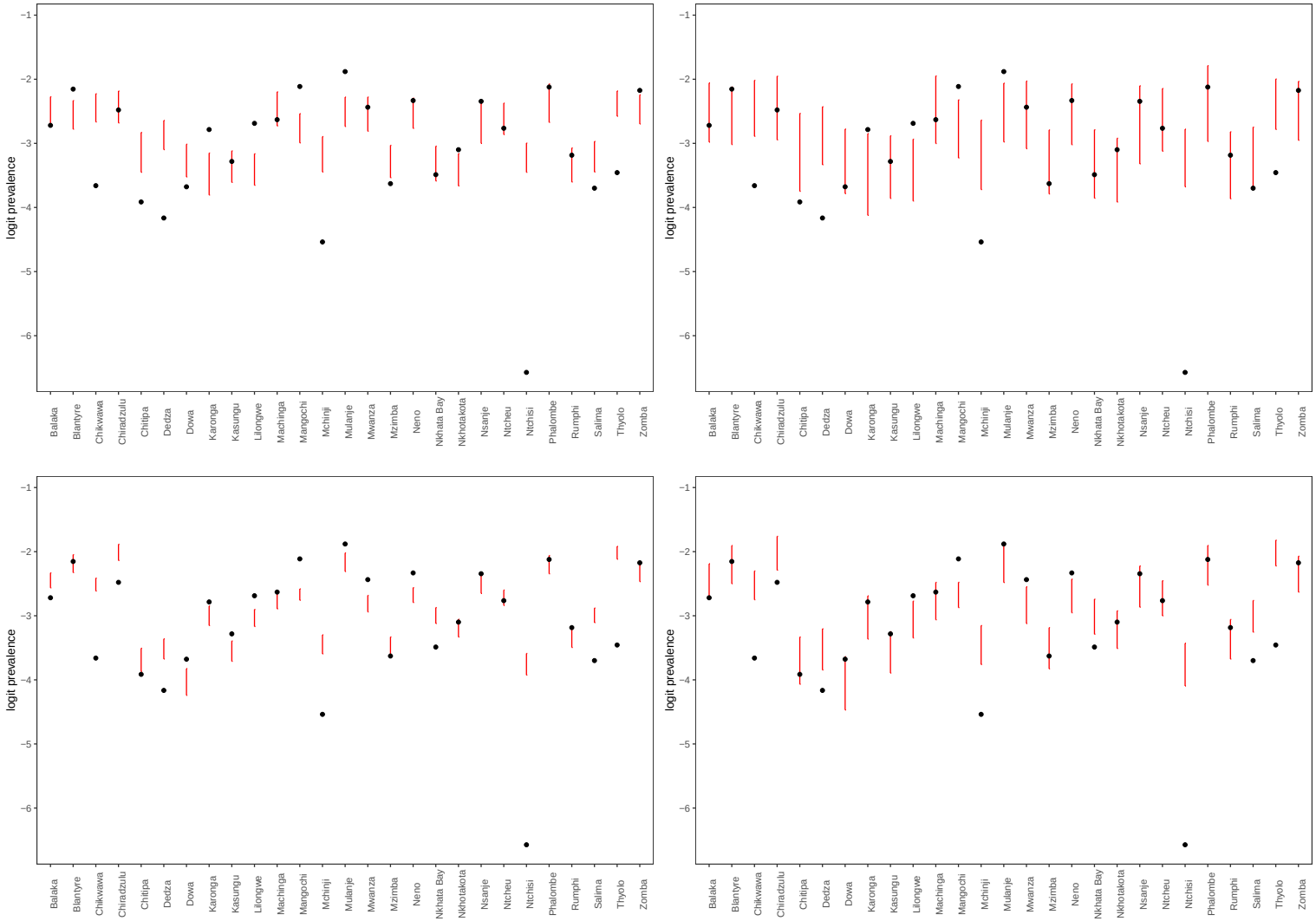


Figure 9: Leave-one cross-validation predictions for the smoothed direct models. Black dots are the direct estimates. Left column: 50% predictive intervals. Right: 80% predictive intervals. Top row: No ANC covariate. Bottom row: ANC covariate.

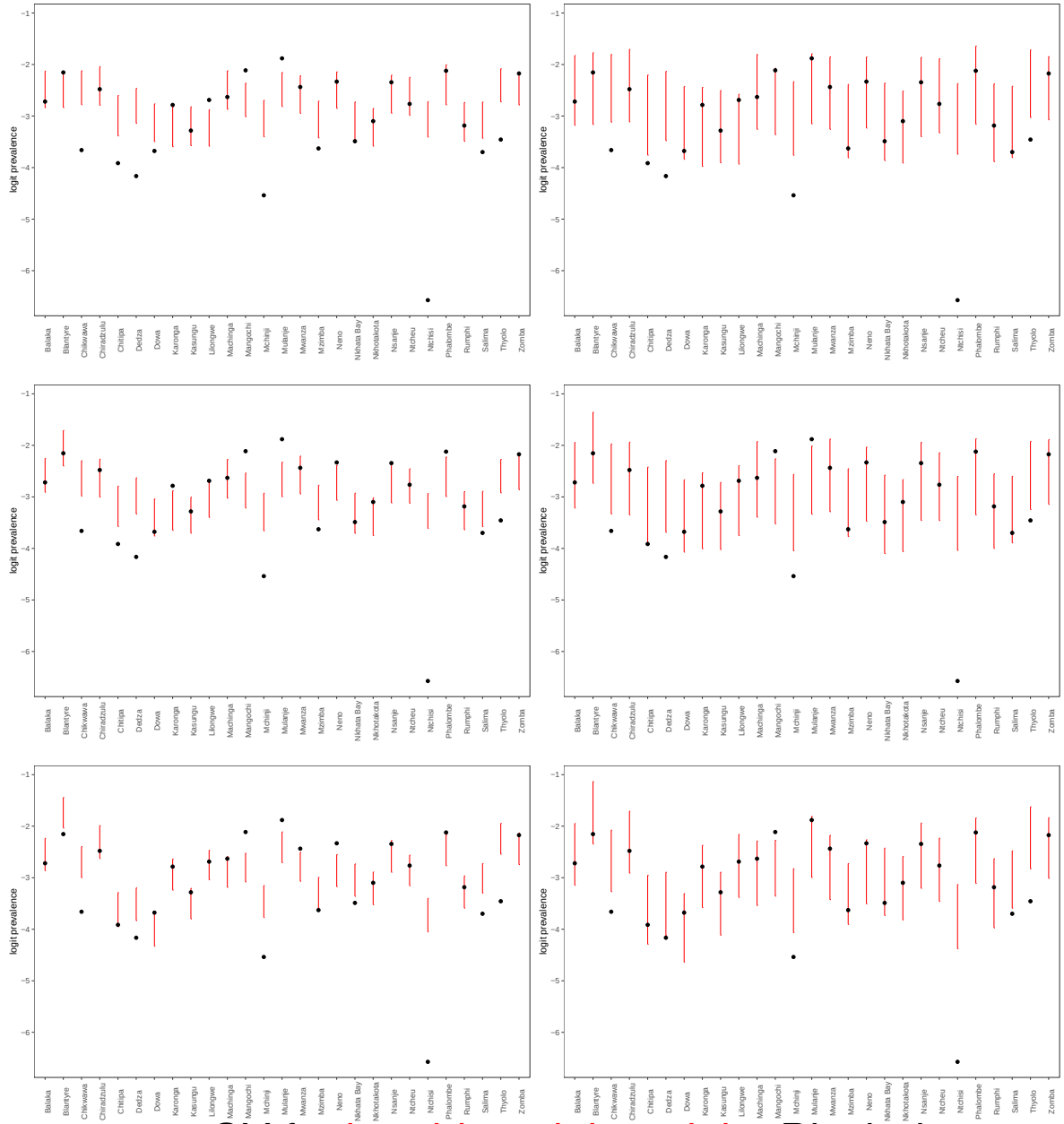


Figure 10: Leave-one CV for betabinomial models. Black dots are direct estimates. Left: 50% predictive intervals. Right: 80% predictive intervals. Top row: No urban/rural, no ANC covariate. Middle row: Urban/rural, no ANC covariate. Bottom row: Urban/rural, ANC covariate.

SPDE Model

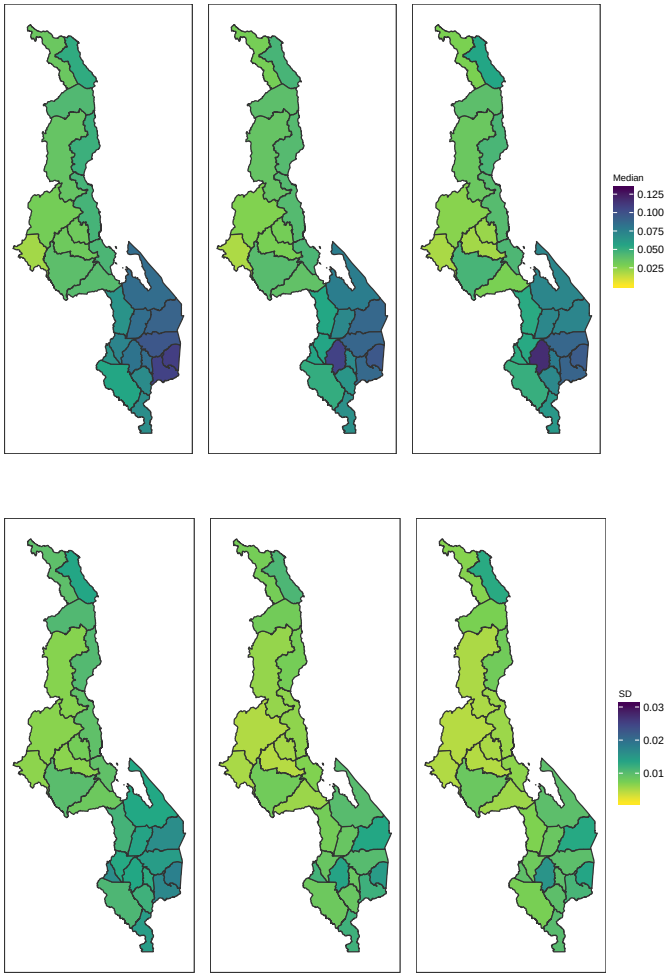


Figure 11: HIV prevalence summaries at the admin-2 level, using SPDE models. Left column is no urban/rural adjustment, no covariate. Middle column: urban/rural, no covariate. Right column: urban/rural, ANC covariate. Top row: posterior medians. Bottom row: posterior standard deviations.

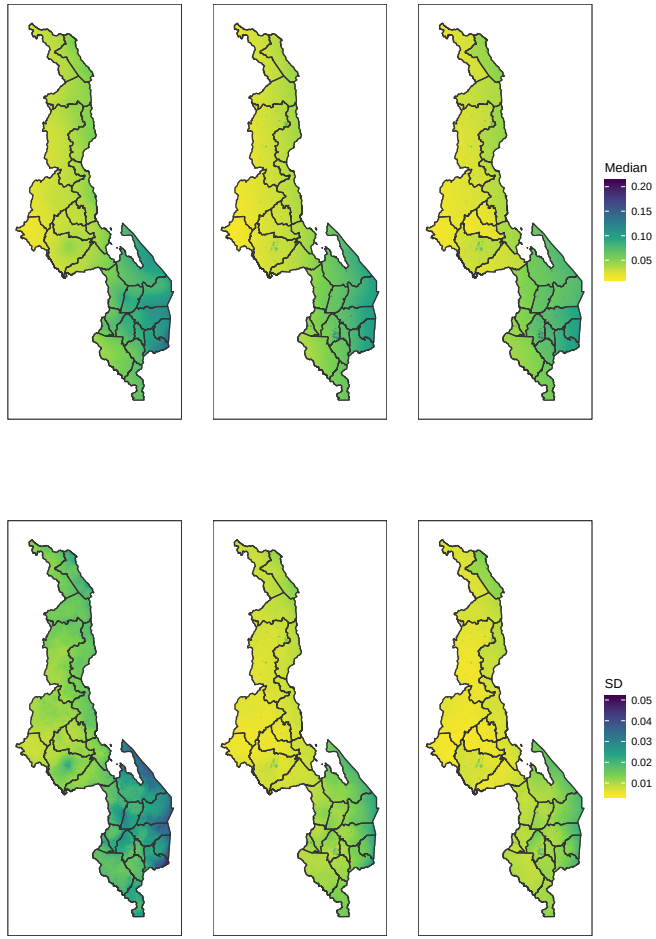


Figure 12: HIV prevalence summaries at the pixel level, using SPDE models. Left column is no urban/rural adjustment, no covariate. Middle column: urban/rural, no covariate. Right column: urban/rural, ANC covariate. Top row: posterior medians. Bottom row: posterior standard deviations.

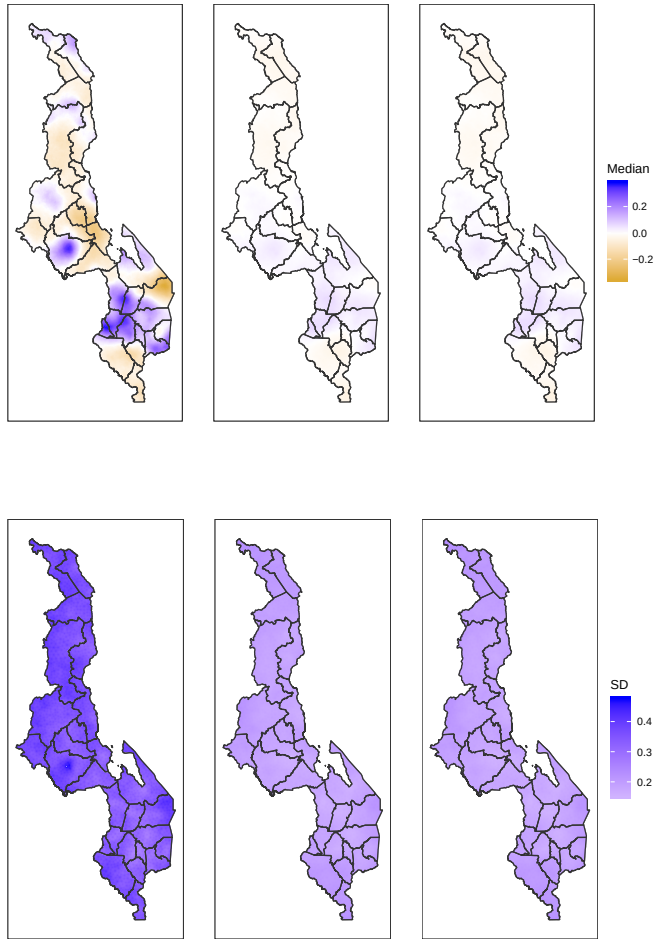


Figure 13: Spatial field summaries estimates at the pixel level, using SPDE models. Left column is no urban/rural adjustment, no covariate. Middle column: urban/rural, no covariate. Right column: urban/rural, ANC covariate. Top row: posterior medians. Bottom row: posterior standard deviations.

Takeaways

- Area-level modeling is more straightforward, if the data are sufficiently abundant.
- Unit-level modeling allow finer-scale modeling, but more sophisticated, and hence trickier; also more computationally expensive.
- If pixel maps are displayed, they should be accompanied by a map of uncertainty. Different methods for showing uncertainty are described in Dong and Wakefield (2021).
- Discrete spatial models always have an ad hoc neighborhood specification, which is unfortunate.
- Continuous spatial model are far more appealing in this respect, and also allow data that are aggregated to different levels to be combined
- The non-GP models can be fit in the `SUMMER` package (Li *et al.*, 2020).

References

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

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

Fay, R. and Herriot, R. (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.

Hájek, J. (1971). Discussion of, "An essay on the logical foundations of survey sampling, part I", by D. Basu. In V. Godambe and D. Sprott, editors, *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto.

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

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

Li, Z., Martin, B., Dong, T., Fuglstad, G.-A., Godwin, J., Paige, J., Riebler, A., Clark, S., and Wakefield, J. (2020). Space-time smoothing of demographic and health indicators using the `R` package `SUMMER`. *https://arxiv.org/abs/2007.05117*.

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

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

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

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

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

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

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, **1**, 16–29.

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

Wakefield, J. (2020). Prevalence mapping. In *Wiley StatsRef: Statistics Reference Online*, pages 1–7. Wiley.

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