2022 SISCER Module 2 Small Area Estimation:
Lecture 3: Unit-Level Modeling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Motivation

Area-level models based on Fay and Herriot (1979), cannot always be used since the weighted estimates, or their variances, may be ill-defined at the geographical level (and time scale) required.

There is also the possibility of getting more informative analyses if one models at the unit-level.

An early important reference on unit-level models is Battese *et al.* (1988).

We describe unit-level models for normal and binomial data.

Basic Unit-Level Model for Normal Data

Basic Unit-Level Model

Suppose area i N_i units and in each area the indices of the sampled indices is S_i with number of elements $|S_i| = n_i$.

For a continuous response, y_{ik}, the basic unit-level model (Rao and Molina, 2015, Chapter 7) is

$$y_{ik} = \beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}_1 + \delta_i + \epsilon_{ik}$$

where

- $\boldsymbol{x}_{ik}$ are unit-level covariates, for example age, sex, race, with associated regression parameters $\boldsymbol{\beta}_1$.
- Area-specific random effects:

  $$\delta_i|\sigma_\delta^2 \sim_{iid} \text{N}(0, \sigma_\delta^2),$$

  which forms the second stage of the model.
- Unit-level errors:

  $$\epsilon_{ik}|\sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2).$$

Model is also known as the basic nested error linear regression model.

Where are the weights?

- No mention of weights!
- The required assumption is that the selection probabilities in area i do not depend on y_{ik}, but may depend on $\boldsymbol{x}_{ik}$.
- For example, for a stratified SRS we should include the stratification (design) variables amongst the $\boldsymbol{x}_{ik}$ variables.
- This model would not be appropriate for data sampled via a cluster design – for such a design, we would expect dependence

Basic Unit-Level Modeling

Inference may be carried out via likelihood or Bayes, with the latter placing priors on $\beta_0, \boldsymbol{\beta}_1, \sigma_\delta^2, \sigma_\epsilon^2$.

If a likelihood approach is taken, the random effect estimates $\widehat{\delta}_1$, are obtained as best linear unbiased predictors (BLUPs), know as Estimated BLUPs (EBLUPs) when $\beta_0, \boldsymbol{\beta}_1$ and variance components $\sigma_\delta^2, \sigma_\epsilon^2$ are estimated.

If there are no data in particular areas we can still make predictions, if we assume the model holds for all areas, but the random effect will be estimated as $\widehat{\delta}_i = 0$, and so we would expect bias.

Spatial random effects can be added to the model, to smooth over neighboring areas, and these may help to alleviate the bias, since they are based on local data.

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

Aggregation for Normal Model

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

Aggregation to obtain the mean follows from,

$$\overline{Y}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} Y_{ik}$$

where N_i is the population size in area i.

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

Remember the model:

$$y_{ik} = \beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\beta}_1 + \delta_i + \epsilon_{ik}.$$

Aggregation for Normal Model

The obvious estimator, under the model, is

$$\begin{aligned}
\widehat{\overline{Y}}_i &= \frac{1}{N_i}\sum_{k=1}^{N_i}\widehat{Y}_{ik} \\
&= \frac{1}{N_i}\sum_{k=1}^{N_i}\left(\widehat{\beta}_0 + \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1 + \widehat{\delta}_i\right) \\
&= \widehat{\beta}_0 + \overline{\boldsymbol{x}}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1 + \widehat{\delta}_i
\end{aligned}$$

where

- $\overline{\boldsymbol{x}}_i$ is the mean, across the complete population, of the covariates in area i.
- $\widehat{\delta}_i$ are the estimated random effects (can be EBLUP or Bayes).

The ϵ_{ik} are viewed as measurement error, and so not included in the predictions for the areas.

Note that if we thought the ϵ_{ik} were "real signal" then we would need to include in the area-level prediction.

Aggregation for Normal Model

Suppose we have a stratified design where the stratification variable is urban/rural.

Let:

- $x_{1ik} = 1$ is unit k in area i is urban, and $x_{1ik} = 0$ if rural.
- $x_{2ik} = 1$ is unit k in area i is rural, and $x_{2ik} = 0$ if urban.
- Note that $x_{1ik} = 1 - x_{2ik}$.
- Let $\overline{x}_{1i}$ and $\overline{x}_{2i}$ be the proportions of urban and rural units in area i, respectively.

Model:

$$y_{ik} = x_{1ik}\beta_1 + x_{2ik}\beta_2 + \delta_i + \epsilon_{ik}.$$

The area mean, under the model, is obtained through aggregation:

$$\overline{Y}_i = \overline{x}_{1i}\beta_1 + \overline{x}_{2i}\beta_2 + \delta_i.$$

Basic Unit-Level Model for Bernoulli Data

BernoulliModel

Suppose we sample indices S_i in area i with $|S_i| = n_i$ units, out of N_i, and we record a binary outcome for each.

For binary data, we may assume the model:

$$\begin{aligned} Y_{ik}|p_{ik} &\sim \text{Bernoulli}(n_{ik}, p_{ik}) \\ p_{ik} &= \text{expit}(\beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}_1 + \delta_i). \end{aligned}$$

for $k \in S_i$ with:

- $\text{expit}(z) = \exp(z)/(1 + \exp(z))$.
- y_{ik} being the k-th response in area i, $k \in S_i$, $i = 1, \dots, n$.
- p_{ik} is the probability of the event of interest – we will refer to this as the risk.
- $\boldsymbol{x}_{ik}$ is a vector of unit-level covariates, with associated regression coefficients $\boldsymbol{\beta}_1$.
- $\delta_i \sim_{iid} \text{N}(0, \sigma_\delta^2)$ are area-specific random effects.

Aggregation for the Bernoulli Model

Aggregated prevalence, under the model,

$$
\begin{aligned}
p_i &= \frac{1}{N_i} \sum_{k=1}^{N_i} p_{ik} \\
&= \frac{1}{N_i} \sum_{k=1}^{N_i} \text{expit}(\beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\beta}_1 + \delta_i)
\end{aligned}
$$

where

- N_i is the total population size in area i.
- Note that we need the covariates for each member of the population, not just the average (which was all that was required for the normal model).

We have assumed the sampled fraction of units in the area is negligible.

Spatial Models

It is straightforward to convert the above models into spatial models.

We simply replace the IID δ_i terms by BYM2 random effects.

The computation is now a bit trickier, but Bayesian inference through INLA is relatively easy, as we will see!

Two-Fold Nested Error Regression Model

Two-Fold Nested Error Regression Model: Normal

In the two-fold nested error regression model K_i primary units (or clusters) are selected in area i and then within these clusters a sample of n_{ik} units are sampled.

For simplicity, assume that units in the same cluster have the same covariates, and the same geographical location.

Normal Model:

$$Y_{ikj} = \beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}_1 + \delta_{i[k]} + \epsilon_{ikj},$$

where

- δ_i are BYM2.
- measurement error is $\epsilon_{ikj} \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2)$.

Two-Fold Nested Error Regression Model: Binomial

Binomial Model: Let $Y_i = \sum_{j=1}^{n_{ik}}$ are the total number of events of interest.

$$Y_{ik}|q_{ik} \sim \text{Binomial}(n_{ik}, q_{ik}),$$

where

- the prevalence model is:

$$q_{ik} = \text{expit}(\beta_0 + \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\beta}_1 + \delta_i + \epsilon_{ik})$$

  where

- δ_i are area-level random effects, for example, having BYM2 structure.

- the terms $\epsilon_{ik} \sim_{iid} \text{N}(0, \sigma_{\epsilon}^2)$ may reflect:
  - Measurement error, i.e., recording errors.
  - Overdispersion.
  - “True signal”, i.e., each cluster has its own unique prevalence, beyond that determined by the spatial random effect and the covariates.

Binomial Model

The interpretation of ϵ_{ik} will affect how we obtain the prevalence p_{ik} in cluster k.

If measurement error, then we can leave off and $q_{ik} = p_{ik}$.

If we assume the ϵ_{ik} are modeling overdispersion then we need to integrate out to obtain the marginal probability of the event in cluster k via

$$p_{ik} = \int_{\epsilon} \text{expit}(\beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\beta}_1 + \delta_i + \epsilon) \text{N}(\epsilon | 0, \sigma_{\epsilon}^2) \; d\epsilon.$$

One way of thinking about overdispersion is that imaginary repeated sampling from the same cluster, we are sampling different groups of units, and the prevalence of these groups is q_{ik}, which is drawn from some distribution – the prevalence we want, p_{ik}, is the average over all possible q_{ik}.

Shortly, we will see an alternative way of modeling overdispersion, via a beta-binomial model.

For more discussion see Dong and Wakefield (2021).

Notation

- Initially we describe the model in space only.
- Suppose K_i clusters are sampled within area i, $i = 1, \ldots, n$ of a study area.
- Let $\boldsymbol{s}_{ik}$ represent the geographical location of cluster k within area i with $k = 1, \ldots, K_i$ so that $K = \sum_i K_i$ is the total number of clusters.

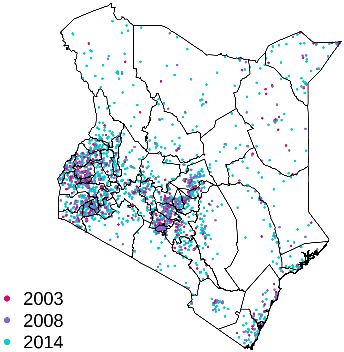


Figure 1: Cluster locations in three Kenya DHS, with county boundaries.

Beta-Binomial Unit-Level Model

- It is common to see overdispersion with spatial health and demographic data.
- One approach to modeling this phenomenon is to assume the cluster-level prevalence q_{ik} that is producing the survey data we see in cluster c, is drawn from a probability distribution.
- If we were to go back in time and draw another random sample, a different q_{ik} would result.
- The common q_{ik} to all units sampled, induces correlation between the responses.
- Overdispersion can be modeled using a random effects distribution for the prevalence.

Beta-Binomial Unit-Level Model

- Here, we suppose the cluster level variability is described by the random effects distribution:

$$q_{ik} | a_{ik}, b_{ik} \sim \text{Beta}(a_{ik}, b_{ik}),$$

with $a_{ik} = dp_{ik}$, $b_{ik} = d(1 - p_{ik})$ so that $d = a_{ik} + b_{ik}$, and

$$\begin{aligned} p_{ik} = \text{E}[q_{ik}] &= \frac{a_{ik}}{d} \\ \text{var}(q_{ik}) &= \frac{p_{ik}(1 - p_{ik})}{d + 1}. \end{aligned}$$

- The overdispersion is described by the scale parameter d.
- The intraclass correlation coefficient is the correlation between two binary outcomes in the same cluster and corresponds to $1/(d + 1)$.
- The parameters a_{ik} and b_{ik} are not the most intuitive, and it is useful to instead think about the two free parameters as being the mean p_{ik} and the scale d.

Beta-Binomial Unit-Level Model

- The sampling model corresponds to

$$\begin{aligned} Y_{ik} | q_{ik} &\sim \text{Binomial}(n_{ik}, q_{ik}) \\ q_{ik} | a_{ik}, b_{ik} &\sim \text{Beta}(a_{ik}, b_{ik}). \end{aligned}$$

- We can integrate over q_{ik} to give the marginal distribution:

$$\Pr(Y_{ik} | p_{ik}, d) = \int_{q_{ik}} \underbrace{\Pr(Y_{ik} | n_{ik}, q_{ik})}_{\text{Binomial}(n_{ik}, q_{ik})} \times \underbrace{p(q_{ik} | p_{ik}, d)}_{\text{Beta}(a_{ik}, b_{ik})} \, dq_{ik},$$

with $p_{ik} = a_{ik}/(a_{ik} + b_{ik})$ and $d = 1/(a_{ik} + b_{ik})$.

Beta-Binomial Unit-Level Model

- Turning the handle,

$$Y_{ik} | p_{ik}, d \sim \text{Beta-Binomial}(n_{ik}, p_{ik}, d),$$

with

$$\begin{aligned} \mathrm{E}[Y_{ik} | p_{ik}, d] &= n_{ik} p_{ik} \\ \mathrm{var}(Y_{ik} | p_{ik}, d) &= n_{ik} p_{ik} (1 - p_{ik}) \times \frac{n_{ik} + d}{1 + d}, \end{aligned}$$

so we have overdispersion for $d > 0$.

Beta-Binomial Unit-Level Model

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

Under the discrete spatial model,

$$p_{ik} = \text{expit}(\beta_0 + \boldsymbol{x}_{ik}^{\top}\boldsymbol{\beta}_1 + \delta_i),$$

where

- p_{ik} is the prevalence for cluster k in area i.
- $\boldsymbol{x}_{ik}$ may include the strata within which cluster k lies.
- δ_i can be BYM2.

Aggregation

- In the DHS application we describe later, we have a single covariate, the urban/rural classification of cluster *k*.

- With this simple spatial form, the modeled area prevalence is

  $$p_i = r_i \times \underbrace{\text{expit}(\beta_0 + \delta_i)}_{\text{Prevalence for Rural}} + (1 - r_i) \times \underbrace{\text{expit}(\beta_0 + \beta_1 + \delta_i)}_{\text{Prevalence for Urban}},$$

  where

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

- The original sampling frame that contains the proportions of urban/rural, is unavailable, though some information is typically available in the DHS reports.

- The proportions r_i can be obtained by thresholding population density surfaces.

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 (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.

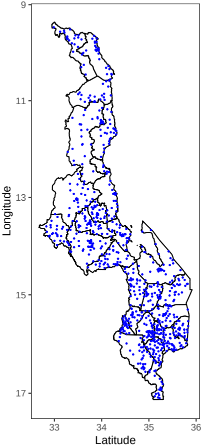


Figure 2: Cluster locations in 2015–16 Malawi DHS.

Spatial Fay-Herriot Model

- An area-level model is,

$$\text{logit}\,(\widehat{p}_i) \quad \sim \quad \text{N}(\theta_i, V_i)$$

  where

  - $\widehat{p}_i$ is the weighted estimate,
  - θ_i is the logit of the true prevalence in area i, and
  - V_i is the variance of the logit estimator.

- We model θ_i via a BYM2 specification:

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

  where δ_i are BYM2 random effects.

- We use the HIV prevalence from antenatal care (ANC) clinics, 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.

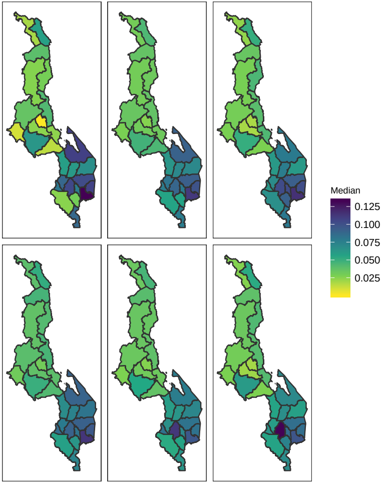


Figure 3: Estimates of HIV prevalence among females aged 15–29 in districts of Malawi in 2015–16. Top row estimates are from area-level models: direct estimates; smoothed direct estimates; smooth direct estimates with antenatal care (ANC) HIV prevalence covariate. Bottom row estimates are from unit-level models: no urban/rural adjustment and no covariate; urban/rural adjustment only; urban/rural adjustment and ANC HIV prevalence covariate.

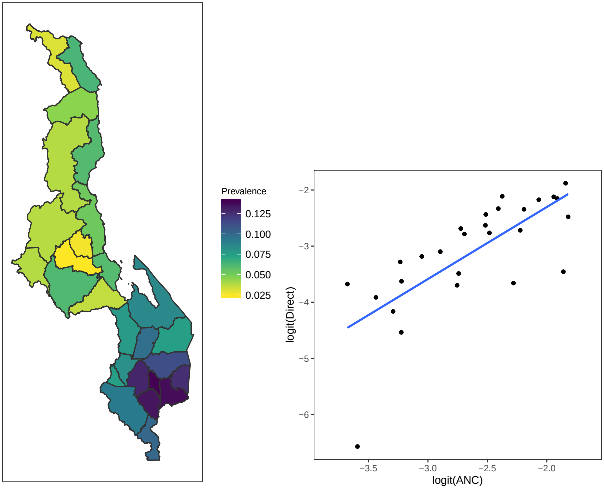


Figure 4: Left: Map of ANC prevalence. Right: logit of direct prevalence estimates versus logit of ANC prevalence estimates.

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:

$$\theta_i = \beta_0 + x_i^{\mathsf{T}} \beta_1 + \delta_i,$$

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

Unit-Level Model

- Y_{ic} and n_{ic} are the number of cases and sample sizes for cluster $k = 1, \dots, K_i$, in area $i = 1, \dots, m$.
- 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 (urban/rural here, in addition to Admin1 areas).
- 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,

$$\begin{aligned} Y_{ik} \mid p_{ik}, d &\sim \text{BetaBinomial}(n_{ik}, p_{ik}, d) \quad (1) \\ p_{ik} &= \text{expit}\left(\beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}_1 + z_{ik}\gamma + \delta_i\right) \quad (2) \end{aligned}$$

where

- z_{ik} is the urban/rural status of sampled cluster k in area i.
- d is the overdispersion parameter.
- We take δ_i as BYM2.

Unit-Level Model

Aggregation to the area-level is carried out via,

$$\begin{aligned} p_i &= r_i \times \underbrace{\text{expit}\left(\beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}_1 + \delta_i\right)}_{\text{Rural Prevalence}} \\ &+ (1 - r_i) \times \underbrace{\text{expit}\left(\beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}_1 + \gamma + \delta_i\right)}_{\text{Urban Prevalence}} \end{aligned}$$

where r_i is the proportion of the area that is rural.

These fractions need to be estimated, which may not be trivial.

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 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 $\exp(\beta_1)$.

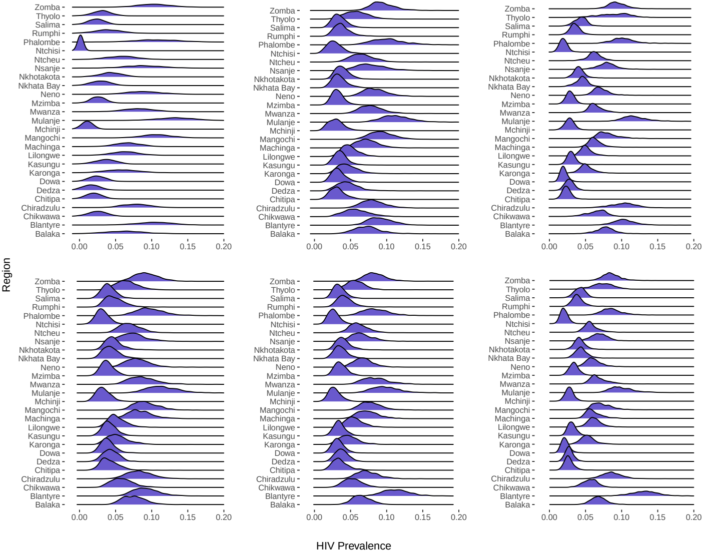


Figure 5: Posterior distributions for HIV prevalence. Top row area-level models: direct; smoothed direct; smoothed direct with ANC covariate. Bottom row unit-level (betabinomial) models: no urban/rural, no covariate; urban/rural only; urban/rural and ANC covariate.

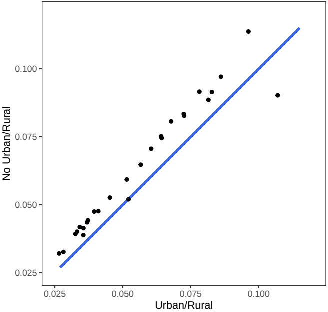


Figure 6: District prevalence estimates from two unit-level models. On the y-axis, the prevalence estimates are from a model with no urban/rural adjustment, while on the x-axis the model has an adjustment.

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

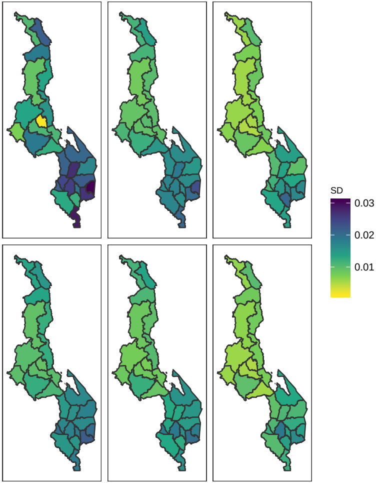


Figure 7: Uncertainty estimates (standard errors for direct estimates, posterior standard deviations for the remainder) of HIV prevalence. Top: area-level models: direct estimates; smoothed direct estimates with no ANC covariate; smooth direct estimates with ANC covariate. Bottom: unit-level models: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

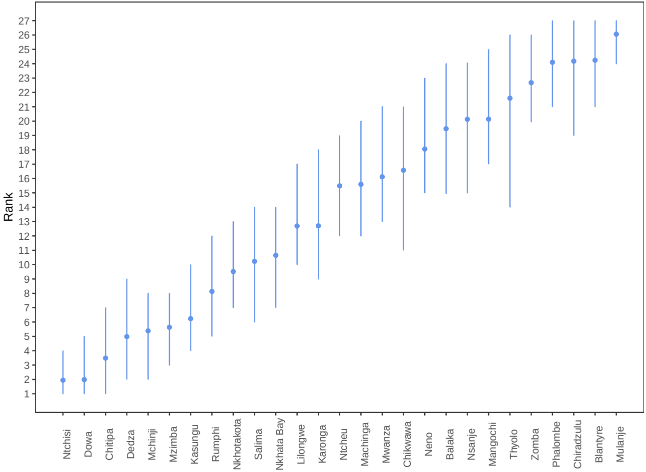


Figure 8: Distributions on the rankings for the smoothed direct estimates with the ANC covariate. The lines represent 90% intervals based on samples from the posterior, with rank = 1 on the y-axis corresponding to the lowest HIV prevalence and rank = 27 corresponding to the highest HIV prevalence.

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

$$\text{logit}(\widehat{p}_i) \sim \text{N}(\theta_i, V_i).$$

Model Assessment

- 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 $\text{logit}(\widehat{p}_i^{\text{HT}})$ (Horvitz-Thompson weighted estimates).

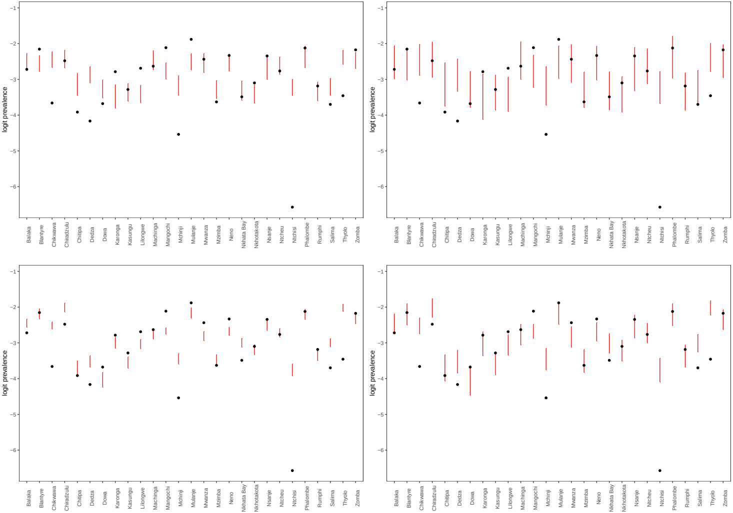


Figure 9: Leave-one cross-validation predictions for the smoothed direct models. Black dots are the direct estimates. Left column: 50% predictive intervals. Right: 80% predictive intervals. Top row: No ANC covariate. Bottom row: ANC covariate.

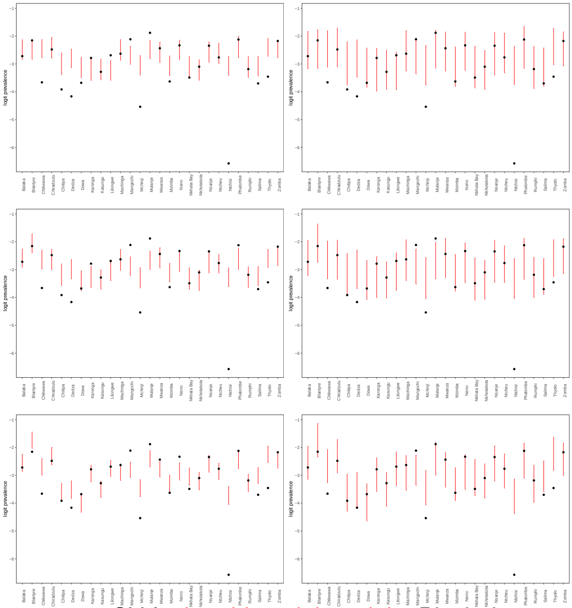
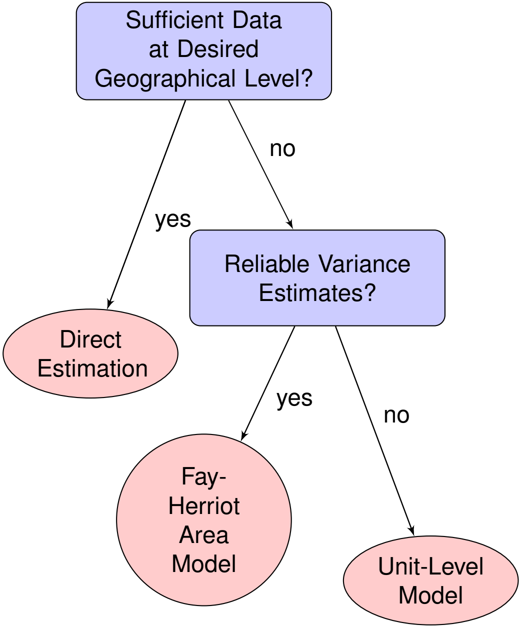


Figure 10: Leave-one CV for betabinomial models. Black dots are direct estimates. Left: 50% predictive intervals. Right: 80% predictive intervals. Top row: No urban/rural, no ANC covariate. Middle row: Urban/rural, no ANC covariate. Bottom row: Urban/rural, ANC covariate.

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 (Gaussian process (GP)) models 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.

Recommended Methods for SAE



Sufficient Data at Desired Geographical Level?

yes

no

Direct Estimation

Reliable Variance Estimates?

yes

no

Fay-Herriot Area Model

Unit-Level Model

Discussion

- The spatial Fay-Herriot modell 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:
    - Discrete spatial models are easier to fit, and aggregation more straightforward.
    - Continuous spatial models are more challenging to fit, and aggregation more challenging.

- 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.

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.

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*.

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

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

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

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