Bayesian SAE with an Emphasis on Low- and Middle-Income Countries: Lecture 3: Area-Level SAE Models

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Beyond Weighted Estimates

- **Aim of Small Area Estimation (SAE):** Provide inferential summaries across geographical areas, using survey data – in a LMICs context, the household surveys use stratified two-stage unequal probability cluster sampling.
- Recall the weighted estimator, which provides a simple approach and has many desirable properties in data-rich situations.
- As data become sparser, weighted estimates can be highly unstable (high variance) or we can't calculate a measure of uncertainty.
- One then needs to consider SAE models, of which there are two distinct types:

1: Area-Level Models	2: Unit-Level Models

The Area-Level Fay-Herriot Model

- An amazingly advanced model was introduced by Fay and Herriot (1979).
- The basic idea behind the area-level Fay-Herriot model, is to simultaneously model the collection of weighted estimates

$$\widehat{\theta}_i^{\mathrm{w}} \quad = \quad \frac{\widehat{T}_i}{\widehat{N}_i} \quad = \quad \frac{\sum_{k \in s_i} w_{ik} Y_{ik}}{\sum_{k \in s_i} w_{ik}}. \tag{1}$$

  from all areas.
- These estimates, $\widehat{\theta}_i^{\mathrm{w}}$, along with their standard errors, $\widehat{V}_i^{1/2}$, constitute the data for $i = 1, \ldots, m$, areas.
- We model between-area variation in $\widehat{\theta}_i^{\mathrm{w}}$ using covariates and by assuming the residuals arise from a common distribution.
- We are effectively increasing the sample size in each area, by leveraging similarity in the means.

The Non-Spatial Area-Level Model

The Most Basic Form of Area-Level Model

- We let $Y_i = \widehat{\theta}_i^{\text{w}}$ correspond to the "data" and $\text{E}[Y_i] = \theta_i$ – here we are performing design-based inference and the expectation is over samples that could have been drawn, and θ_i is the true finite population mean in area i.

- We describe a two-stage model in which we have a sampling model and a linking model.

- If there are sufficient samples within the area, a design-based asymptotic arguments (Breidt and Opsomer, 2017) gives $Y_i \mid \theta_i \sim \text{N}(\theta_i, \widehat{V}_i)$ where $\widehat{V}_i$ is the design-based variance.

- When we fit this model, it is an example of indirect estimation because the data from all areas is being used to enhance the estimates in each area.

The Most Basic Form of Area-Level Model

The simplest version of the area-level model is:

$$Y_i = \theta_i + \epsilon_i, \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_i) \tag{2}$$

$$\theta_i = \alpha + u_i, \qquad u_i \sim_{iid} \mathrm{N}(0, \sigma_u^2). \tag{3}$$

- In the first stage of the model, in line (2) we have the sampling model with variances $\widehat{V}_i$ treated as known.
- The notation ind is short for independent.
- In the second stage of the model, in line (3) we have the linking model which ties together $\theta_1, \ldots, \theta_m$.
- The notation iid is short for independent and identically distributed.
- Parameters: $\alpha = \mathrm{E}[\theta_i]$ is the intercept and is the average of the population means, $\sigma_u^2 = \mathrm{var}(\theta_i)$ is the variance of the second-stage residuals u_i.

Estimation for the Area-Level Model

- Suppose for simplicity that α and σ_u^2 are known.

- In this case the posterior mean estimate of θ_i is the so-called Best Linear Unbiased Predictor (BLUP), which is the posterior mean,

$$\begin{aligned}\widehat{\theta}_i^{\text{BLUP}} = \mathsf{E}[\theta_i \mid y_i] &= Y_i - \frac{\widehat{V}_i}{\widehat{V}_i + \sigma_u^2}(Y_i - \alpha) \\ &= F_i \times Y_i + (1 - F_i) \times \alpha\end{aligned}$$

  where

$$F_i = \frac{\sigma_u^2}{\widehat{V}_i + \sigma_u^2} \leq 1$$

  is the shrinkage factor – estimates are shrunk towards α.

- The intuition is that the estimate is a compromise between the observed data in the area, Y_i, and the average of the data over the complete set of areas, α.

Estimation for the Area-Level Model

Recall,

$$\widehat{\theta}_i^{\text{BLUP}} \quad = \quad F_i \times Y_i + (1 - F_i) \times \alpha, \qquad F_i = \frac{\sigma_u^2}{\widehat{V}_i + \sigma_u^2}$$

- A measure of uncertainty is the posterior variance,

  $$\text{var}(\theta_i \mid y_i) = F_i \times \widehat{V}_i,$$

  so that the uncertainty is reduced, compared to the direct (weighted) estimate variance ($\widehat{V}_i$).

- Notice that if σ_u^2 is big and/or $\widehat{V}_i$ is small, we have $F_i \approx 1$ so that

  $$\widehat{\theta}_i^{\text{BLUP}} \approx Y_i, \qquad \text{var}(\theta_i \mid y_i) \approx \widehat{V}_i.$$

  If we imagine the sample size increasing in an area, then we are in this situation as $\widehat{V}_i \to 0$.

- Informally, this explains why Fay-Herriot estimates are design consistent.

Fay-Herriot Model with Covariates

To model between-area variation in Y_i we may introduce area-level covariates $\boldsymbol{x}_i$, a $p \times 1$ vector.

With covariates:

$$Y_i = \theta_i + \epsilon_i, \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_i) \tag{4}$$

$$\theta_i = \alpha + \boldsymbol{x}_i^{\mathsf{T}}\beta + u_i, \qquad u_i \sim_{iid} \mathrm{N}(0, \sigma_u^2). \tag{5}$$

- The BLUP becomes:

$$\begin{aligned} \widehat{\theta}_i^{\mathrm{BLUP}} = \mathrm{E}[\theta_i \mid y_i] &= Y_i - \frac{\widehat{V}_i}{\widehat{V}_i + \sigma_u^2}(Y_i - \alpha - \boldsymbol{x}_i^{\mathsf{T}}\beta) \\ &= F_i \times Y_i + (1 - F_i) \times (\alpha + \boldsymbol{x}_i^{\mathsf{T}}\beta) \end{aligned}$$

with $F_i = \sigma_u^2/(\widehat{V}_i + \sigma_u^2)$, as before.

Including Covariates

- Covariates derived from a census are desirable but in LMICs censuses are infrequently carried out, and the data may not be easily accessible at the level required.

- Instead, satellite data are often used.

- It is appealing to use survey covariates, but there are two issues that require consideration:

    1. The covariates are not population-level summaries, but rather estimates – this errors-in-variables aspect requires careful thought see, for example, Ybarra and Lohr (2008).
    2. Correlated errors at the first stage of the model, if response and covariate are taken from the same survey – we discuss this a little further.

Survey-Based Covariates

- For simplicity assume a univariate x_i – if the same response data Y_i and covariate data X_i (the version with error) are from the same survey then we violate a key assumption of the model.

- We have

$$\begin{aligned} Y_i &= \theta_i + \epsilon_i, \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_{yi}) \\ X_i &= x_i + \eta_i, \qquad \eta_i \sim_{ind} \mathrm{N}(0, \widehat{V}_{xi}) \end{aligned}$$

- Because both the area-level response Y_i and the auxiliary covariate X_i are direct (weighted) estimates derived from the same survey sample, the standard Fay-Herriot assumption of mutually independent sampling errors $\mathrm{cov}(\epsilon_i, X_i) = 0$ is violated.

- The shared sampling variance induces an endogeneity problem via area-level measurement error.

The Linear Mixed Effects Model

We can write the two-stage model (4) and (5) as

$$Y_i = \alpha + \boldsymbol{x}_i^{\mathsf{T}}\beta + u_i + \epsilon_i,$$

with

$$u_i \mid \sigma_u^2 \sim_{iid} \mathrm{N}(0, \sigma_u^2), \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_i), \qquad i = 1, \ldots, m.$$

In this model:

- α and β are fixed effects.
- $u_1, \ldots, u_m$ are random effects – they have a distribution!
- Hence, we have a mixed effects model, and it is linear, and so, more specifically, it is a linear mixed effects model (LMEM).
- Note: This is a frequentist distinction since in a Bayesian framework, all parameters are random.

The Spatial Area-Level Model

Area-Level (Fay-Herriot) Spatial Model

Recall the area-level Fay-Herriot model:

$$\begin{aligned} Y_i &= \theta_i + \epsilon_i, \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_i) \\ \theta_i &= \alpha + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + u_i, \end{aligned}$$

where previously we assumed $u_i \mid \sigma_u^2 \sim_{iid} \mathrm{N}(0, \sigma_u^2)$, $i = 1, \ldots, m$.

- In the spatial version of the model, the "residuals" (random effects)

  $$\boldsymbol{u} = (u_1, \ldots, u_m)$$

  are modeled in such a way that if areas i and j are "close" they are more likely to be similar.

- The most common way in which closeness is defined, is to say that areas i and j are neighbors if they share a common boundary.

Neighborhood Graph in Rwanda

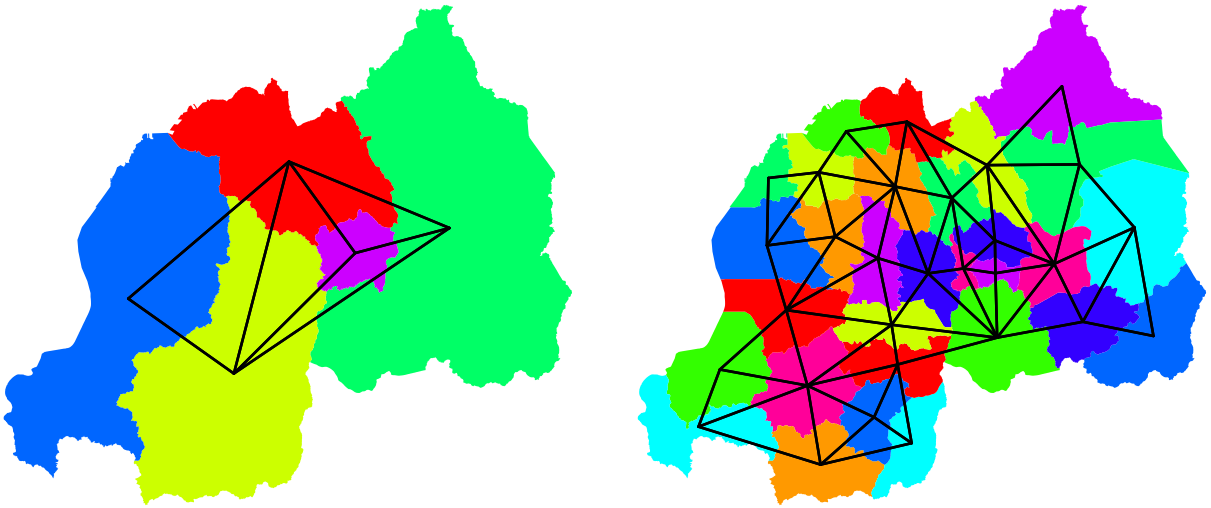

Figure 1: Neighborhood structure for 5 Admin1 areas and 30 Admin2 areas in Rwanda – this is used in the spatial smoothing models. Black lines link neighbors.

Area-Level (Fay-Herriot) Spatial Model

There are many spatial modeling possibilities for $\boldsymbol{u}$ (Banerjee *et al.*, 2015) including:

- Conditional Autoregressive (CAR) model.
- Simultaneous Autoregressive (SAR) model.
- Intrinsic CAR (ICAR) model.
- Combining an independent and identically distributed (IID) model with an ICAR model, to give the famous Besag, York, Mollié (BYM) model (Besag *et al.*, 1991).
- We describe a reformulation of the BYM model which has certain advantages – known as the BYM2 model and due to Riebler *et al.* (2016).

Area-Level (Fay-Herriot) Spatial Model

We write

$$\boldsymbol{u} \mid \sigma_u^2, \phi \sim \text{BYM2}(\sigma_u^2, \phi).$$

The spatial residual is decomposed as:

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} = \lambda \left(\sqrt{\phi} \begin{bmatrix} s_1 \\ \vdots \\ s_m \end{bmatrix} + \sqrt{1-\phi} \begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix} \right)$$

where

- The s_i are the spatial terms and the $e_i \sim \text{N}(0, 1)$ are the non-spatial (IID) terms, $i = 1, \ldots, m$.
- Two parameters in the model: ϕ is the proportion of the variation that is spatial and λ is the residual standard deviation – tells us how bumpy the surface is and corresponds to the smoothing parameter.

The ICAR Spatial Model

- The ICAR model is very simple – the "best guess" of the level is taken as the mean of the neighbors, with variation allowed around that level.

- The explicit form for the ICAR spatial terms is,

$$s_i \mid s_j, j \in \text{ne}(i), \sigma_s^2 \sim_{ind} \text{N}\left(\overline{s}_i, \frac{\sigma_s^2}{m_i}\right),$$

  which is a local smoothing model in which:

  - $\text{ne}(i)$ is the set of neighbors of i, of which there are m_i,
  - $\overline{s}_i = \frac{1}{m_i} \sum_{j \in \text{ne}(i)} s_j$ is the mean of the neighbors,
  - σ_s^2 is the variance that describes the variation around the conditional mean.

The ICAR Spatial Model

- We can write the spatial model as:

$$\begin{aligned}\pi(\boldsymbol{s} \mid \sigma_s^2) &\propto (\sigma_s^2)^{-(m-1)/2} \exp\left(-\frac{1}{2\sigma_s^2}\sum_{i\sim j}(s_i - s_j)^2\right) \\ &= (\sigma_s^2)^{-(m-1)/2} \exp\left(-\frac{\boldsymbol{s}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{s}}{2\sigma_s^2}\right)\end{aligned}$$

- The "precision" matrix $\boldsymbol{Q}$ associated has elements

$$Q_{ij} = \begin{cases} m_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

- This model is extremely computationally efficient to work with, because $\boldsymbol{Q}$ is sparse.

The ICAR Spatial Model

- In the weeds, in the implementation we use, the ICAR precision is scaled so that the variance σ_s^2 has a marginal interpretation, see Riebler *et al.* (2016) for details.

- This leads to a precision matrix $\boldsymbol{Q}_\star$ and scaled random effects $\boldsymbol{s}_\star$.

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} = \sigma_u^2 \left(\sqrt{\phi} \begin{bmatrix} s_{\star,1} \\ \vdots \\ s_{\star,m} \end{bmatrix} + \sqrt{1-\phi} \begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix} \right)$$

  where $e_i \sim_{iid} \mathrm{N}(0,1)$, $i = 1, \ldots, m$, as before.

- We have

$$\mathrm{var}(\boldsymbol{u} \mid \sigma_u^2, \phi) = \sigma_u^2 [\, \phi \boldsymbol{Q}_\star^- + (1-\phi)\boldsymbol{I} \,]$$

  where $\boldsymbol{Q}_\star^-$ is the generalized inverse.

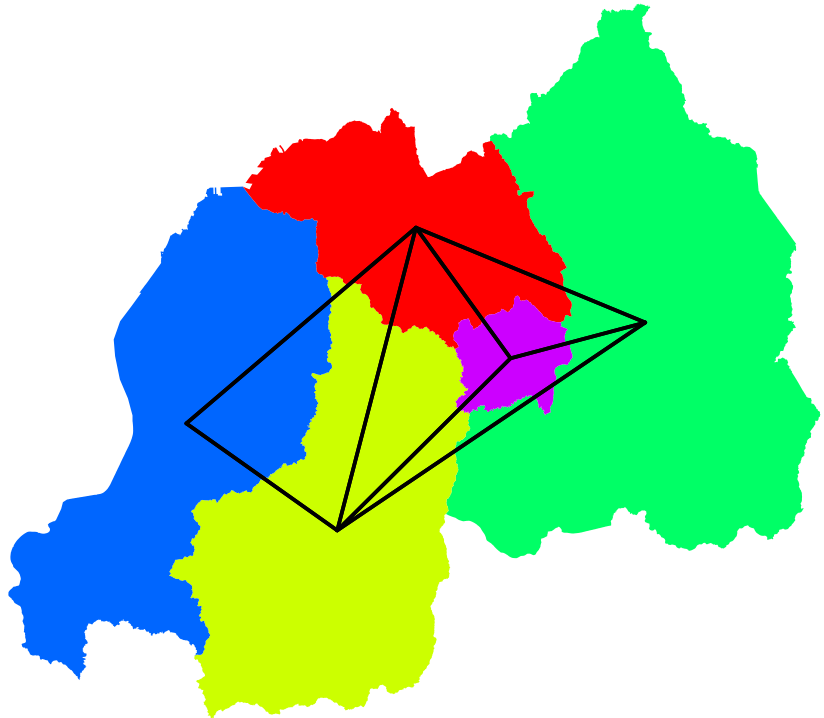

FIGURE 2: Neighborhood structure for 5 Admin1 areas. Black lines link neighbors.

- For example, the level for area 1 (note: numbering is arbitrary) is being pulled towards the average of the levels in areas 2, 3, 4 (its neighbors):

$$\mathrm{E}[S_1 \mid S_2, S_3, S_4] = \frac{1}{3}(S_2 + S_3 + S_4)$$

Area-Level Prevalence Models

Area-Level Prevalence Models

- When the target variable is not on the whole of the real line, we may choose to take a transformation.
- Fay and Herriot (1979) carried out SAE for per capita income for local geographic areas like townships, cities, and boroughs level — the response variable is positive and the log transformation was taken.
- Prevalences are by far and away the most common response types in the DHS.
- Prevalances lie between 0 and 1, which makes modeling a little more tricky – various transformations are possible and we choose the logit transform; the inverse arcsin is an alternative.

Area-Level Prevalence Models

Let,

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right), \qquad i = 1, \ldots, m,$$

be the log odds of the observed prevalence, which is on the real line.

Modeling proceeds as before, but on the logit scale.

The response variable is taken as,

$$Y_i = \log\left(\frac{\widehat{p}_i^{\,w}}{1-\widehat{p}_i^{\,w}}\right), \qquad \text{for } i = 1, \ldots, m.$$

Input Data:

$$\underbrace{(Y_1, \widehat{V}_1), (Y_2, \widehat{V}_2), \quad \ldots, \quad , (Y_{m-1}, \widehat{V}_{m-1}), (Y_m, \widehat{V}_m)}_{\text{AREAS ARE LINKED THROUGH COMMON MODEL}}$$

Area-Level Prevalence Models

In large samples,

$$Y_i = \underbrace{\theta_i}_{\text{TRUTH}} + \underbrace{\epsilon_i}_{\text{SAMPLING ERROR}},$$

with

- ϵ_i following a N(0, $\widehat{V}_i$) distribution,
- $\widehat{V}_i$ estimated using formulas relevant to the survey design.
- We might expect the logit transform to improve the normal approximation to the sampling distribution, but note that $\mathrm{E}[Y_i] \neq \theta_i$ though Y_i is a consistent estimator of θ_i – later we will discuss this aspect further in the context of unmatched linking models.

Area-Level Prevalence Models

- As with the area-level model we described previously, we can have IID or spatial random effects and can include area-level covariates.

- In the Fay-Herriot spatial model with covariates we have:

$$
\begin{aligned}
Y_i &= \theta_i + \epsilon_i, \qquad \epsilon_i \sim_{ind} \mathrm{N}(0, \widehat{V}_i) \\
\theta_i &= \underbrace{\alpha}_{\text{OVERALL LEVEL}} + \underbrace{\boldsymbol{x}_i\beta}_{\text{COVARIATE MODEL}} + \underbrace{u_i}_{\text{SPATIAL RESIDUAL}}, \qquad i = 1, \ldots, m
\end{aligned}
$$

The spatial terms $u_1, \ldots, u_m$ are assigned a BYM2 model with parameters σ_u^2, ϕ.

Area-Level (Fay-Herriot) Model

Advantages:

- Weighted estimates and uncertainty measures that feed into the hierarchical model account for the design, which avoids potential bias due to (say) stratified and PPS sampling, and accounts for the clustering.
- As the sample size in each area increases, the estimates tend to the true prevalences, which is known as design consistency.
- By modeling all the data in unison, uncertainty in each area is reduced (on average).
- The benefits arise from both the random effects and the covariates.

Area-Level (Fay-Herriot) Model

Disadvantages:

- The modeling is based on the weighted estimates and standard errors in each area, so when the data are sparse and estimates are not available or unreliable, the method cannot be used.
- 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).

Subnational NMR Estimation in Rwanda: Area-Level Models

Rwanda DHS

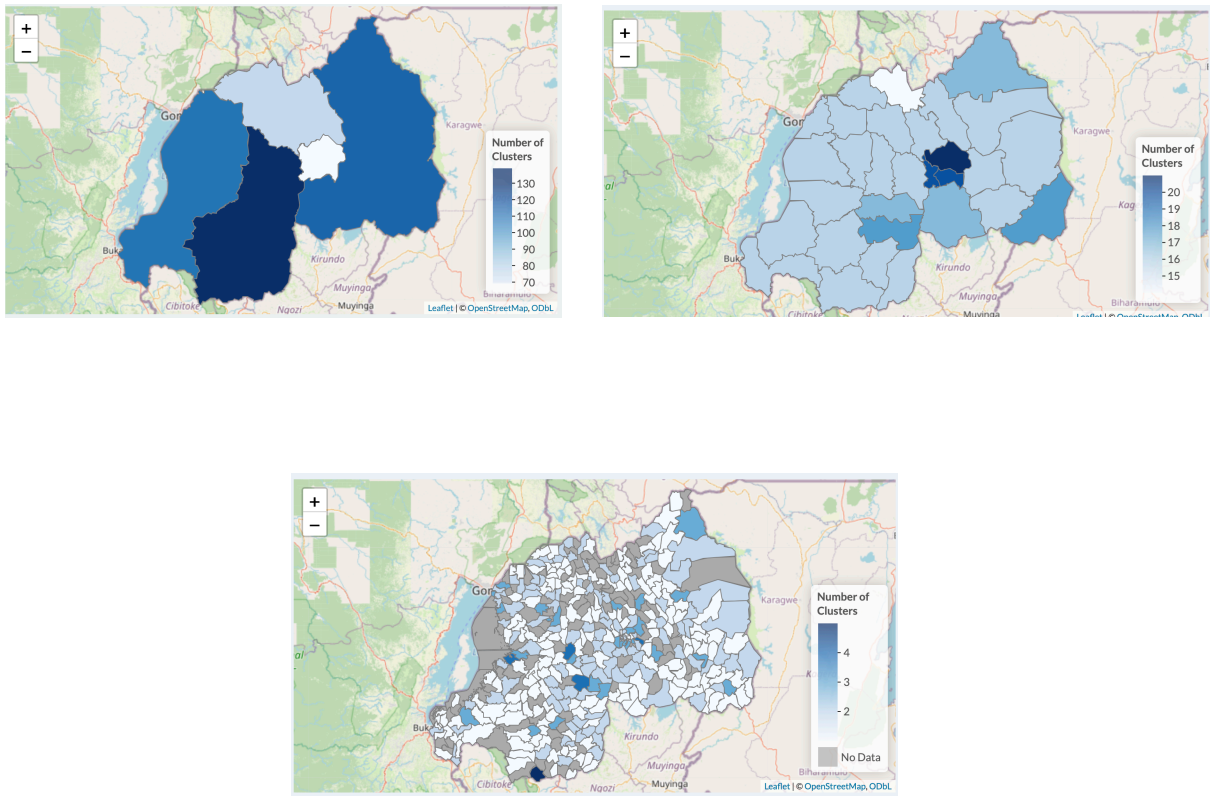

Figure 3: Number of sampled clusters, for Admin1/Admin2/Admin3 areas, from Rwanda 2019 DHS.

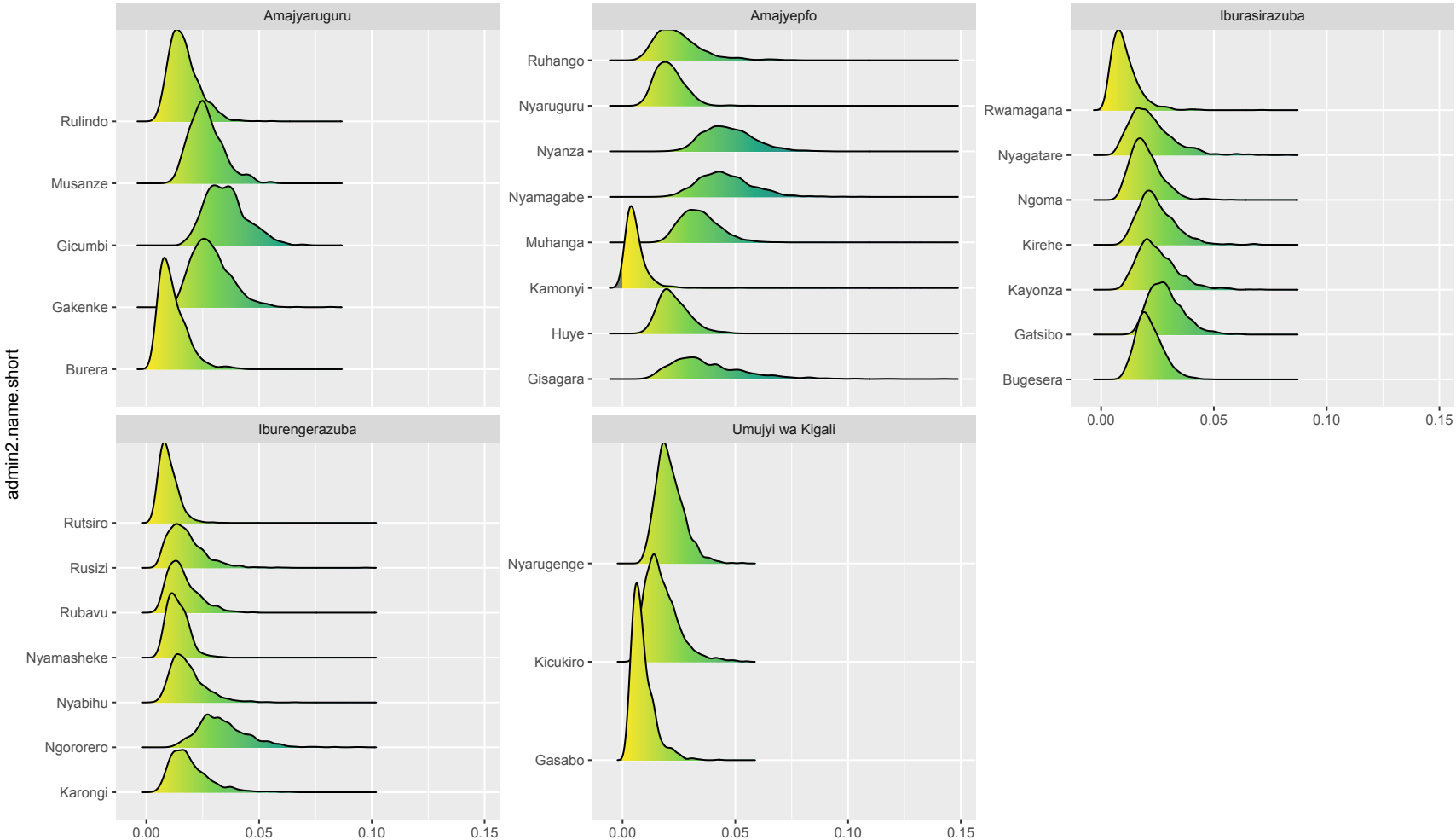

FIGURE 4: Ridgeplots for Admin2 areas *within* Admin1, direct estimates.

- For some Admin1, the constituent Admin2 estimates have similar distributions, while for others they are quite different.

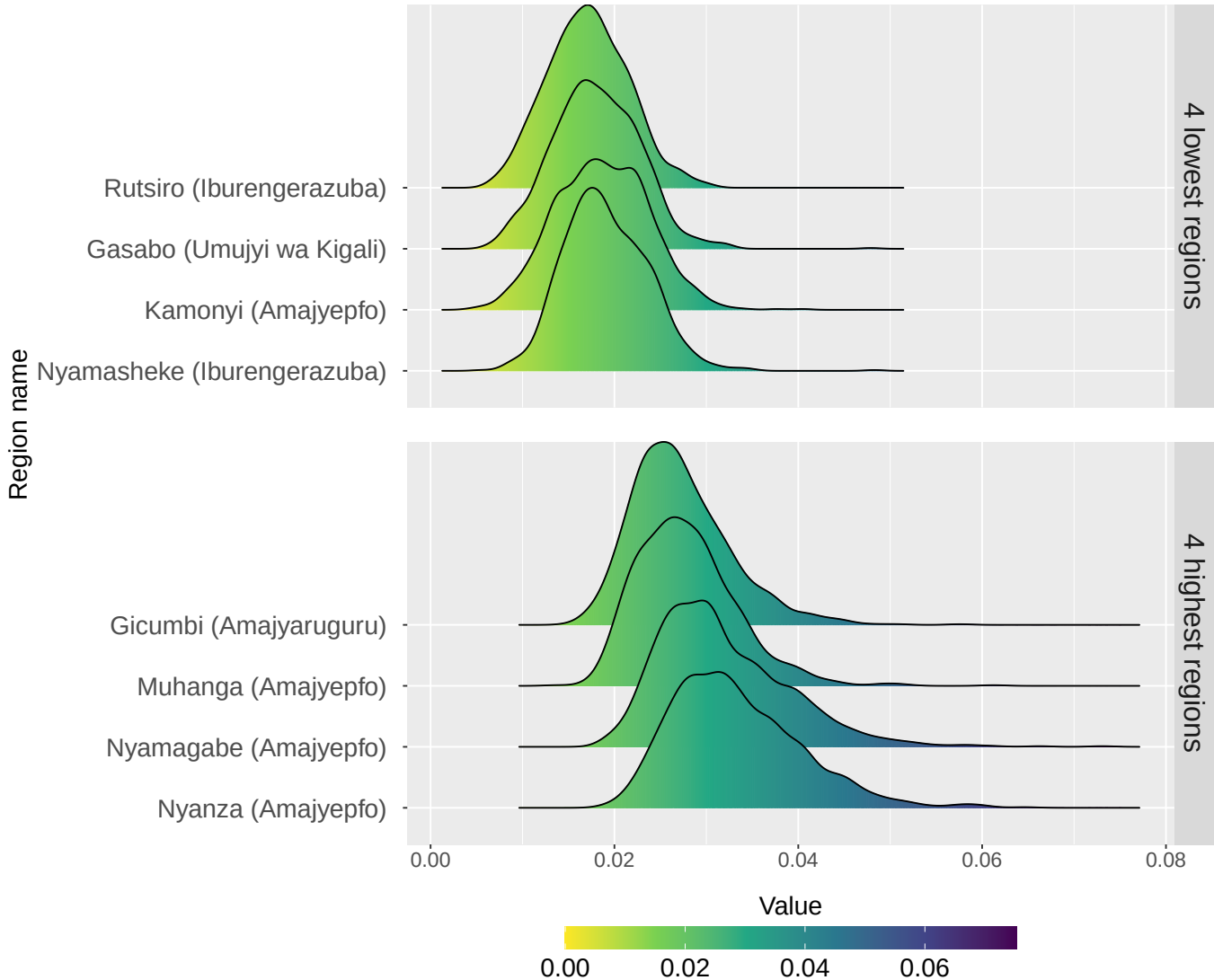

FIGURE 5: Ridgeplots for extreme (low and high) Admin2 areas.

- Lack of information at Admin2, which explains the relatively wide distributions and large overlap across areas.

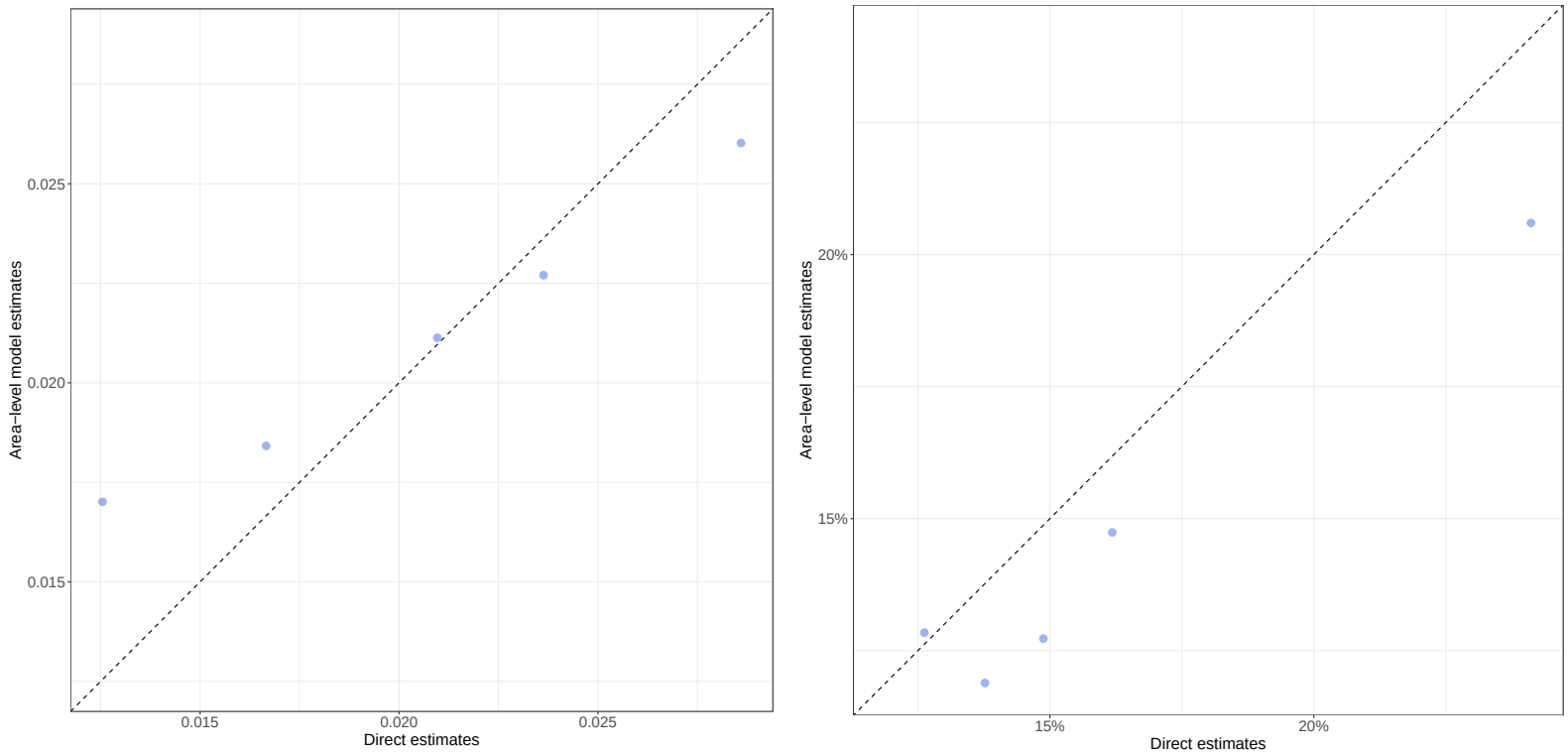

FIGURE 6: Area-level estimates versus direct estimates (left) and area-level CVs versus direct CVs (right), at Admin1.

- Some shrinkage (narrowing of range of estimates) of estimates at Admin1 (left).
- Gains in precision in all but one Admin1 area (right).

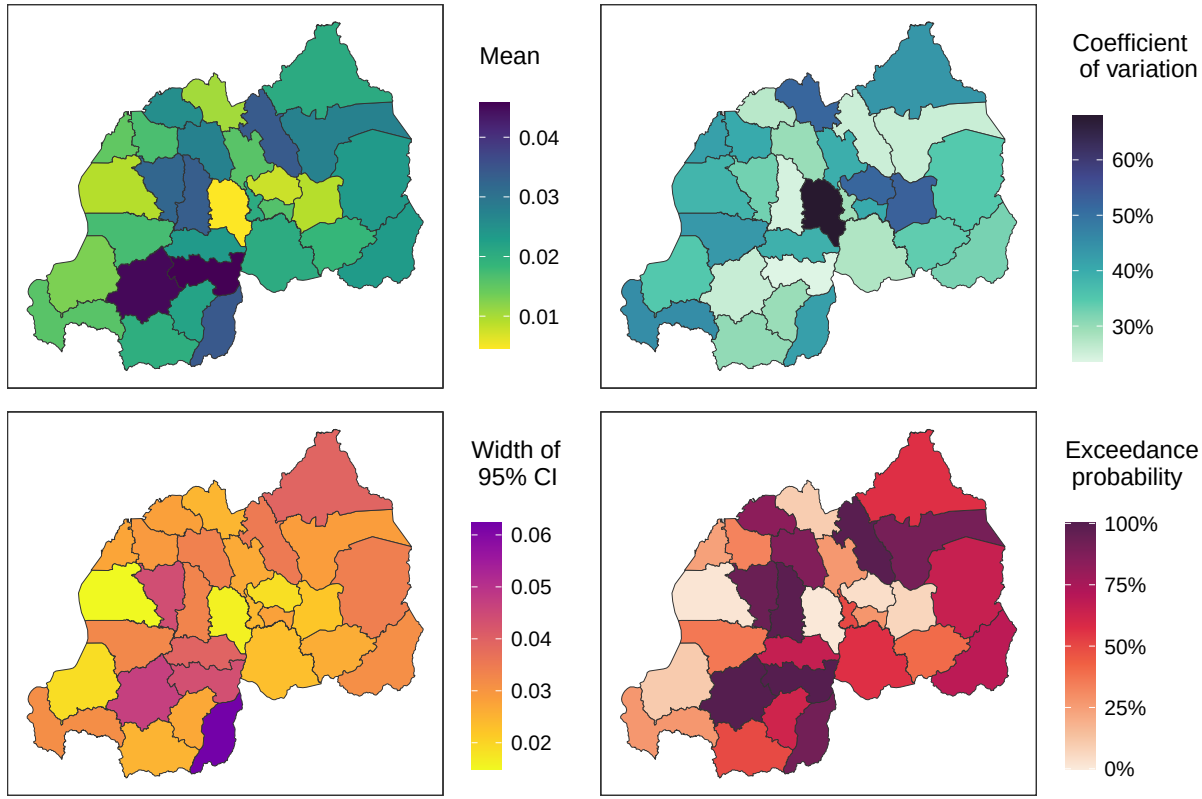

FIGURE 7: Summary measures for direct estimates at Admin2.

- Relatively large range in point estimates.
- Uncertainty measure show large variation also, and for some areas, estimates are quite imprecise.

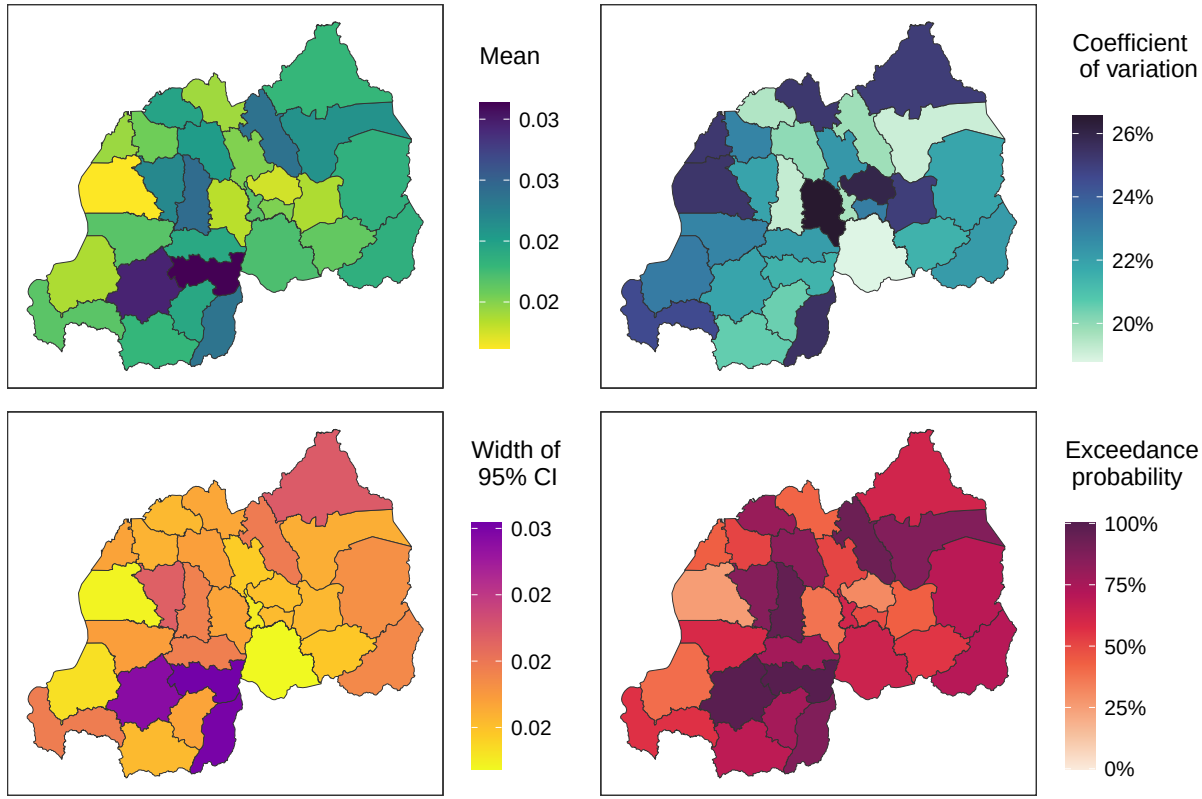

FIGURE 8: Summary measures for area-level (Fay-Herriot) estimates at Admin2.

- Compared to direct estimates (Figure 7) a narrower range in the point estimates.
- Uncertainty measure show less variation also, due to sharing of information.

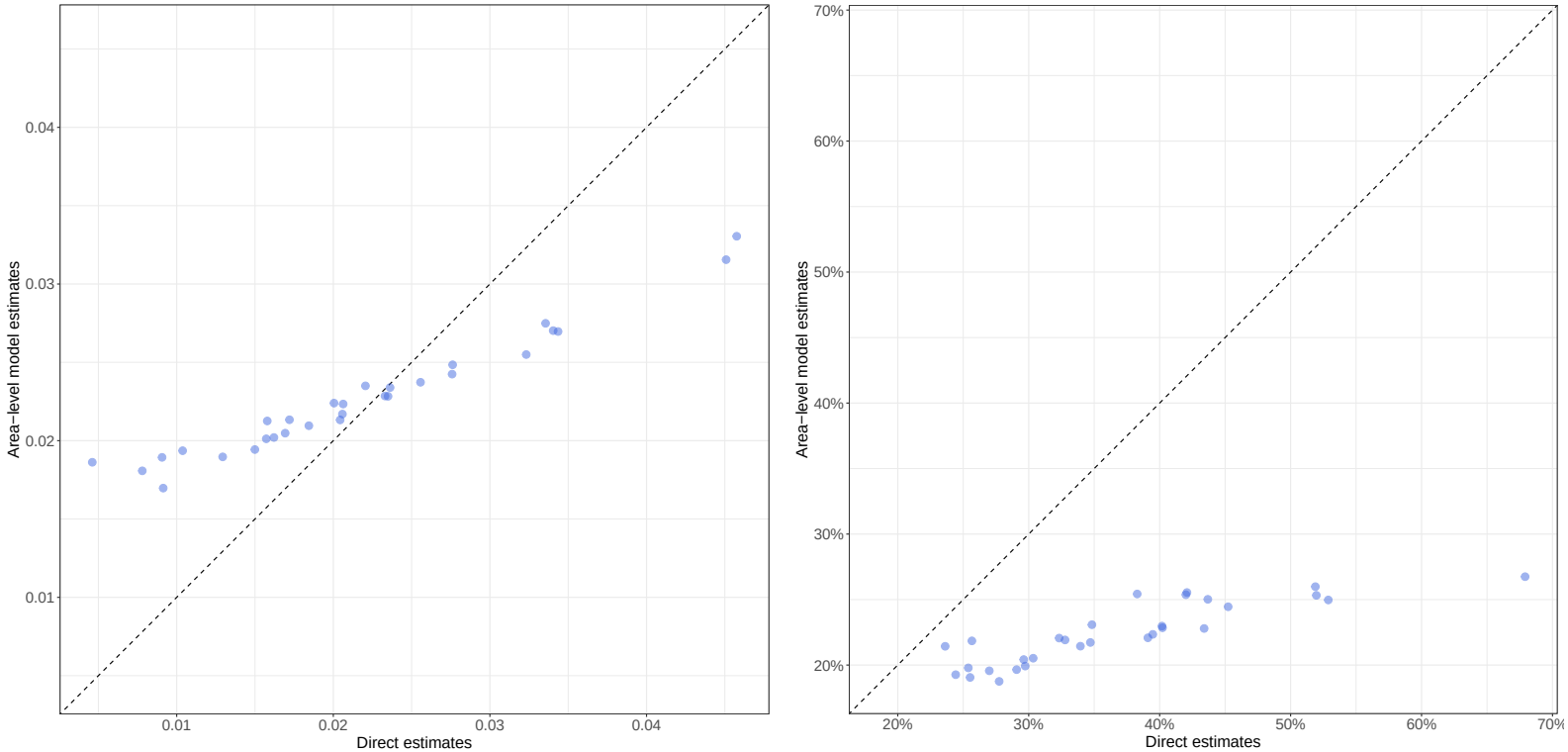

FIGURE 9: Area-level estimates versus direct estimates (left) and area-level CVs versus direct CVs (right), at Admin2.

- As compared to Figure 6, much greater shrinkage at Admin2 (left).
- As compared to Figure 6, much greater gains in precision in all Admin2 areas (right).

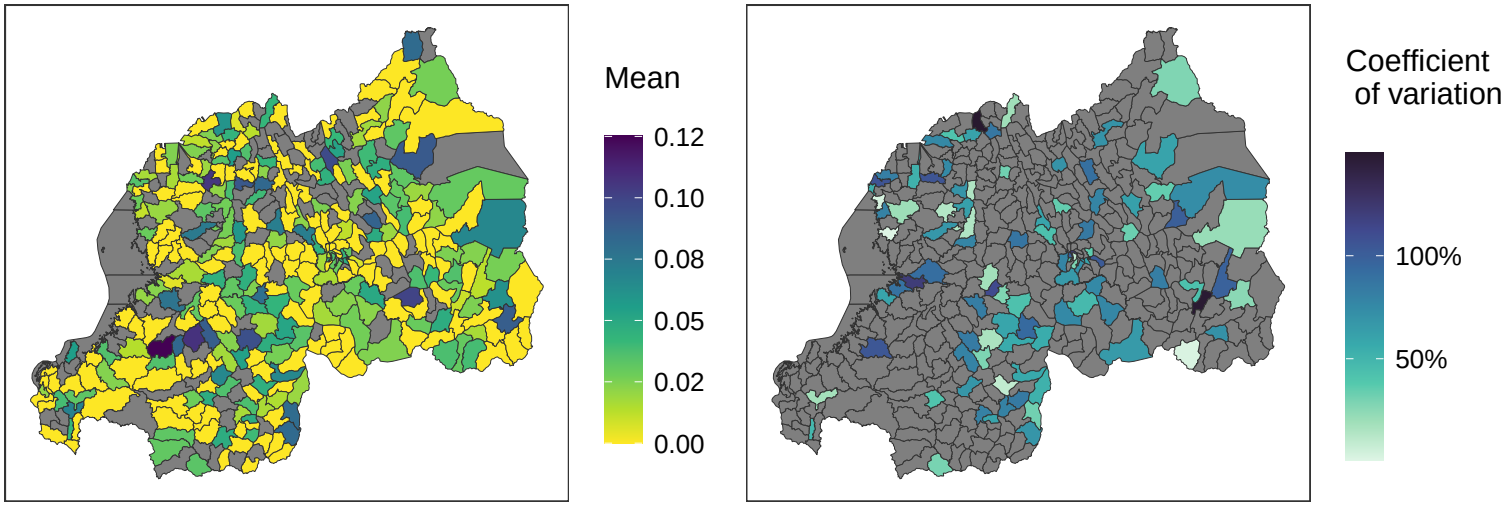

FIGURE 10: Direct prevalence estimates and CV maps at Admin3.

- Direct estimates are not available in quite a few areas (since no data).
- CVs are available in very few, because even if data are available, some configurations do not allow calculation of a variance, e.g., all responses zero.
- So unable to fit area-level models at Admin3 – unit-level models are the only possibility.
- The shinyApp gives warnings.

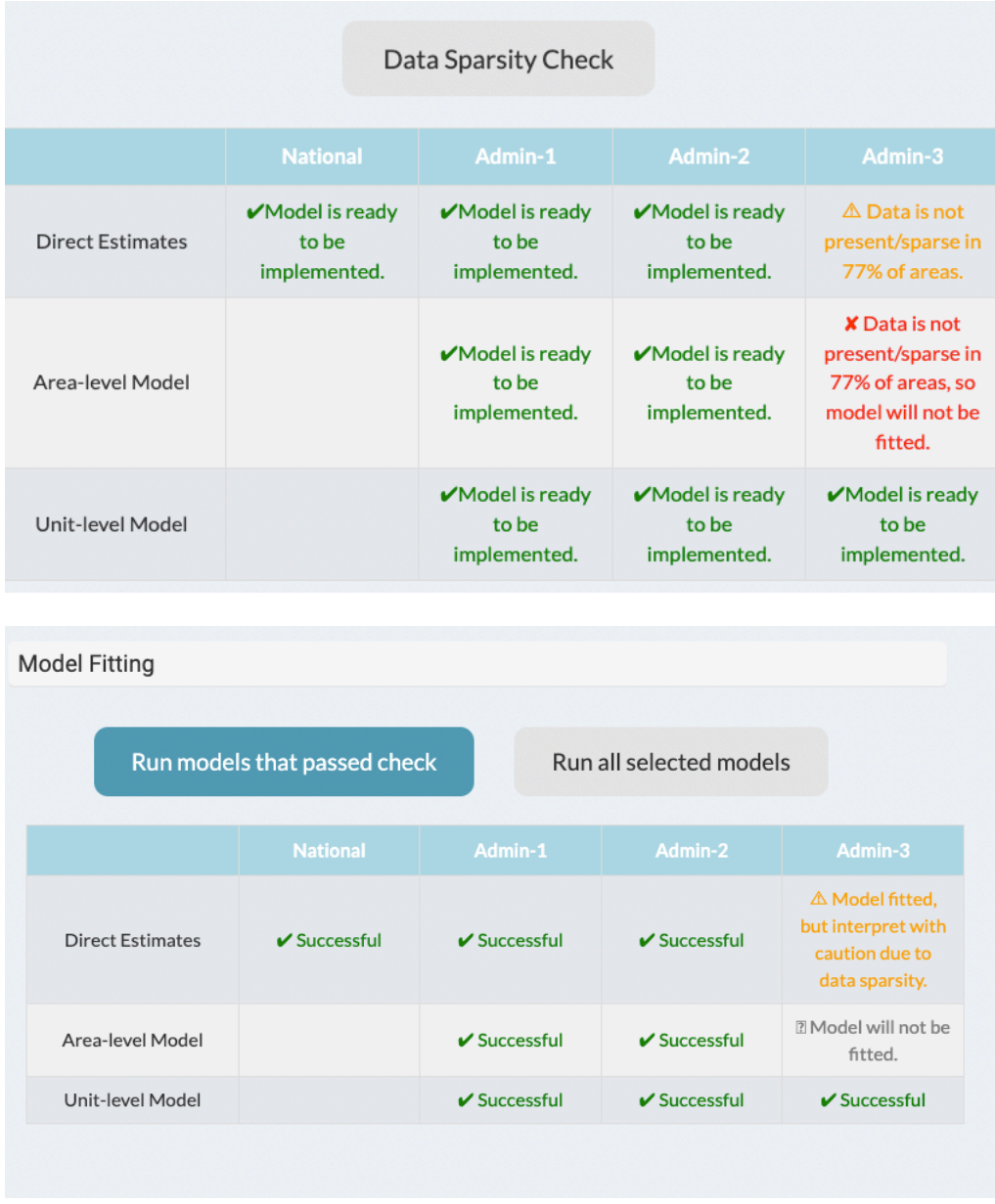

FIGURE 11: The shinyApp gives a warning that direct estimates not available in many Admin3 areas, and so this does not allow the fitting of an Admin3 area-level Fay-Herriot model.

Discussion

- Fay-Herriot area-level models are by far the most popular SAE models.
- Overshrinkage a worry: nested spatial models (which we discuss later) are a worthwhile alternative.
- Model checking still in infancy: we use cross-validation to predict weighted estimates.
- Machine learning approaches are very intoxicating, but very difficult to quantify uncertainty.

Discussion

- Benchmarking to known totals is less popular in LMICs because the population information may be inaccurate or out of date, see Okonek and Wakefield (2022) for references to the benchmarking literature.

- Variance modeling (which we discuss later) can be carried out to increase the utility of this method (Gao and Wakefield, 2023) – a method for variance modification is available in the `surveyPrev` package.

References

Banerjee, S., Carlin, B., and Gelfand, A. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press.

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

Breidt, J. and Opsomer, J. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*.

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.

Gao, P. A. and Wakefield, J. (2023). A spatial variance-smoothing area level model for small area estimation of demographic rates. *International Statistical Review*, **91**, 493–510.

Okonek, T. and Wakefield, J. (2022). A computationally efficient approach to fully Bayesian benchmarking. *Journal of Official Statistics*. Published online: May 27, 2024.

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

Ybarra, L. M. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919–931.