Bayesian SAE with an Emphasis on Low- and Middle-Income Countries: Lecture 4: Unit-Level SAE Models

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Motivation

- We have seen that direct (weighted) estimation is reliable and relatively assumption-free when there are abundant data in each area.
- If direct estimates with associated variances (perhaps after variance smoothing) but they are imprecise, then the area-level (Fay-Herriot) model provides a good alternative – it is quiet robust to model mis-specification, and as a linear mixed effects model (LMEM), is well understood.
- When the data are sparse in an area, the weighted estimates may be on the boundary (for a prevalence) and variance estimates may be undefined or difficult to calculate.
- In these circumstances, unit-level models can be fitted, but great care is required since obtaining valid inference is far more tricky.

Overview of the Unit-Level Model

- The area-level model adjusts for the design, both in the mean and the variance of the sampling model, and produces design-consistent inference.
- The standard unit-level does not use the weights and so not such guarantees exist.
- Instead, one must account for the design (for example stratification and clustering in DHS and MICS data) via terms in the model.
- We will interchange the standard SAE terminology of unit-level model with cluster-level model – in low- and middle-income countries (LMICs) all units (households) in the same cluster have the same weight and design variables.
- We focus on discrete spatial models rathe than the continuous spatial models that are favored by model-based geostatistics ((MBG) approaches.

Unit-Level Models

The Data and the Target

The data consist of:

- the response for unit k in area i, y_{ik}, with $\boldsymbol{x}_{ik}$ being a vector of unit-level covariates and with $k = 1, \ldots, n_i$, and $i = 1, \ldots, m$.

The target in SAE is the area-level finite population mean:

$$\theta_i = \overline{y}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} y_{ik},$$

where N_i is the size of the population.

Basic Unit-Level Model (Battese *et al.*, 1988)

The unit-level nested error linear regression model is specified as:

$$
\begin{aligned}
Y_{ik} &= \mu_{ik} + \epsilon_{ik}, && && (1)\\
\epsilon_{ik} \mid \sigma^2_\epsilon &\sim_{iid} \mathrm{N}(0, \sigma^2_\epsilon), && k = 1, \ldots, n_i, && (2)\\
\mu_{ik} &= \alpha + \boldsymbol{x}^\mathsf{T}_{ik}\boldsymbol{\beta} + u_i, && && (3)\\
u_i \mid \sigma^2_u &\sim_{iid} \mathrm{N}(0, \sigma^2_u), && i = 1, \ldots, m, && (4)
\end{aligned}
$$

where

- α is the intercept and $\boldsymbol{\beta}$ are regression coefficients associated with the covariates, $\boldsymbol{x}_{ik}$.
- u_i are area-specific random effects.
- ϵ_{ik} are unit-specific errors terms, with u_i and ϵ_{ik} being independent.

We have $\mathrm{E}[Y_{ik} \mid \mu_{ik}] = \mu_{ik}$, where the expectation is over a hypothetical infinite population of units in area i and we have moved from a design-based perspective to a model-based perspective.

The Basic Unit-Level Model

- Note the weights are nowhere to be seen.
- A key assumption here is that model (1)–(4) holds for both the sampled units and the non-sampled units – i.e., for the complete population.
- The requirement is that, conditional on covariates, the sampling of a unit does not depend on the response of that unit.
- Most household surveys in LMICs are stratified by Admin1 crossed with urban/rural, and often urban clusters are oversampled.
- The prevalence of many indicators will be associated with urban/rural and if so, modeling as a function of urban/rural is important.
- Stratification variables can appear in $\boldsymbol{x}_{ik}$ though stratification for Admin1 may be accounted for using area-level random effects.

Aggregation for Basic Unit-Level Model

An additional significant complication when using unit-level models, is how to make inference for areas – this requires aggregation.

The definition of the area-level finite population mean is,

$$\theta_i = \frac{1}{N_i} \sum_{k=1}^{N_i} y_{ik}$$

where N_i is the population size in area i.

In the following we assume the sampled fraction of the units in each area is negligible.

Aggregation for the Basic Unit-Level Model

The model is,

$$\begin{aligned} \mathrm{E}[Y_{ik} \mid \mu_{ik}] &= \mu_{ik} \\ &= \alpha + \boldsymbol{x}_{ik}^{\mathsf{T}}\beta + u_i. \end{aligned}$$

Under this model,

$$\begin{aligned} \theta_i &= \frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik} \\ &\approx \frac{1}{N_i}\sum_{k=1}^{N_i} \mu_{ik} \qquad \text{since} \qquad \frac{1}{N_i}\sum_{k=1}^{N_i} \epsilon_{ik} \approx 0 \\ &= \frac{1}{N_i}\sum_{k=1}^{N_i} \left(\alpha + \overline{\boldsymbol{x}}_{ik}^{\mathsf{T}}\beta + u_i\right) \\ &= \alpha + \overline{\boldsymbol{x}}_i^{\mathsf{T}}\beta + u_i \end{aligned}$$

where $\overline{\boldsymbol{x}}_i$ is the mean, across the population which is practically important – inference depends on the mean (e.g., from a census), and $\boldsymbol{x}_{ik}$ for each unit in the population is not needed.

Areas with No Data

If there are no data in area $i^\star$ (say) we can still make predictions, using the vector of area covariate means, $\overline{\boldsymbol{x}}_{i^\star}$ to give samples from the posterior,

$$\theta_{i^\star}^{(s)} = \alpha^{(s)} + \overline{\boldsymbol{x}}_{i^\star}^{\mathsf{T}} \beta^{(s)} + u_{i^\star}^{(s)}, \qquad s = 1, \ldots, S$$

This of course depends on the original model holding for all the areas.

Under the Bayesian approach we follow, the random effect will be estimated via draws from,

$$u_{i^\star}^{(s)} \mid \sigma_u^{2(s)} \sim \mathsf{N}(0, \sigma_u^{2(s)}),$$

where

$$\sigma_u^{2(s)} \sim p(\sigma_u^2 \mid \boldsymbol{y}), \qquad s = 1, \ldots, S.$$

The posterior mean of $\theta_{i^\star}$ will not be contributed to by $u_{i^\star}$ but the spread of the posterior will be.

Inference for the Unit-Level Model

- Rao and Molina (2015) contains extensive details on estimation for the unit-level model.
- A Bayesian model is completed by adding a prior on $\alpha, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_u^2$ – we use penalized complexity (PC) priors (Simpson *et al.*, 2017).
- In the `survey` and `SUMMER` `R` packages, we use INLA to carry out the required computations for posterior summarization.
- With samples from the posterior $\{\alpha^{(s)}, \boldsymbol{\beta}^{(s)}, u_i^{(s)}, s = 1, \ldots, S\}$ we can produce samples from the posterior for the area level mean via:
$$\theta_i^{(s)} = \alpha^{(s)} + \overline{\boldsymbol{x}}_i^\intercal \boldsymbol{\beta}^{(s)} + u_i^{(s)}, \qquad s = 1, \ldots, S$$
- We can then summarize the posterior in the usual way, for example, via posterior quantiles.

The Spatial Unit-Level Model

The spatial version of the model is,

$$\begin{aligned} Y_{ik} &= \alpha + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta} + u_i + \epsilon_{ik} \\ \epsilon_{ik} \mid \sigma_\epsilon^2 &\sim \mathrm{N}(0, \sigma_\epsilon^2) \\ \boldsymbol{u} = [u_1, \ldots, u_m] \mid \sigma_u^2, \phi &\sim \mathrm{BYM2}(\sigma_u^2, \phi) \end{aligned}$$

If there are no data in area $i^\star$ (say) we can still make predictions, if we assume the model holds for all areas, but the random effect will be estimated via draws from $u_{i^\star}^{(s)} \mid \sigma_u^{2(s)}, \phi^{(s)} \sim \mathrm{BYM2}(\sigma_u^{2(s)}, \phi^{(s)})$, where

$$\sigma_u^{2(s)}, \phi^{(s)} \sim p(\sigma_u^2, \phi \mid \boldsymbol{y}), \qquad s = 1, \ldots, S.$$

Spatial random effects are especially beneficial when compared to the IID model, since they can exploit local dependencies.

But, results should be viewed skeptically if there are many areas with no data.

Binary Data Unit-Level Models

Bernoulli Unit-Level Model

- Many indicators in LMICs are binary and so linear models are appealing because probabilities are constrained to $[0, 1]$ and the variance is a function of the mean (rather than be constant, as in the linear model).
- A natural sampling model to consider is a binomial – we begin with this model and then extend to account for cluster sampling.
- Let Y_{ik} be a binary indicator for the event of interest, for $k = 1, \ldots, n_i$ units in area i, $i = 1, \ldots, m$.
- Also, suppose we have a vector of unit-level covariates $\boldsymbol{x}_{ik}$.
- We will construct a risk model, with r_{ik} being the risk of the event – imagine a hypothetical superpopulation of individuals and the proportion who manifest the event is r_{ik}.

Bernoulli Unit-Level Model

For binary data, we may assume the model:

$$\begin{aligned} Y_{ik} \mid r_{ik} &\sim \text{Bernoulli}(r_{ik}), \\ \log\left(\frac{r_{ik}}{1-r_{ik}}\right) &= \alpha + \boldsymbol{x}_{ik}^{\mathsf{T}}\beta + u_i, \qquad k = 1, \ldots, n_i, \end{aligned}$$

with,

- β being a vector of regression coefficients,
- u_i being an area-level random effect, with

$$u_i \mid \sigma_u^2 \sim_{iid} \text{N}(0, \sigma_u^2), \qquad i = 1, \ldots, m,$$

or

$$\boldsymbol{u} = [u_1, \ldots, u_m] \mid \sigma_u^2, \phi \sim \text{BYM2}(\sigma_u^2, \phi).$$

Aggregation for the Bernoulli Model

Aggregated risk, under the model, and assuming the sampled fraction of units in the area is negligible, is

$$
\begin{aligned}
r_i &= \frac{1}{N_i}\sum_{k=1}^{N_i} r_{ik} \\
&\underbrace{\approx}_{\substack{\text{Modeling} \\ \text{Approximation}}} \frac{1}{N_i}\sum_{k=1}^{N_i} \text{expit}(\alpha + \boldsymbol{x}_{ik}^{\mathsf{T}}\beta + u_i).
\end{aligned}
$$

Notice that, in contrast to the linear model case, we require the covariates on each member of the population.

In practice, we can aggregate over a grid:

$$
r_i \approx \sum_{g=1}^{G_i} F_{ig} \times \text{expit}(\alpha + \boldsymbol{x}_{ig}^{\mathsf{T}}\beta + u_i), \tag{5}
$$

with F_{ig} the fraction of the population units that are located in grid cell g, for example, from WorldPop population estimates (Tatem, 2017).

Inference for the Bernoulli Model

- We have a generalized linear mixed model (GLMM) and so inference is relatively straightforward with both frequentist and Bayesian approaches being possible.
- We continue with a Bayesian standpoint, using INLA (Rue *et al.*, 2017) and penalized complexity (PC) priors on variance components.
- Again no design weights in the model so we need to include terms in the model to acknowledge the design – for example, including an urban/rural indicator in $\boldsymbol{x}_{ik}$; we then need an urban/rural surface to perform aggregation via (7).
- We need to consider the clustering aspect also – this will be done shortly via considering overdispersion models.

Binary Data Cluster-Level Models

Cluster Level Models

- In LMICs, household surveys sample clusters (enumeration areas) and then household within clusters.
- Let the clusters be labeled by c, with $c = 1, \ldots, C_i$ clusters in area i and all the units sampled in cluster c are regarded as having the same location and the same covariates.
- The geographical location of cluster c is denoted $\boldsymbol{s}_{ic}$; this is a 2-dimensional point but will actually represent the complete sampling cluster (i.e., the enumeration area).
- We focus on binary outcomes – given the shared location, we sum up Bernoulli responses in cluster c of area i, and let

$$Y_{ic} = Y(\boldsymbol{s}_{ic}) = \sum_{k \in c} Y_{ik}$$

be the sum of the events in the cluster, out of a total,

$$n_{ic} = n(\boldsymbol{s}_{ic}) = \sum_{k \in c} 1.$$

Notation

- We suppose there are C_i clusters sampled within area i, $i = 1, \ldots, m$.
- And $\boldsymbol{s}_{ic}$ is the geographical location of cluster c within area i, $c = 1, \ldots, C_i$, $i = 1, \ldots, m$.
- At location $\boldsymbol{s}_{ic}$, the number of deaths and the number of births in cluster c, area i are denoted $Y(\boldsymbol{s}_{ic})$ and $n(\boldsymbol{s}_{ic})$, respectively.
- The risk $r(\boldsymbol{s}_{ic})$ at location $\boldsymbol{s}_{ic}$ is the proportion of hypothetical events in an infinite population at that location with the characteristic of interest.

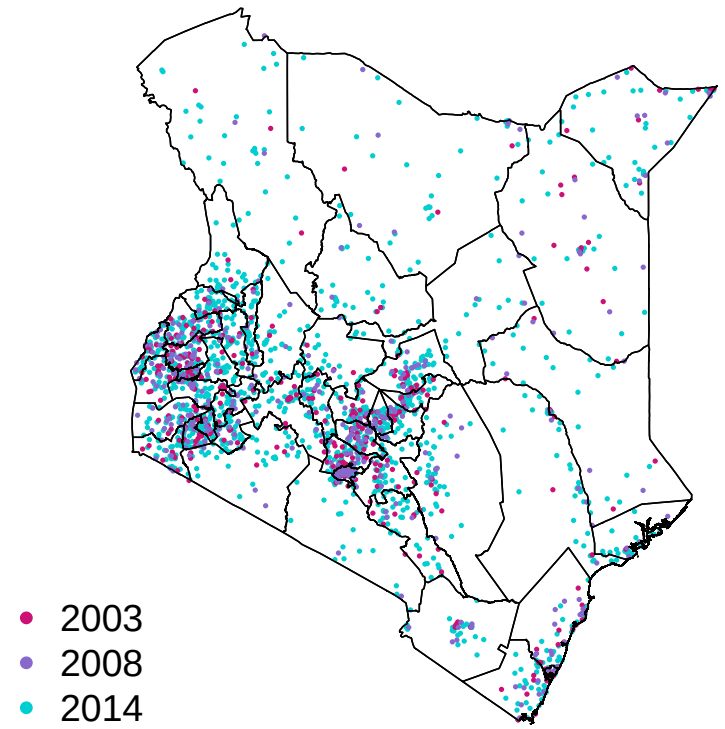


Figure 1: Cluster locations in three Kenya DHS, with county boundaries.

Excess-Binomial Variation (Overdispersion)

- Two responses in the same cluster are more likely to be similar than two responses in different clusters – this positive dependence induces excess-binomial variation.
- This is also known as overdispersion.
- The way we model this is by envisaging a two-stage sampling scheme:
  1. Generate a random variable δ from some distribution.
  2. Generate a random variable Y from a binomial conditioned on the δ sampled at stage 1.
- At stage 1. we describe two alternative models – generating a normal random variable, or generating a beta random variable.

Excess-Binomial Variation (Overdispersion)

- If we marginalize over δ we obtain an overdispersed binomial distribution – the key point is that the same δ is used for all units in the same cluster.
- This is a little confusing!
- We label the marginal risk as r_{ic} while the conditional risk is $r^{\star}_{ic}$.

LogitNormal-Binomial Cluster-Level Model

A popular cluster level model is,

$$\begin{aligned} Y_{ic} \mid r^{\star}_{ic} &\sim \text{Binomial}(\, n_{ic}, r^{\star}_{ic}\,) \\ \text{logit}(r^{\star}_{ic}) &= \alpha + \boldsymbol{x}_{ic}^{\mathsf{T}}\beta + u_i + \epsilon_{ic} \end{aligned}$$

where,

- $\boldsymbol{x}_{ic}$ is the vector of covariates.
- u_i are area-level random effects, for example, having BYM2 structure.
- The cluster level error $\epsilon_{ic} \mid \sigma^2_{\epsilon} \sim_{iid} \text{N}(0, \sigma^2_{\epsilon})$.
- Since ϵ_{ic} is shared by all units in the same cluster, positive dependence in their responses is induced.
- An example of a generalized linear mixed model (GLMM).

Cartoon for the GLMM

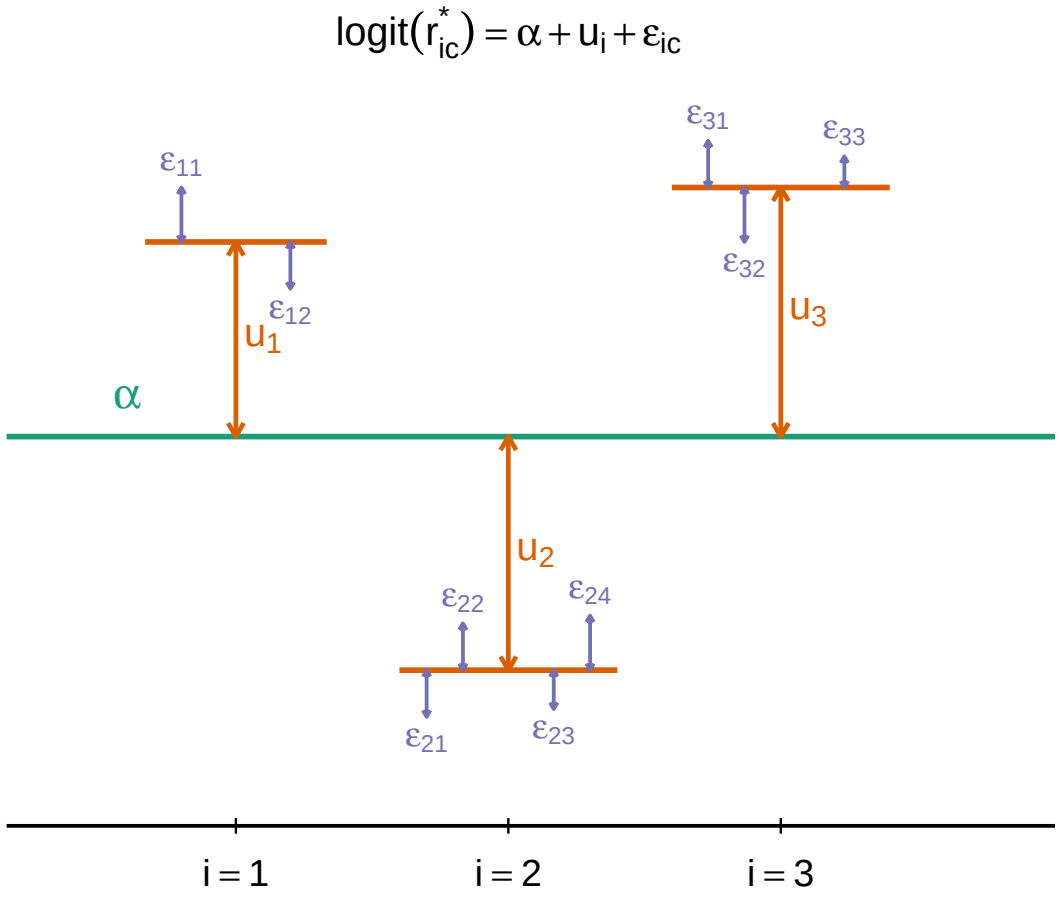


Figure 2: One-dimensional representation of the cluster-level model (no covariates, for simplicity).

Aggregation for the GLMM

- Suppose there are C_i^{POP} clusters in the sampling frame with N_{ic} units in cluster c, $c = 1, \ldots, C_i^{\text{POP}}$.
- Again, assume a negligible sampling fraction:

$$r_i = \frac{1}{N_i} \sum_{c=1}^{C_i^{\text{POP}}} \underbrace{N_{ic} \times r_{ic}}_{\text{Number of Events}}$$

$$\underbrace{\approx}_{\substack{\text{Modeling} \\ \text{Approximation}}} \sum_{c=1}^{C_i^{\text{POP}}} \frac{N_{ic}}{N_i} \underbrace{\int_{\epsilon} \overbrace{\text{expit}(\alpha + \boldsymbol{x}_{ic}^{\mathsf{T}}\beta + u_i + \epsilon)}^{\text{Conditional Risk } r_{ic}^{\star}} \times \text{N}(\epsilon \mid 0, \sigma_{\epsilon}^2)\, d\epsilon}_{\text{Marginal Risk } r_{ic}}.$$

Aggregation for the GLMM

- In practice, C_i^{POP} unknown and often N_{ic} and N_i are unknown too, so we can aggregate over a grid:

$$r_i \approx \sum_{g=1}^{G_i} F_{ig} \times \int_{\epsilon} \text{expit}(\alpha + \boldsymbol{x}_{ig}^{\mathsf{T}}\beta + u_i + \epsilon) \times \text{N}(\epsilon \mid 0, \sigma_\epsilon^2)\, d\epsilon, \quad (6)$$

with F_{ig} the fraction of the population units in area i that are located in grid cell g.

Beta-Binomial Cluster-Level Model

- In the GLMM model we have:

  1. Generate $\epsilon_{ic} \mid \sigma^2_\epsilon \sim_{iid} \mathrm{N}(0, \sigma^2_\epsilon)$.

  2. Generate $Y_{ic} \mid \epsilon_{ic} \sim \mathrm{Binomial}(n_{ic}, r^\star_{ic})$ with

$$r^\star_{ic} = r^\star_{ic}(\epsilon_{ic}) = \mathrm{expit}(\alpha + \boldsymbol{x}_{ic}\boldsymbol{\beta} + u_i + \epsilon_{ic}).$$

- Integrating over ϵ gives the marginal risk:

$$\begin{aligned} r_{ic} &= \int_\epsilon \mathrm{expit}(\alpha + \boldsymbol{x}_{ic}^\mathsf{T}\boldsymbol{\beta} + u_i + \epsilon) \times \mathrm{N}(\epsilon \mid 0, \sigma^2_\epsilon)\, d\epsilon \\ &\approx \mathrm{expit}\left(\frac{\alpha + \boldsymbol{x}_{ic}^\mathsf{T}\boldsymbol{\beta} + u_i}{\sqrt{1 + a^2\sigma^2_\epsilon}}\right) \end{aligned}$$

  where $a = 16\sqrt{3}/15\pi$.

- The marginal variance does not have a closed form but displays excess-binomial variation.

Beta-Binomial Cluster-Level Model

- We now describe a similar construction but rather than inducing dependence through a shared normal effect, we instead have a shared beta random effect – this has a simpler interpretation and gives computational advantages.

- Suppose the cluster level variability is described by the beta random effects distribution:

$$r_{ic}^{\star} \mid r_{ic}, d \sim \text{Beta}\left(r_{ic} \times d, (1 - r_{ic}) \times d\right),$$

  which gives,

$$\begin{aligned} \text{E}[r_{ic}^{\star}] &= r_{ic} \\ \text{var}(r_{ic}^{\star}) &= \frac{r_{ic}(1 - r_{ic})}{d + 1}. \end{aligned}$$

- The overdispersion is described by the scale parameter $d > 0$.

- The intraclass correlation coefficient is the correlation between two binary outcomes in the same cluster and corresponds to $1/(d + 1)$.

Beta-Binomial Cluster-Level Model

- The sampling model is,

$$\begin{aligned} Y_{ic} \mid r^{\star}_{ic} &\sim \text{Binomial}(n_{ic}, r^{\star}_{ic}) \\ r_{ic} \mid r_{ic}, d &\sim \text{Beta}\left(r^{\star}_{ic} \times d, (1 - r^{\star}_{ic}) \times d\right). \end{aligned}$$

- Integrating over r_{ic} gives the marginal sampling distribution:

$$\Pr(Y_{ic} \mid r_{ic}, d) = \int_{r_{ic}} \underbrace{\Pr(Y_{ic} \mid n_{ic}, r^{\star}_{ic})}_{\text{Binomial}(n_{ic}, r^{\star}_{ic})} \times \underbrace{p(r^{\star}_{ic} \mid r_{ic}, d)}_{\text{Beta}(r_{ic} \times d, (1 - r_{ic}) \times d)} dr^{\star}_{ic}.$$

- Turning the handle, gives

$$Y_{ic} \mid r_{ic}, d \sim \text{Beta-Binomial}(n_{ic}, r_{ic}, d),$$

with marginal mean and variance:

$$\begin{aligned} \text{E}[Y_{ic} \mid r_{ic}, d] &= n_{ic} r_{ic} \\ \text{var}(Y_{ic} \mid r_{ic}, d) &= n_{ic} r_{ic}(1 - r_{ic}) \times \frac{n_{ic} + d}{1 + d}, \end{aligned}$$

so we have overdispersion for $d > 0$.

Beta-Binomial Cluster-Level Model

We still need to specify a form for the mean, and a logistic model is natural:

$$r_{ic} = \text{expit}(\alpha + \boldsymbol{x}_{ic}^{\mathsf{T}}\beta + u_i),$$

where

- r_{ic} is the risk,
- $\boldsymbol{x}_{ic}$ may include the urban/rural strata within which cluster c lies,
- u_i can be IID or BYM2.

Aggregation

- For simplicity suppose we have a single covariate, the urban/rural classification of cluster c in area i so that $x_{ic} = 0/1$ for urban/rural.

- The modeled area marginal risk is

$$r_i = F_i \times \underbrace{\text{expit}(\alpha + u_i)}_{\text{Risk for Urban}} + (1 - F_i) \times \underbrace{\text{expit}(\alpha + \beta + u_i)}_{\text{Risk for Rural}},$$

  where

  - F_i is the proportion of the area that is urban, and
  - $1 - F_i$ is the proportion that is rural.

- The original sampling frame that contains the proportions of urban/rural in each, are typically unavailable.

- The F_i can be obtained by thresholding population density surfaces – Wu and Wakefield (2022) give a fancier method.

Aggregation for the Beta-Binomial Model

- The area risk is,

$$r_i = \sum_{c=1}^{C_i^{\text{POP}}} \frac{N_{ic}}{N_i} \times r_{ic}$$

$$\underbrace{\approx}_{\substack{\text{Modeling} \\ \text{Approximation}}} \sum_{c=1}^{C_i^{\text{POP}}} \frac{N_{ic}}{N_i} \times \underbrace{\text{expit}(\alpha + \boldsymbol{x}_{ic}^{\mathsf{T}}\beta + u_i)}_{\text{Marginal Risk } r_{ic}}.$$

- Note: unlike the normal random effect case, the marginal risk is available in closed form.

- In practice, we can aggregate over a grid:

$$r_i \approx \sum_{g=1}^{G_i} F_{ig} \times \text{expit}(\alpha + \boldsymbol{x}_{ig}^{\mathsf{T}}\beta + u_i), \tag{7}$$

with F_{ig} the fraction of the population units in area i that are located in grid cell g.

Cluster-Level Models

Advantages:

- If the data are sparse, this is the only possible method, since it can deal with situations in which the responses are either all 0s or all 1s.
- By modeling all the data in unison, uncertainty in each area is reduced (on average).
- Being a standard parametric model, there is a huge literature on inference and model assessment to lean on.

Cluster-Level Models

Disadvantages:

- Since the weights are not used, one must adjust for the design within the model specification, for example, by modeling the association between risk and urban/rural and adding overdispersion.
- Unit-level models do not provide design-consistent results in general – we are not guaranteed to converge to the true prevalence as the within-area sample size tends to the population size.
- This cluster effect is intended to account for within-cluster dependence in the outcomes. May also pick up within-area variation (interpretation tricky).

Cluster-Level Models

Disadvantages:

- The difficulties of aggregation, especially when the model is nonlinear (such as the nonlinear model we have used with binary data) should not be underestimated.
- The population information required in LMICs may have substantial error – we routinely perform aggregation of Admin2 level models from Admin2 to Admin1 to national to assess consistency with other estimates, e.g., weighted at Admin1 and area-level at Admin2.
- When the data are sparse, the estimates may be overshrunk since the data are not sufficiently informative to discriminate from the "flat" map case.
- The model produces shrinkage which can lead to interpretation being more tricky (since bias is introduced in each estimate).
- Is the unit-level uncertainty for area-level estimates too small? This is a very difficult question to answer.

A Case Study: HIV Prevalence in Malawi

Malawi DHS HIV Prevalence Example

- We consider SAE of HIV prevalence among females aged 15–29, in districts of Malawi, using data from the 2015–16 Malawi DHS – see Wakefield *et al.* (2020).
- 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.

Malawi DHS HIV Prevalence Example

- The sampling frame contained 12,558 clusters and our analyses use data from 827 sampled clusters.
- 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.

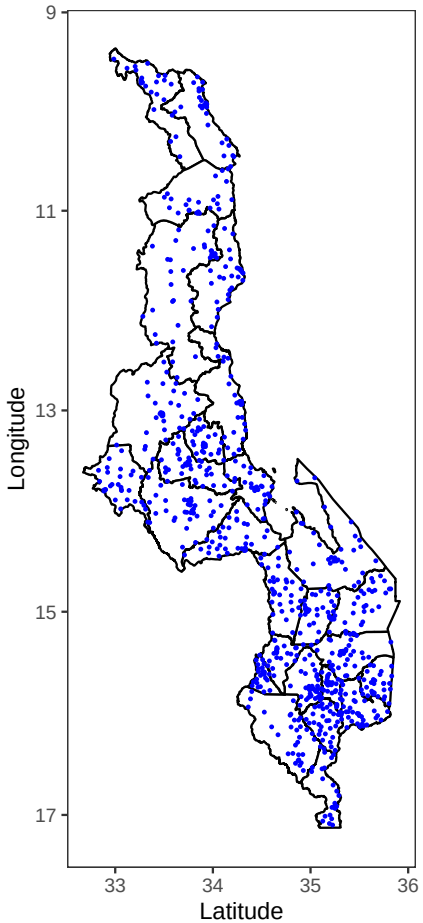


Figure 3: Cluster locations in 2015–16 Malawi DHS.

area-level Model

- Let $\widehat{\theta}_i^{\mathrm{w}} = \log\left(\frac{\widehat{p}_i^{\mathrm{w}}}{1-\widehat{p}_i^{\mathrm{w}}}\right)$ denote the logit of the weighted estimate.
- The area-level model begins with the sampling model

$$\widehat{\theta}_i^{\mathrm{w}} \mid \theta_i \quad \sim \quad \mathrm{N}(\theta_i, \widehat{V}_i)$$

where θ_i is the logit of the true proportion in area i, and $\widehat{V}_i$ is the design variance of the logit estimator.

- The linking model for θ_i is,

$$\theta_i = \alpha + x_i\beta + u_i,$$

where $[u_1, \ldots, u_m]$ are BYM2(σ_u^2, ϕ).

- We use the HIV prevalence from antenatal care (ANC) clinics in area i, as covariate x_i.

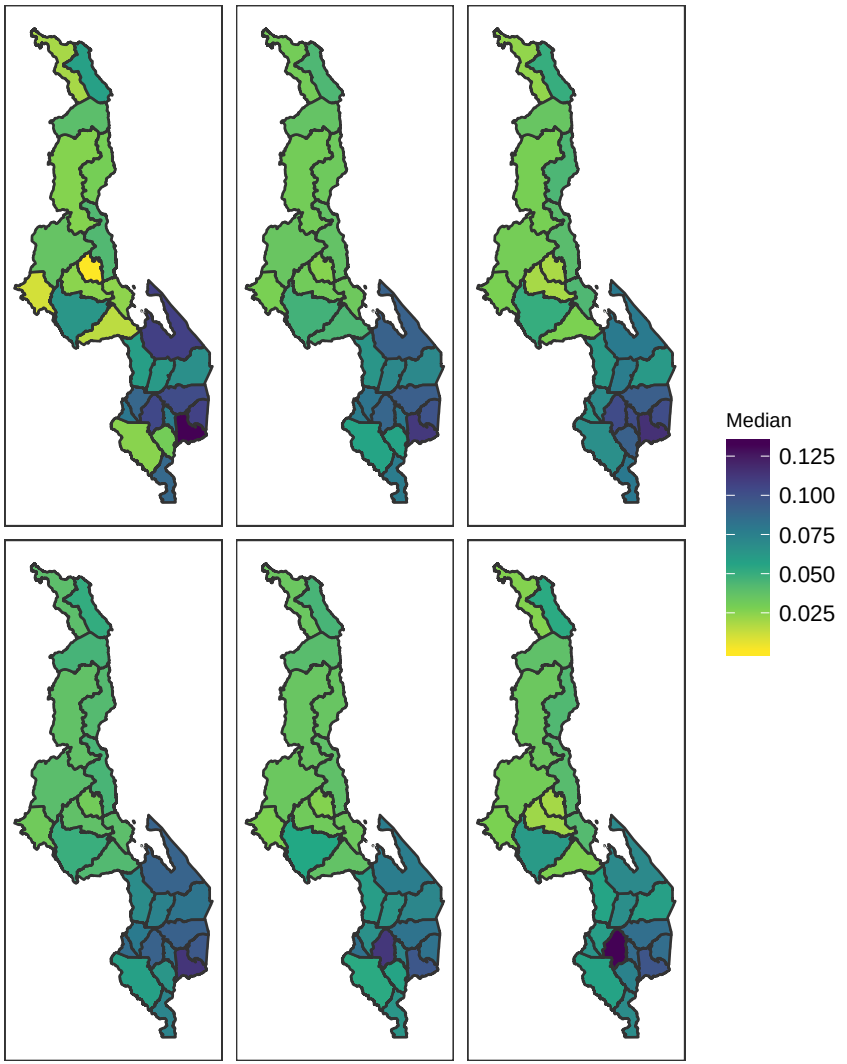
The Data Aggregated to Districts

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.

Comparison of Point Estimates

- 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; Fay-Herriot estimates, no covariate; Fay-Herriot estimates with 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.



Model		2.5%	Median	97.5%
No Covariates				
	BYM2 variance σ_u^2	0.07	0.19	0.48
	Proportion spatial ϕ	0.14	0.57	0.94
logit(ANC)				
	BYM2 variance σ_u^2	0.00	0.04	0.19
	Proportion spatial ϕ	0.01	0.17	0.85
	logit(ANC): odds ratio	1.59	2.72	4.03

TABLE 2: Posterior quantiles for the area-level BYM2 models. The BYM2 total variance is σ_u^2, the proportion spatial is ϕ, and the logit ANC (odds ratio) is $\exp(\beta)$.

The linear predictor is:

$$\theta_i = \alpha + x_i\beta + u_i,$$

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

Unit-Level Model

- We again 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

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

$$Y_{ic} \mid r_{ic}, d \sim \text{BetaBinomial}(n_{ic}, r_{ic}, d) \quad (8)$$

$$r_{ic} = \text{expit}\left(\alpha + x_i\beta + z_{ic}\gamma + u_i\right) \quad (9)$$

where

- x_i is the area-level (logit) ANC prevalence covariate, with associated odds ratio $\exp(\beta)$,
- z_{ic} is the urban/rural status (in the frame), with associated odds ratio $\exp(\gamma)$,
- d is the overdispersion parameter and
- we take $[u_1, \ldots, u_m]$ as BYM2(σ_u^2, ϕ).

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

$$r_i^\star \approx \sum_{g=1}^{G_i} F_{ig} \times r_{ig} \tag{10}$$

where

- the grid cell risk is,

$$r^\star = \text{expit}(\alpha + x_i\beta + z_{ig}\gamma + u_i),$$

  is the risk at cell g.

- F_{ig} is the population density at grid cell g, which is needed at all locations on the approximating mesh, $g = 1, \ldots, G_i$.

Model		2.5%	Median	97.5%
No Covariates				
	Overdispersion d	0.01	0.02	0.05
	BYM2 total σ_u^2	0.05	0.14	0.35
	Proportion spatial ϕ	0.15	0.62	0.96
U/R In				
	Overdispersion d	0.01	0.02	0.04
	BYM2 variance σ_u^2	0.05	0.13	0.33
	Proportion spatial ϕ	0.20	0.71	0.98
	Urban: OR $\exp(\gamma)$	1.73	2.29	3.00
U/R In, logit(ANC)				
	Overdispersion d	0.01	0.02	0.04
	BYM2 variance σ_u^2	0.00	0.02	0.12
	Proportion spatial ϕ	0.01	0.22	0.91
	Urban: OR $\exp(\gamma)$	1.70	2.24	2.94
	logit(ANC): OR $\exp(\beta)$	1.59	2.32	3.35

TABLE 3: Posterior quantiles for the unit-level betabinomial models. The overdispersion parameter is d, 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 $\exp(\beta)$.

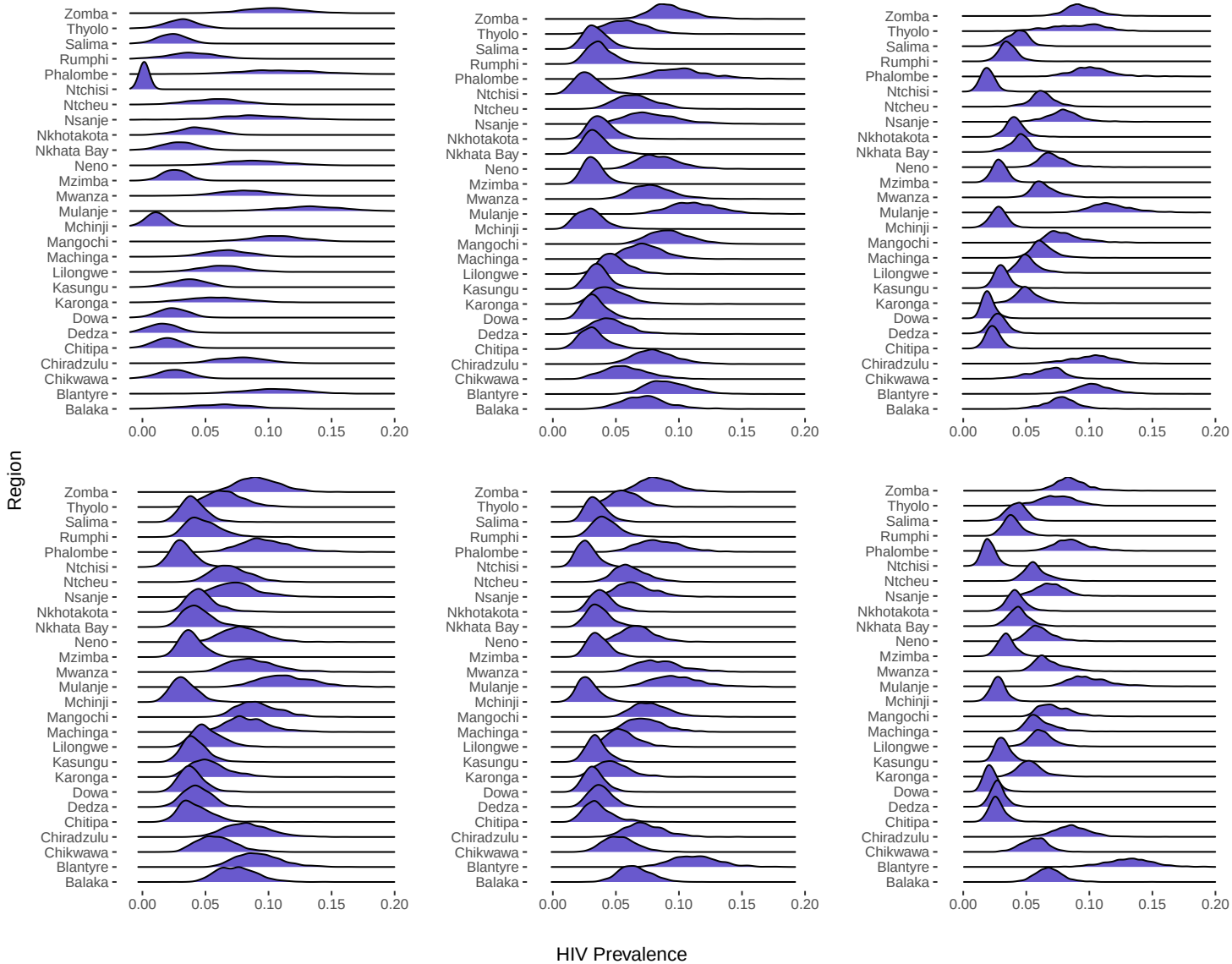


FIGURE 4: Posterior distributions for HIV prevalence/risk. Top row area-level models: direct; area-level, no covariates; area-level with ANC covariate. Bottom row unit-level (betabinomial) models: no urban/rural, no covariate; urban/rural only; urban/rural and ANC covariate.

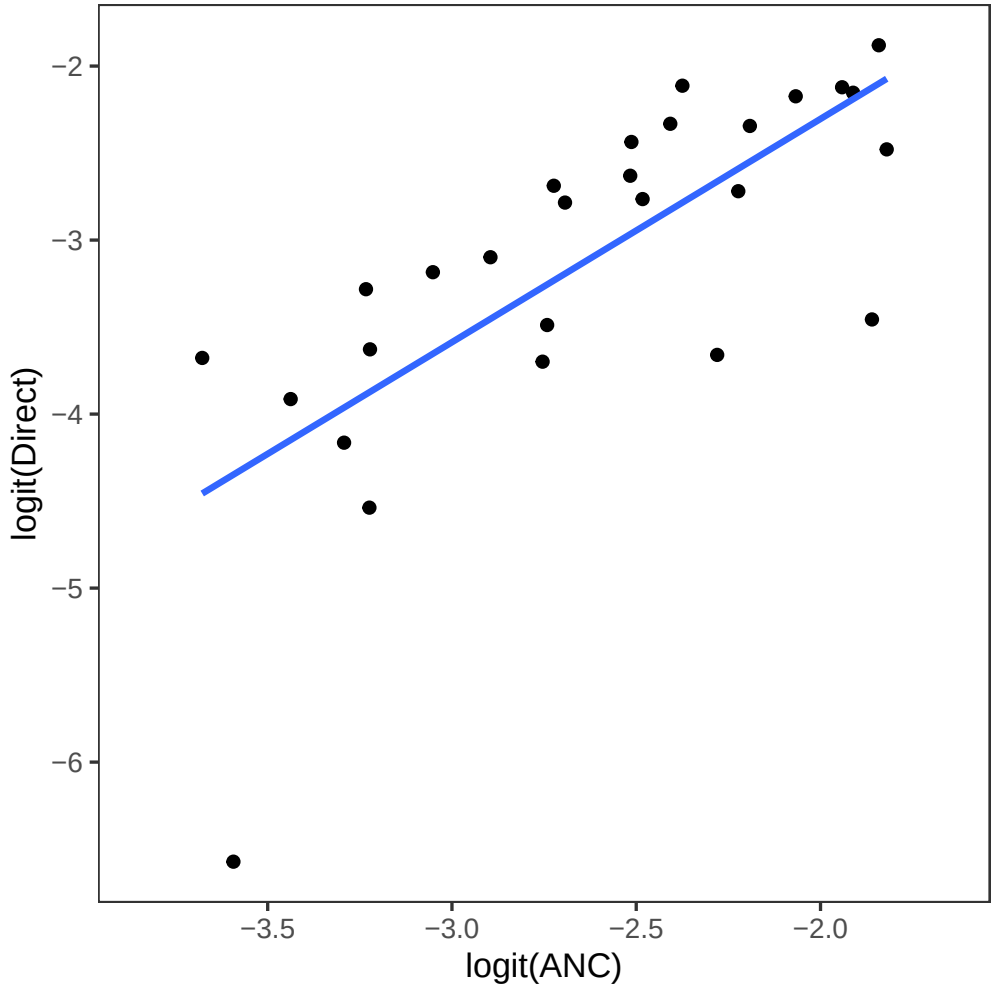


FIGURE 5: Logit of direct risk estimates versus logit of ANC risk estimates from unit-level model.

- Unsurprisingly, there is a strong association between the logit of ANC in this population of women and the logit of prevalence in pregnant women.

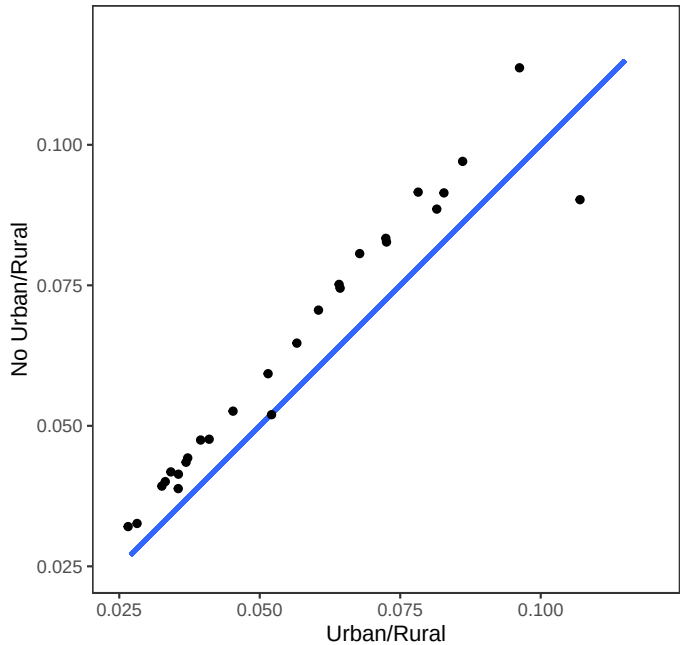
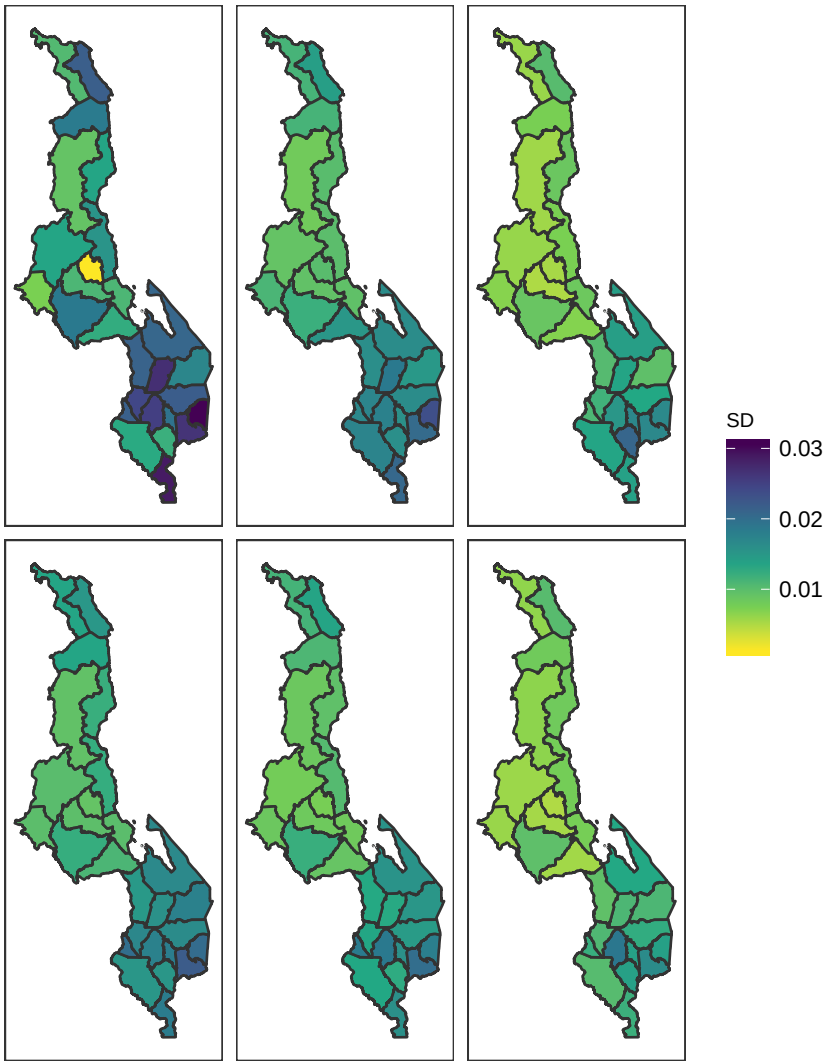


FIGURE 6: District prevalence estimates from two unit-level models. On the y-axis, the prevalence estimates are from a model with no urban/rural adjustment, while on the x-axis the model has an adjustment.

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

Comparison of Estimates

- Uncertainty estimates (standard errors for direct estimates, posterior standard deviations for the remainder) of HIV prevalence.
- Top: area-level models: direct estimates; area-level 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.



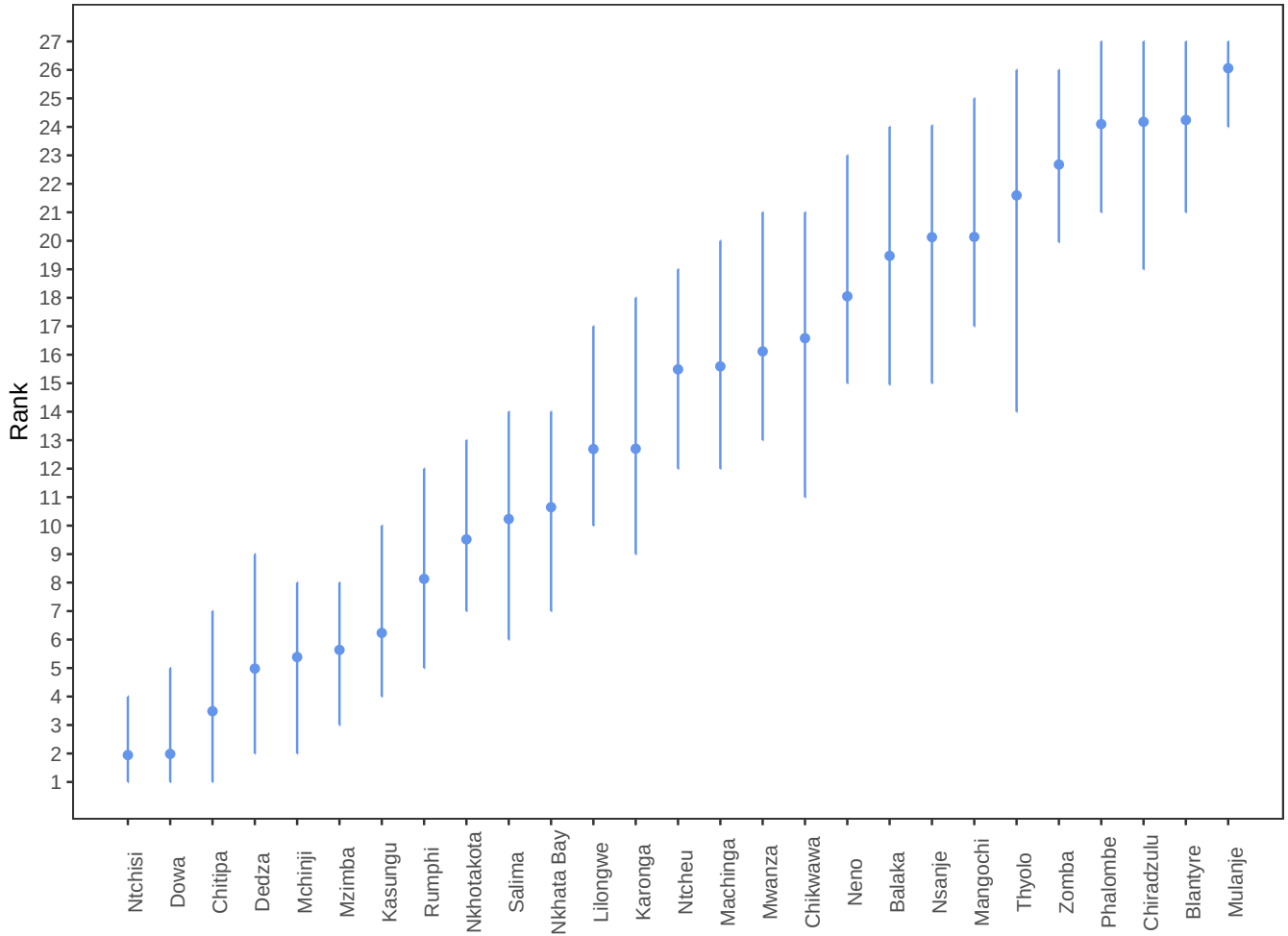


FIGURE 7: Distributions on the rankings for the area-level estimates with the ANC covariate. The lines represent 90% intervals based on samples from the posterior, with rank = 1 on the y-axis corresponding to the lowest HIV prevalence and rank = 27 corresponding to the highest HIV prevalence.

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

$$\theta_i^{\mathsf{w}} = \mathsf{logit}(\widehat{p}_i^{\mathsf{w}}),$$

  based on the remaining areas.
- The asymptotic distribution of this direct estimate is

$$\widehat{\theta}_i^{\mathsf{w}} \mid \theta_i \sim \mathsf{N}(\theta_i, V_i).$$

Model Assessment

- We simulate samples from the approximation to the posterior of θ_i that is provided by INLA, and add iid 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 $\widehat{\theta}_i^{\mathsf{W}}$.

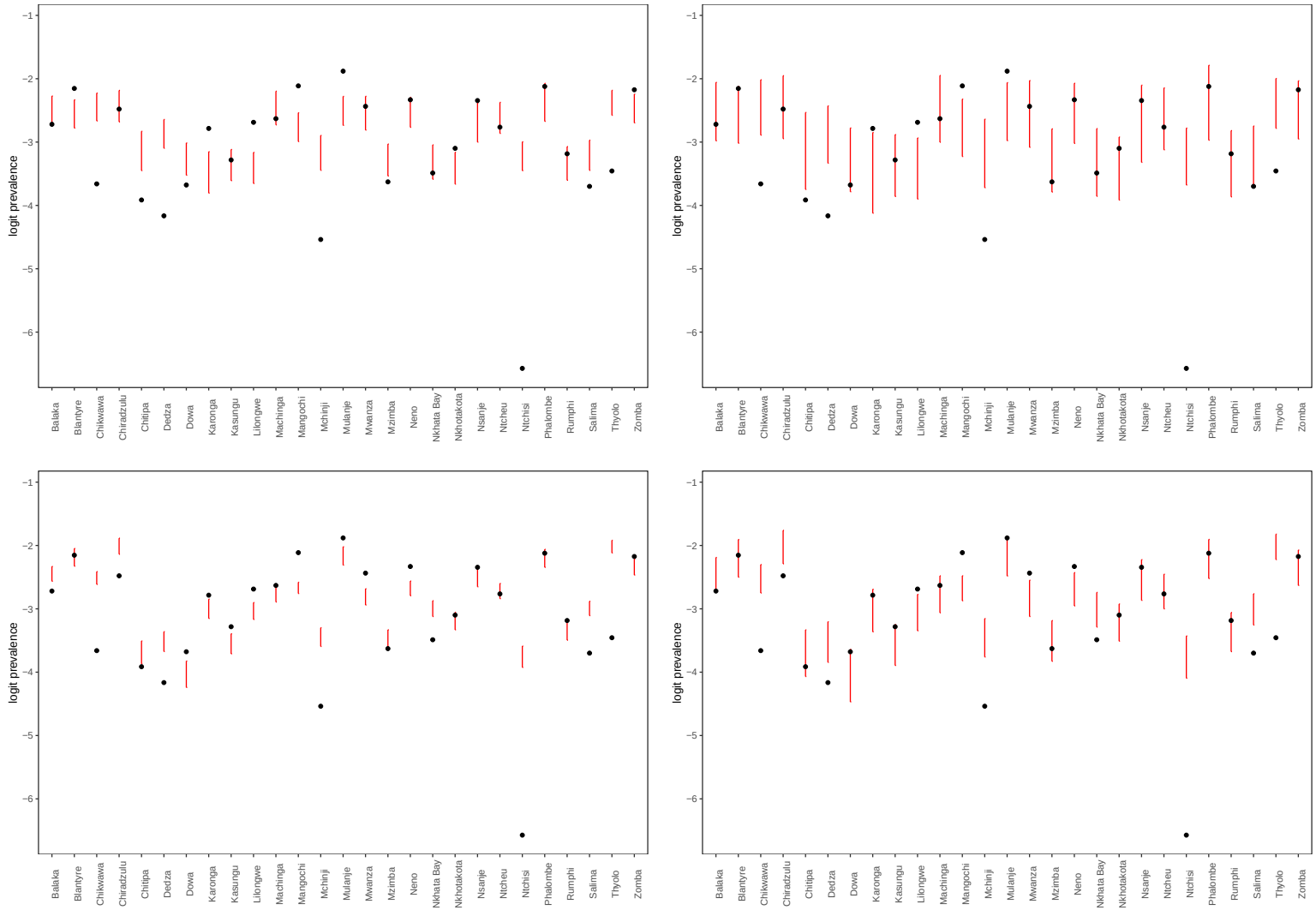


FIGURE 8: Leave-one cross-validation predictions for the area-level models. Black dots are the direct estimates. Left column: 50% predictive intervals. Right: 80% predictive intervals. Top row: No ANC covariate. Bottom row: ANC covariate.

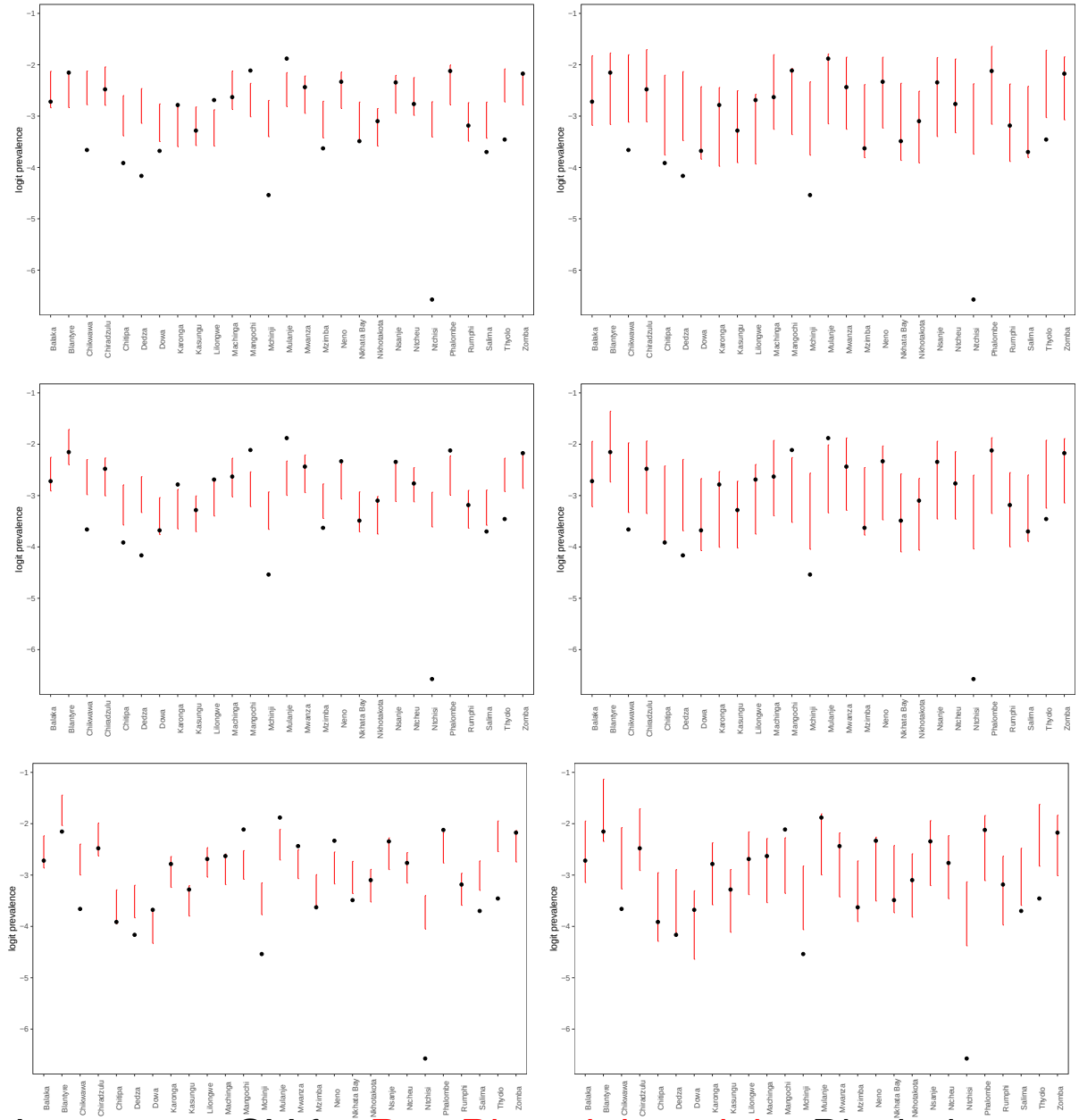
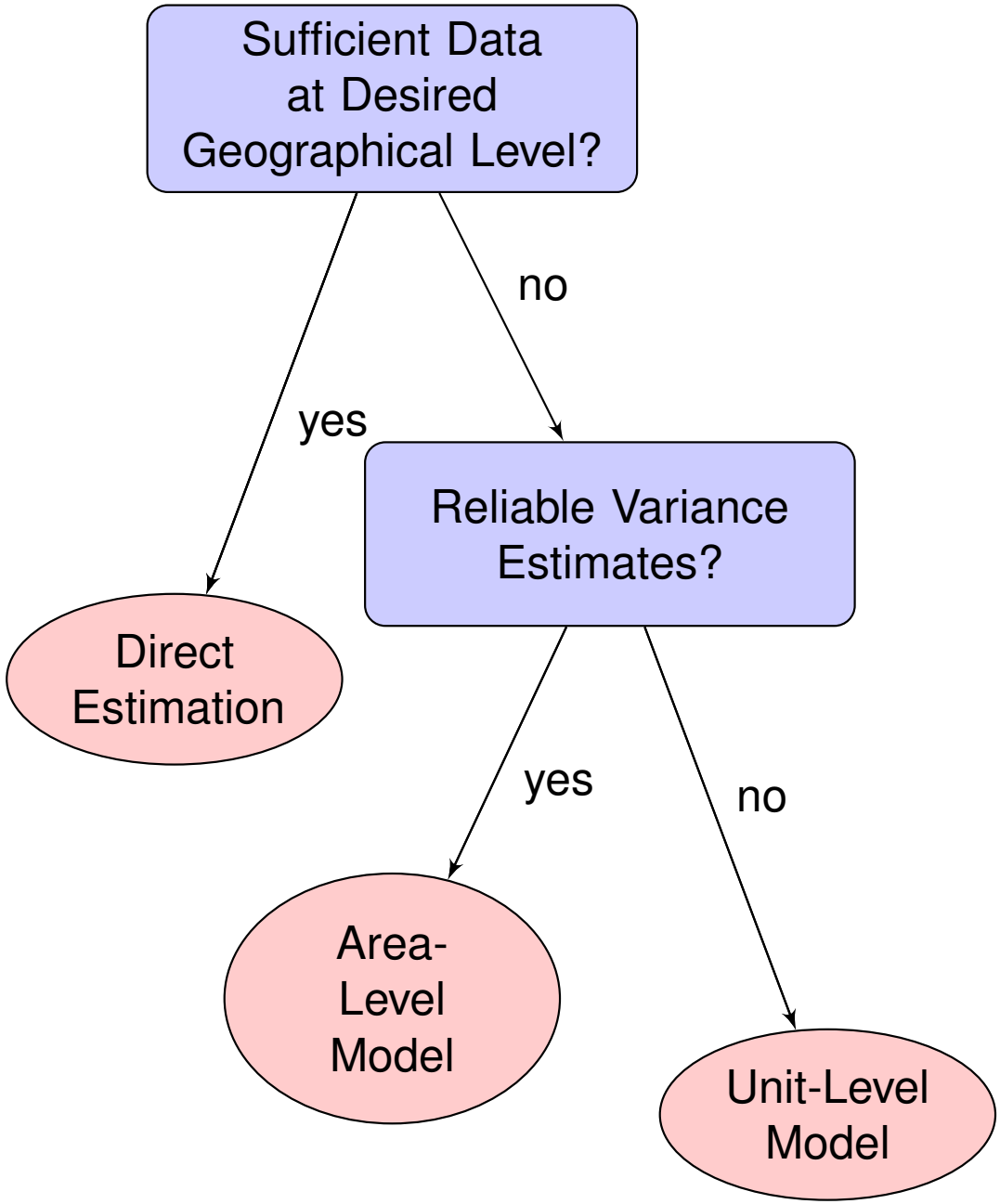


FIGURE 9: Leave-one CV for BetaBinomial models. Black dots are direct estimates. Left: 50% predictive intervals. Right: 80% predictive intervals. Top row: No urban/rural, no ANC covariate. Middle row: Urban/rural, no ANC covariate. Bottom row: Urban/rural, ANC covariate.

Discussion

Recommended Methods for SAE



Discussion

- The area-level (Fay-Herriot) model builds on the strengths of direct (weighted) estimates and spatial smoothing models.
- In the limit, as we obtain larger data in an area, the weighted estimates will dominate, which is exactly what we want!
- If insufficient samples in areas, then estimated variance is unacceptably large (or undefined), and then we need to resort to the cluster-level models.
- Unit-level models allow finer-scale modeling, but are more sophisticated, and hence trickier in practice to implement.
- Model checking techniques are still undeveloped in the SAE context.
- Prevalence mapping is still in its infancy, and currently no agreed upon “best” approach.

References

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

Malawi DHS (2016). Malawi Demographic Health Survey 2016–17. Technical report, ICF.

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

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, **4**, 395–421.

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

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

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

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

Wu, Y. and Wakefield, J. (2022). Modeling urban/rural fractions in low- and middle income countries. *Submitted*.