Bayesian SAE with an Emphasis on Low- and Middle-Income Countries: Lecture 1: Motivation and Overview

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Logistics

Course Outline

- 9:00–9:45: Lecture 1: Motivation and Overview
  - Context and examples
  - Weighted estimation
  - Overview of approaches to SAE
- 9:45–10:15: Lecture 2: Bayesian Smoothing
  - Bayes theorem
  - Summarization of the posterior
  - Bayesian smoothing models in time: random walk models
- 10:15–10:45: Lecture 3: Area-Level Models, Basics
  - Fay-Herriot model formulation
  - Prior specification
  - Computation
- 10:45–11:00: Software Demo
- 11:00–11:30 Coffee Break

Course Outline

- 11:30–12:15: Lecture 4: Unit-Level Models
  - Linear unit-level models
  - Adjustment for the design
  - Non-linear models and aggregation
- 12:15–12:45: Lecture 5: Further Topics
  - Nested models
  - Variance smoothing
  - Unmatched linking models
  - Model-based geostatistics
  - SAR spatial models
  - Space-time models
  - Spatial machine learning
- 13:15–13:30 Software Demo

Context and Overview

Motivation

- Small area estimation (SAE) entails estimating characteristics of interest for domains, often geographical areas, in which there may be few or no samples available – "small" refers to the number of samples in, and not to the geographical size of, the areas.

- SAE has a long history and a wide variety of methods have been suggested, from a bewildering range of philosophical standpoints.

- Application areas include: epidemiology, public and global health, agriculture, economics, education,...

- The classic text on SAE is Rao and Molina (2015).

- Background reading on the Bayesian models that will be described: Wakefield *et al.* (2019), Wakefield *et al.* (2020) and Dong and Wakefield (2021).

Health and Demographic Indicators

My own research focuses on health/demographic indicators in low- and middle-income countries (LMICs):

- Characterizing and understanding subnational variation is an important public health endeavor.
- For example, in the Sustainable Development Goals (SDGs), Goal 3.2 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".
- Many other indicators have SDG targets and there is an explicit ask for assessing geographic disparities.

Prevalence Mapping and Geostatistics

- Examination of proportions across space, is known as prevalence mapping – we may map continuously in space, or across discrete administrative areas – we focus on the latter.
- SAE methods provide one approach to performing prevalence mapping, for administrative areas.
- "The term geostatistics is a short-hand for the collection of statistical methods relevant to the analysis of geolocated data, in which the aim is to study geographical variation throughout a region of interest, but the available data are limited to observations from a finite number of sampled locations." (Diggle and Giorgi, 2019).
- Model-based geostatistics (MBG) provide another approach to performing prevalence mapping, over continuous space, though these continuous surfaces can be averaged for area-level inference.

Characterization of Methods and Approaches

Some important distinctions:

Direct	or	Indirect	Estimation
Unit-level	or	Area-level	Modeling
Linear	or	Non-Linear	Modeling
Non-spatial Mixed	or	Spatial Mixed	Modeling
No Auxiliary	or	Auxiliary	Data
Design-based	or	Model-based	Inference
Frequentist	or	Bayesian	Inference

In general, the lack of information in small samples is compensated for by:

- The use of covariates (auxiliary variables) in a regression model.
- Employing smoothing via mixed effects models, including spatial smoothing.

Overview of SAE

- Data: We consider the common situation in which the available data arise from surveys with a complex design.
- A Problem: If small sample sizes in some areas/time periods, there is high instability. In the limit, there may be no data...
- Auxiliary Data: On covariates to aid in modeling.
- Survey Sampling Methodology: Required for design and analysis.
- Shrinkage and Spatial Smoothing: To reduce instability, use the totality of data to smooth both locally and globally over space.
- Different Approaches to SAE: Both traditional and Bayesian methods that use spatial smoothing.
- Implementation: In `R` programming environment, using the `SUMMER`, `surveyPrev`, `surveyMICS`, `survey` packages and various web-based Apps, see `https://sae4health.stat.uw.edu/`
- Visualization: Maps of uncertainty, accompanied with uncertainty, produced using the GIS capabilities of `R`.

My SAE Background

My own interest in SAE:

- Started with work on Behavioral Risk Factor Surveillance System (BRFSS), with King County (in Washington state) government researchers.
- Moved to estimating subnational estimation of:
  - under-5 child mortality,
  - neonatal mortality,
  - vaccination,
  - HIV prevalence,
  - female educational attainment,...

  in LMICS, working with UNICEF, WHO, DHS and MICS.

The Team

Space Time Analysis Bayes (STAB) Lab: Orvalho Augusto, Adi Cohen, Tariq Cisse, Willow Crawford-Crudell, Qianyu Dong, Ameer Dharamshi, Geir-Arne Fuglstad, Peter Gao, Jessica Godwin, Jitong Jiang, Ziyu Jiang, Victoria Knutson, Zehang Richard Li, Alana McGovern, Josh Manandhar, Taylor Okonek, Albert Osom, Bumjun Park, Katie Paulson, Paige Tomer, Jon Wakefield, Yunhan Wu, Jieyi Xu, Joshua Yang, Zihang Yu.

DHS: Trevor Croft.

MICS: Nazim Gashi, Yadigar Coskun, Ivana Bjelic, Attila Hancioglu, Bo Pedersen.

UNICEF/UN IGME TAG: Lucia Hug, Liliana Carvajal, Patrick Gerland, Jon Pedersen, Leontine Alkema, Dave Sharrow, Bruno Masquelier, Monica Alexander, Danzhen You, Jimmy Ayalew, Charlotte Lie-Piang.

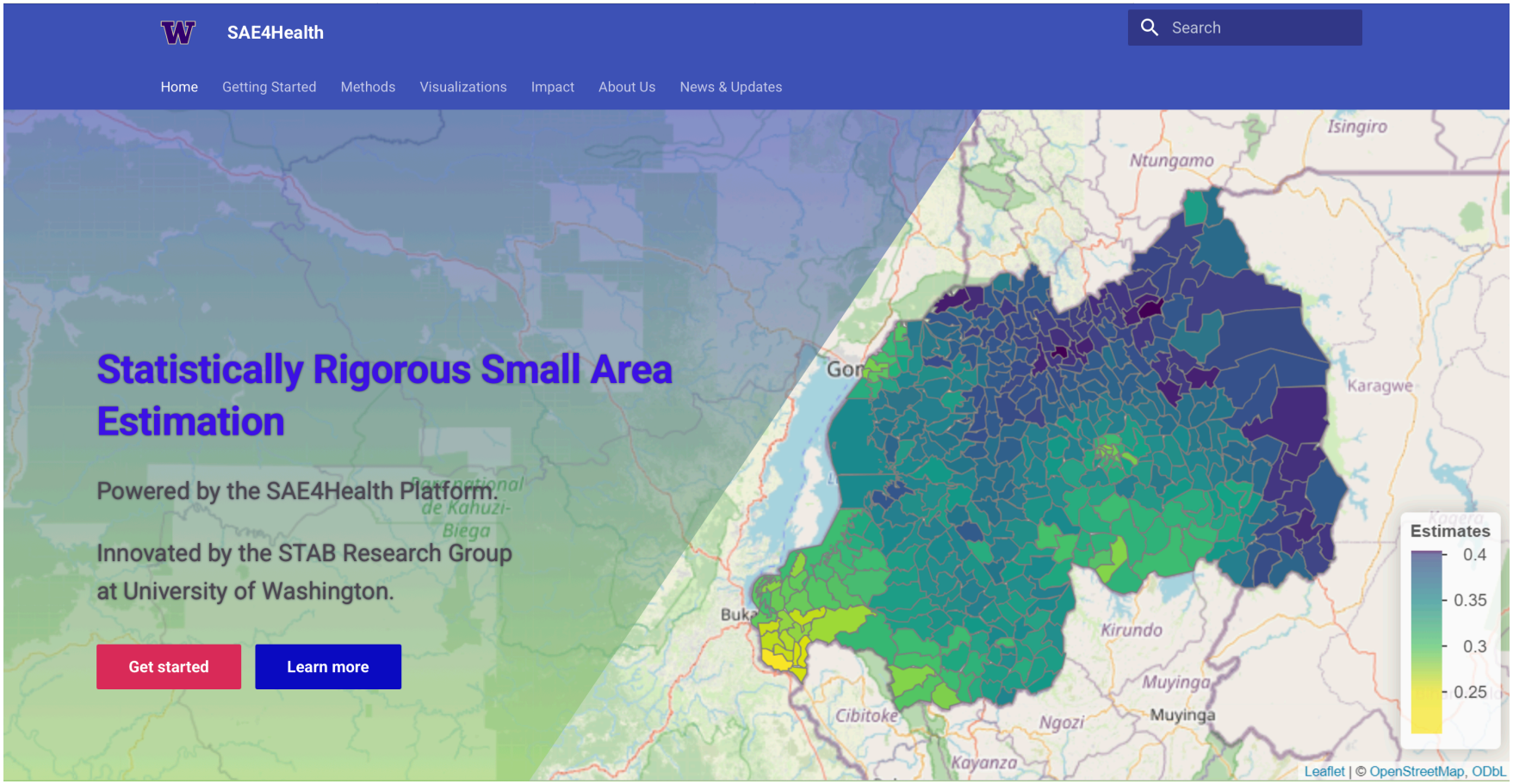
WHO: Charlton Callendar, Haidong Wang.

WorldPop: Andy Tatem, Edson Utazi.

Gates: Ash Shah, Nicole Danfakha.

Details on methods, videos, latest updates, and much more can be found at:

https://sae4health.stat.uw.edu/



Complex Survey Sampling and Motivating Examples

Demographic and Health Surveys (DHS)

- Motivation: In many developing world countries, vital registration is not carried out, so that births and deaths go unreported.
- Objective: To provide reliable estimates of demographic/health indicators at the (say) Admin1 or Admin2 level[1], at which policy interventions are often carried out.
- We will illustrate using data from Demographic and Health Surveys (DHS).
- DHS Program: Typically stratified cluster sampling to collect information on population, health, HIV and nutrition; more than 450 surveys carried out in over 90 countries, beginning in 1984.
- The Problem: Data are sparse, at the Admin2 level in particular.
- SAE: Leverage spatial similarity and covariates to construct a Bayesian smoothing model.

[1] Admin0 = country level boundaries, Admin1 = first level administrative boundaries (states in US), Admin 2 = second level administrative boundaries (counties in US)

The Complex Survey Design of the DHS

- The DHS uses: **Stratified two-stage unequal probability cluster sampling**.
- Strata consist of urban/rural crossed by geographic administrative areas.
- In each strata, enumeration areas (EAs) (also referred to as clusters) are selected with probability proportional to size sampling (number of households being the size variable) using a sampling frame developed from the most recent census (this is the unequal probability sampling part).
- In each of the clusters, households are selected.
- Within each household, women (and a smaller fraction of men) between the ages of 15 and 49 are interviewed.
- Based on these steps we can calculate π_{ik}, the probability that individual k in area i is selected into the study.

The Rwandan 2019 DHS

- Sampling frame: 2012 census; stratification: urban/rural crossed with 30 districts.
- 500 sampled clusters: 112 were in urban areas and 388 were in rural areas.
- 26 households were selected from each selected EA, to give 13,000 in total.

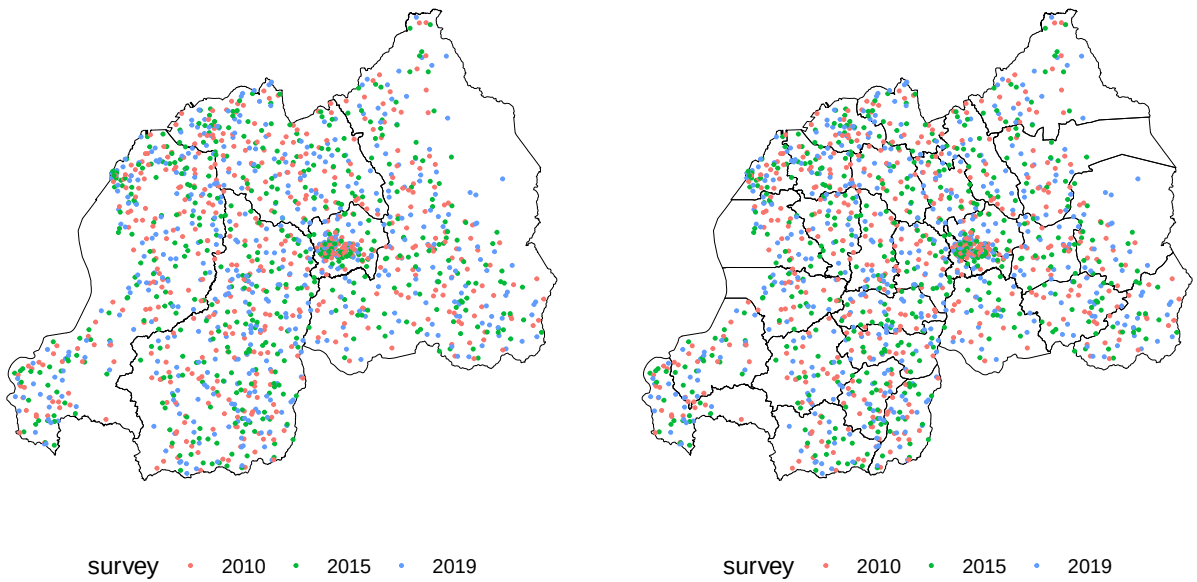


Figure 1: Rwandan cluster locations: 5 provinces (Admin1) and 30 districts (Admin2).

Acknowledging the Design: Stratification

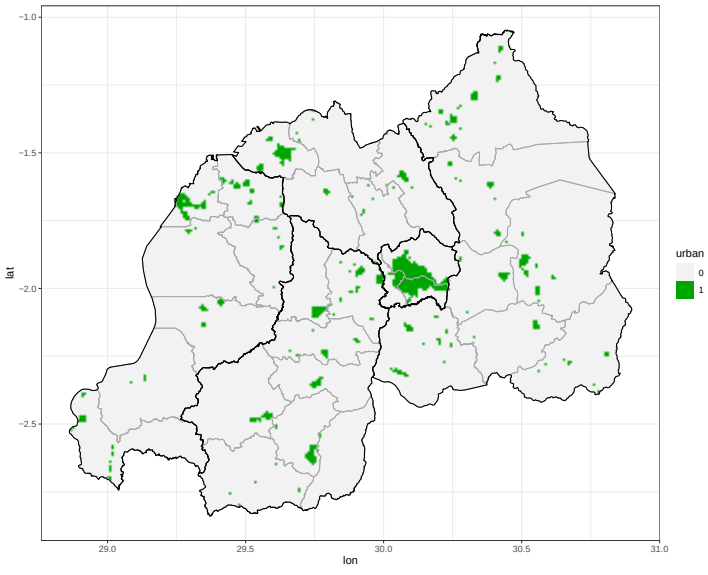


Figure 2: In the DHS in Rwanda, stratification is based on provinces and urban/rural.

- Suppose we are interested in the proportion of women aged 20–29 who complete secondary education.
- The prevalence of secondary education is higher in urban areas than in rural.
- If we oversample urban areas but ignore this when we analyze the data we will overestimate the prevalence of women who complete secondary education, i.e., we will introduce bias.
- Taking into account the stratification also reduces the variance of the estimator.
- In the design-based approach to inference, the stratification is accounted for via design weights.
- In the model-based approach to inference, the stratification is accounted for in the mean model.

Acknowledging the Design: Cluster Sampling

- The DHS also employs cluster sampling, in which multiple units (individuals) within the same cluster are interviewed.
- Units within the same cluster tend to be more similar than units in different clusters, which reduces the information content of the clustered sample, relative to independently sampled units.
- In the design-based approach to inference, the clustering is accounted for in the variance calculation that is carried out.
- In the model-based approach to inference, the clustering is accounted for by including a cluster-specific term in the model.
- An excellent text on survey sampling is Lohr (2010), my own take is summarized in Skinner and Wakefield (2017).

Aim: Inference for U5MR over Counties and Years

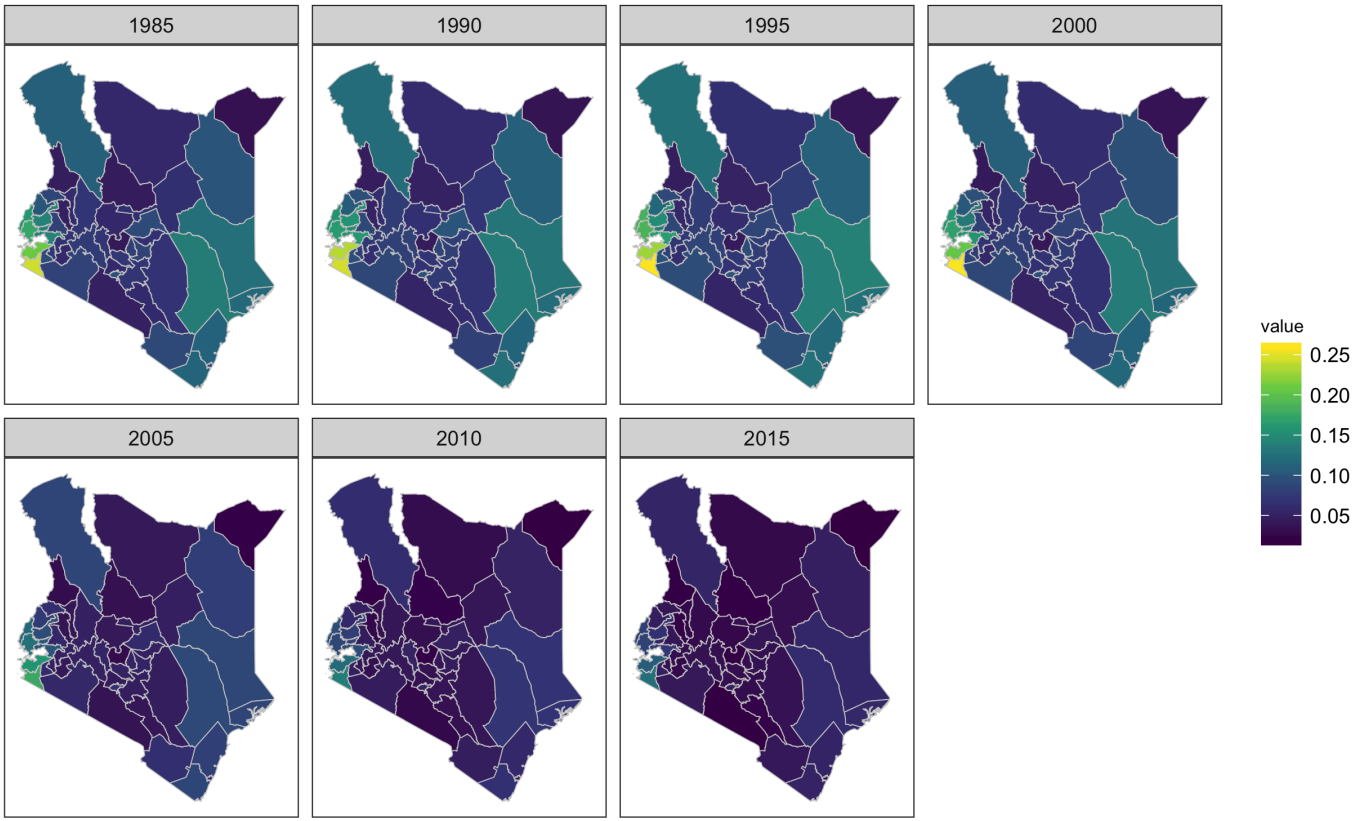
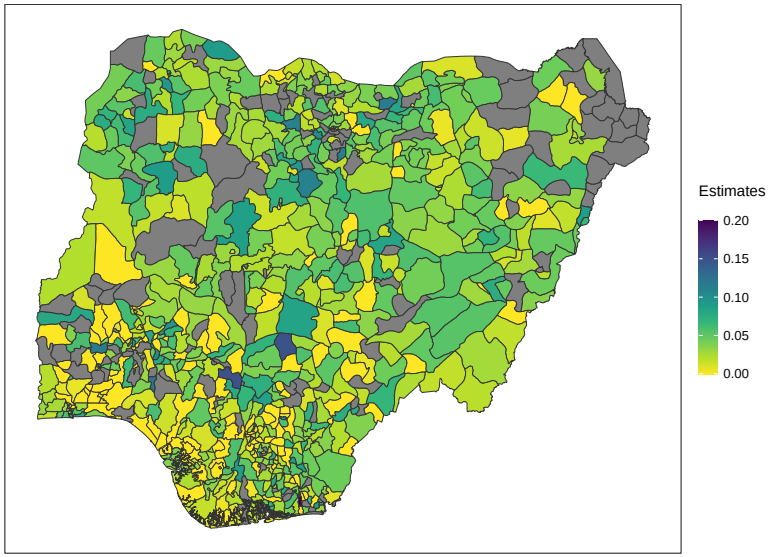
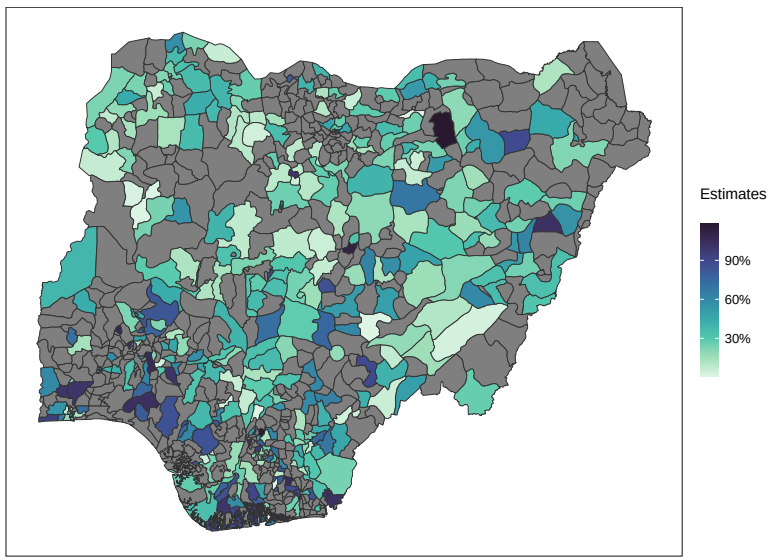


Figure 3: SAE estimates of under-5 mortality risk, across time, and Kenyan counties. These estimates were obtained using the SUMMER package.

Sparsity at Admin2

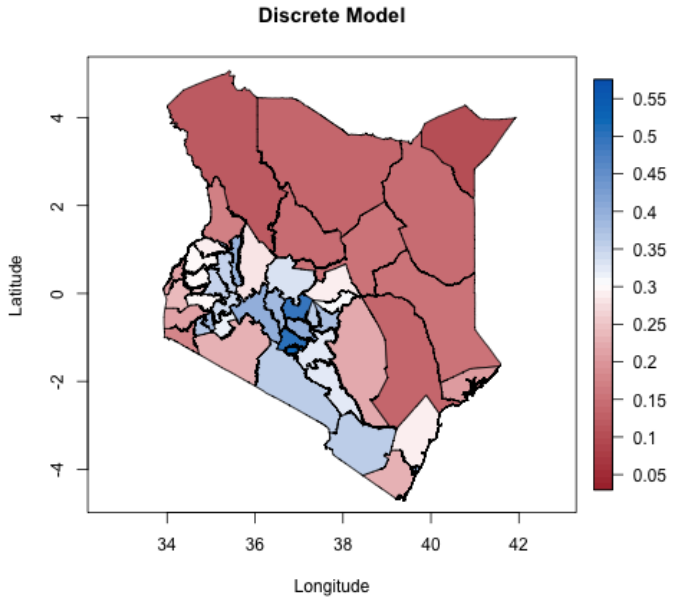
- As another DHS example, we consider neonatal mortality prevalence in Nigeria, from the 2018 Nigerian DHS.
- In Nigeria, the Admin2 areas correspond to Local Government Areas (LGAs) and there are 774 in total – with such a large number there are many LGAs with little/no data.
- In the mean estimate figure at the top, areas in gray have no clusters.
- In the in the CV figure at the bottom, areas in gray have no clusters, or configurations of data that do not allow a variance estimate to be calculated.



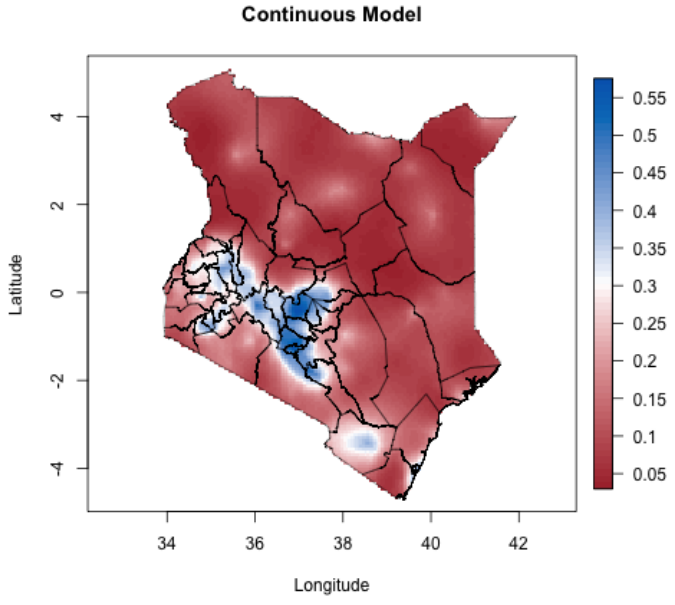


Two Approaches to Spatial Smoothing

- Model at the area level using a discrete spatial model. These are the SAE models that are implemented in the sae4health software.



- Model at the point level using a continuous spatial model. Model-based geostatistics is a popular approach.



Weighted Estimation

Overview of SAE Methods

There are a multitude of methods for small area estimation, but a big distinction is between:

- Area-level models.
- Unit-level models.

We begin with a discussion of the weighted (direct) estimate before looking at each of these model types in detail.

Two Inferential Settings

In a non-complex survey sampling setting:

- The data are (often implicitly) assumed to arise from simple random sample (SRS).
- The target of interest is a parameter in a hypothetical infinite population.

In a survey sampling setting:

- The data arise from a complex design, such as stratified two-stage unequal probability cluster sampling.
- The target of interest is a finite population characteristic.

In this second setting, statistical methods are required which acknowledge these two aspects.

Inference in SRS Setting

- When faced with 0/1 responses from a SRS, the usual approach is to assume an underlying risk parameter in area i is r_i.
- The risk r_i is the proportion of events of interest (e.g., neonatal deaths), in a hypothetical infinite "superpopulation" residing in area i.
- Let $Y_{ik} = 0/1$ be the indicator for the event k in area i with $k = 1, \ldots, n_i$, so that n_i is the number of samples in area i.
- In a model-based approach, it is often assumed that the total number of 1's, $Y_i = \sum_{k=1}^{n_i} Y_{ik}$, is distributed as Binomial(n_i, r_i).
- Under SRS in area i, the estimator is the simple mean

$$\widehat{r}_i^{\,\mathrm{SRS}} = \frac{Y_i}{n_i} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}, \qquad \text{Std Err}(\widehat{r}_i^{\,\mathrm{SRS}}) = \sqrt{\frac{\widehat{r}_i^{\,\mathrm{SRS}}(1 - \widehat{r}_i^{\,\mathrm{SRS}})}{n_i}}.$$

Weighted Estimates

- **Prevalence** p_i is the empirical proportion in a finite population with a certain condition.
- In area i, if N_i is the population size and T_i is the number of events, then the prevalence is

$$p_i = \frac{T_i}{N_i}.$$

- The distinction between risk and prevalence is important in the survey sampling literature – the target in SAE is generally the prevalence.
- If we were to take a complete census in area i, i.e., sample all N_i individuals, then we would know the prevalence, and no statistics required!
- But we would not know the risk, as this is a summary from a hypothetical infinite population.

Weighted Estimates

- Let $Y_{ik} = 0/1$ again be the indicator for the event k in area i, with $w_{ik} = 1/\pi_{ik}$ being the design weight, that accounts for the (probabilistic) sample selection.

- w_{ik} can be thought of as the number of people represented by person k.

- For area i, the weighted estimator is

$$\widehat{p}_i^{\,w} \quad = \quad \frac{\widehat{T}_i}{\widehat{N}_i} \quad = \quad \frac{\sum_{k \in S_i} w_{ik} Y_{ik}}{\sum_{k \in S_i} w_{ik}}. \tag{1}$$

  where S_i is the set of sampled births.

- Known as a **direct** estimate since based on data from the area in question only.

- Standard errors can be obtained using design-based theory.

The weighted mean is the direct estimator that is the first choice for producing subnational estimates

Direct Estimates

Advantages:

- Estimates and uncertainty measures account for design via weighting.
- Inference is based on minimal assumptions, since there is no explicit model for the data.
- Often produces reasonable inference for Admin1 areas (unless outcome is very rare). Sometimes OK for Admin2 also.

Disadvantages:

- When data are sparse in an area, then estimator will have uncertainty that is high, or not possible to estimate (variance formula breaks).
- For example, for the SRS estimator in a model-based framework, the standard error is $\sqrt{\widehat{r}_i^{\text{SRS}}(1-\widehat{r}_i^{\text{SRS}})/n_i}$ which is 0 if $\widehat{r}_i^{\text{SRS}}$ is equal to 0 or 1.
- If no data in an area, then no estimate available.

Summary Measures

Summarizing Information

- A map of the prevalence estimates is visually appealing, but in isolation is impossible to interpret.
- To make interpret and take action, it is vital to understand the inherent uncertainty in the estimate, with possibilities:
  - The standard error – useful if unceratinty distribution is symmetric.
  - An uncertainty interval.
  - The width of an uncertainty interval.
  - The coefficient of variation.
  - An examination of ridgeplots.
  - Calculating exceedence probabilities.
- These measures are of generally interest, not just for weighted estimates.

Uncertainty Measures

- The variance of $\widehat{p}_i^{\mathrm{w}}$ can be obtained using survey sampling techniques which are tuned to the specific type of survey – call this $\widehat{V}_i$.
- The standard error is simply $\sqrt{\widehat{V}_i}$.
- If the sample size n_i is sufficiently large, an accurate 95% confidence interval is

$$\left[\widehat{p}_i^{\mathrm{w}} - 1.96 \times \sqrt{\widehat{V}_i} \, , \, \widehat{p}_i^{\mathrm{w}} + 1.96 \times \sqrt{\widehat{V}_i} \right]$$

Uncertainty Measures

- The width of the confidence interval (CI) is

$$2 \times 1.96 \times \sqrt{\widehat{V}_i} \approx 4 \times \sqrt{\widehat{V}_i}$$

- Notice that the width of the interval does not depend directly on the point estimate, which may be deceptive since, for example, an interval of width 0.05 can mean very different things:
  - $\widehat{p}_i^{\mathrm{W}} = 0.50$, interval: (0.475,0.525), versus
  - $\widehat{p}_i^{\mathrm{W}} = 0.05$, interval: (0.025,0.075).

  When prevalence is close to 0 or 1, we might prefer an alternative measure.

Uncertainty Measures

- The coefficient of variation (CV) is popular for SAE, particularly when the prevalence is close to zero.
- The CV is often expressed as a percentage and is defined as

$$\text{CV} = 100 \times \frac{\text{Standard Error}}{\text{Point Estimate}} = 100 \times \frac{\sqrt{V_i}}{\widehat{p}_i^{\text{W}}}.$$

- Examples:
  - if the point estimate is $\widehat{p}_i^{\text{W}} = 0.05$ and the standard error is 0.025, then we have a CV of 50% which would be viewed as large uncertainty.
  - By contrast, if the point estimate is $\widehat{p}_i^{\text{W}} = 0.25$ and the standard error is again 0.025, then we have a CV of 10% which may be viewed as reasonable uncertainty.

Uncertainty Measures

- Some agencies have criteria for reporting, for example, Statistics Canada have guidelines:
  - reliable if coefficient of variation $< 17\%$,
  - view with caution if in the range $(17\%, 33\%)$ and
  - conclude unreliable if $> 33\%$.
- Since the CV is an explicit function of $\widehat{p}_i^{\mathrm{w}}$ it may be preferable as a summary to the standard error or width of a CI – I find it most useful when $\widehat{p}_i^{\mathrm{w}}$ is small.
- Notice the CI depends on a level (95%) while the CV does not.

Ridgeplots

- Ridgeplots, as we use them in SAE, show the complete range of uncertainty for a collection of areas.
- We may examine, for example,
  - All Admin1 areas stacked on top of each other
  - Admin2 areas within a particular Admin1
- The ordering may be via the point estimate (alphabetical in Figure 4).

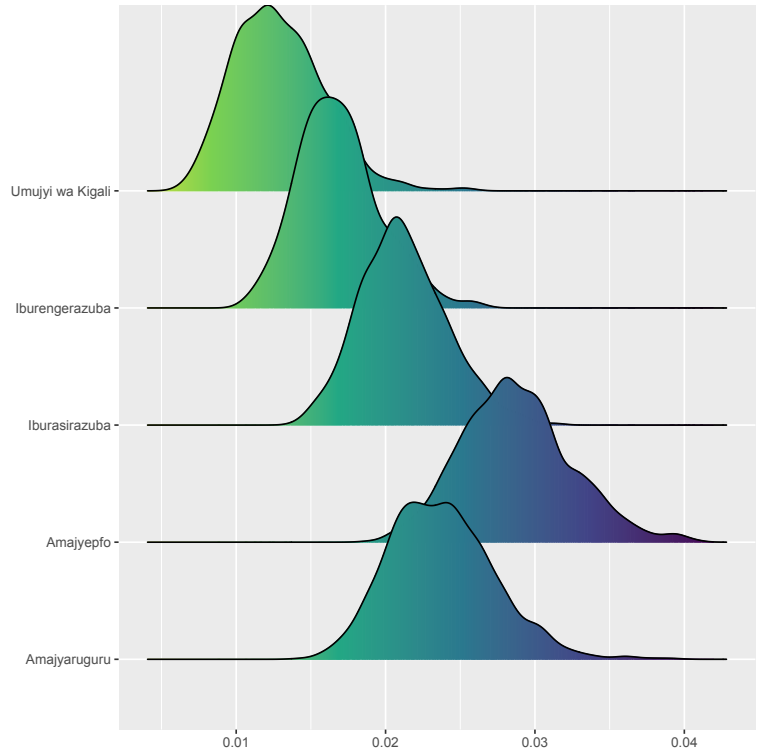


Figure 4: Ridgeplots for weighted estimates of NMR prevalence for provinces in Rwanda.

SAE for Neonatal Mortality in Rwanda

Subnational Neonatal Mortality in Rwanda

- Initial Aim: Estimate prevalence of neonatal deaths in the 10 years before the survey, at Admin1.
- Results from the 2019 Rwanda DHS, analyzed using the `DHS R shinyApp`.
- The national weighted estimate is:
  - 21 deaths per 1000 live births, with a standard error of 1.4
  - 95% confidence interval of (18,24) deaths per 1000 live births (a width of 6 deaths)
  - Coefficient of variation is $100 \times 1.4/21 = 6.8\%$

Rwanda 2019 DHS

- DHS use stratified two-stage cluster sampling. The strata consist of urban/rural crossed with 30 Admin1 areas.
- In each strata, a total of 500 enumeration areas (EAs) are selected with probability proportional to size using a sampling frame developed from the most recent census.
- In each of the clusters, 26 households are selected. Within each household, women between the ages of 15 and 49 are interviewed.

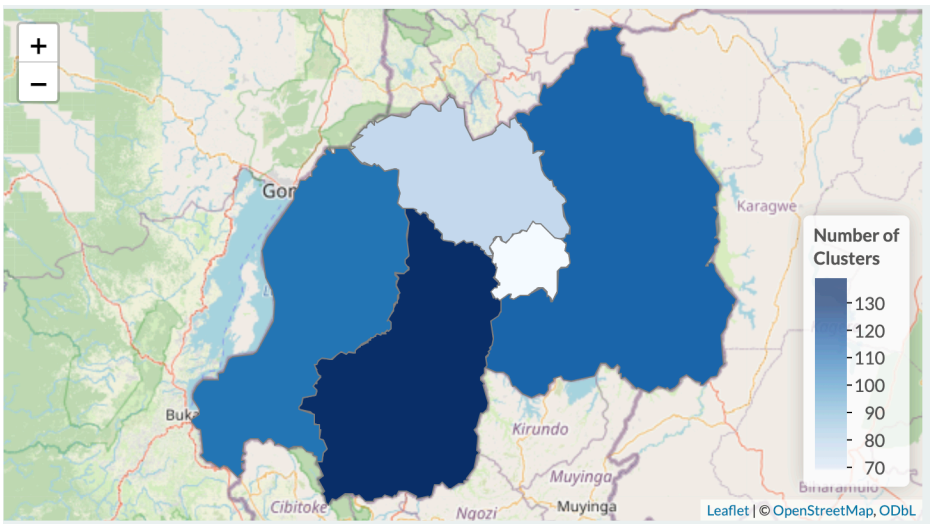


Figure 5: Number of sampled clusters, by Admin1, from Rwanda 2019 DHS.

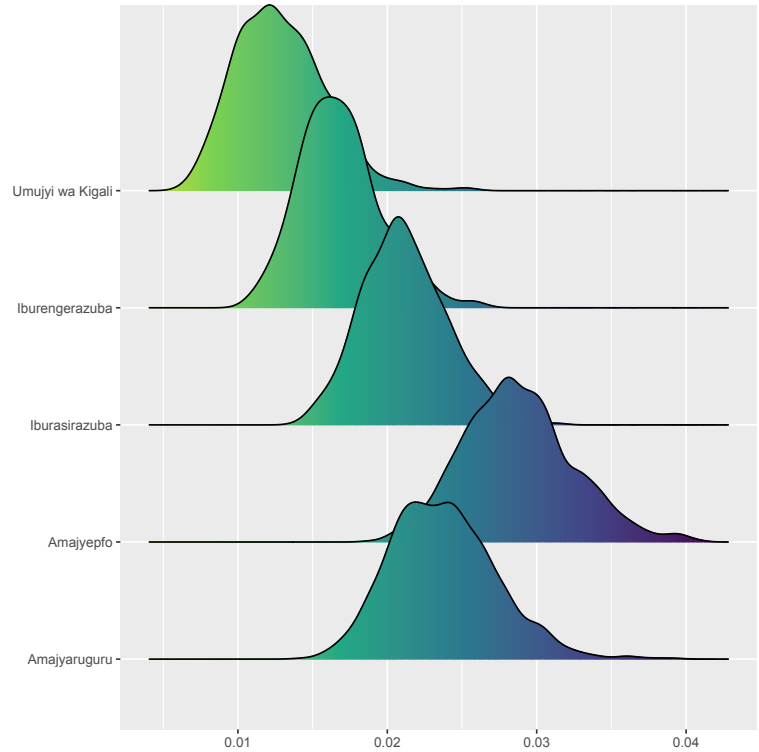


FIGURE 6: Ridgeplots for weighted estimates of NMR prevalence for provinces in Rwanda.

- Look at the overlap! The magnitude of uncertainty is clear, even at Admin1.
- This is sobering: we often have to deal with large levels of uncertainty.

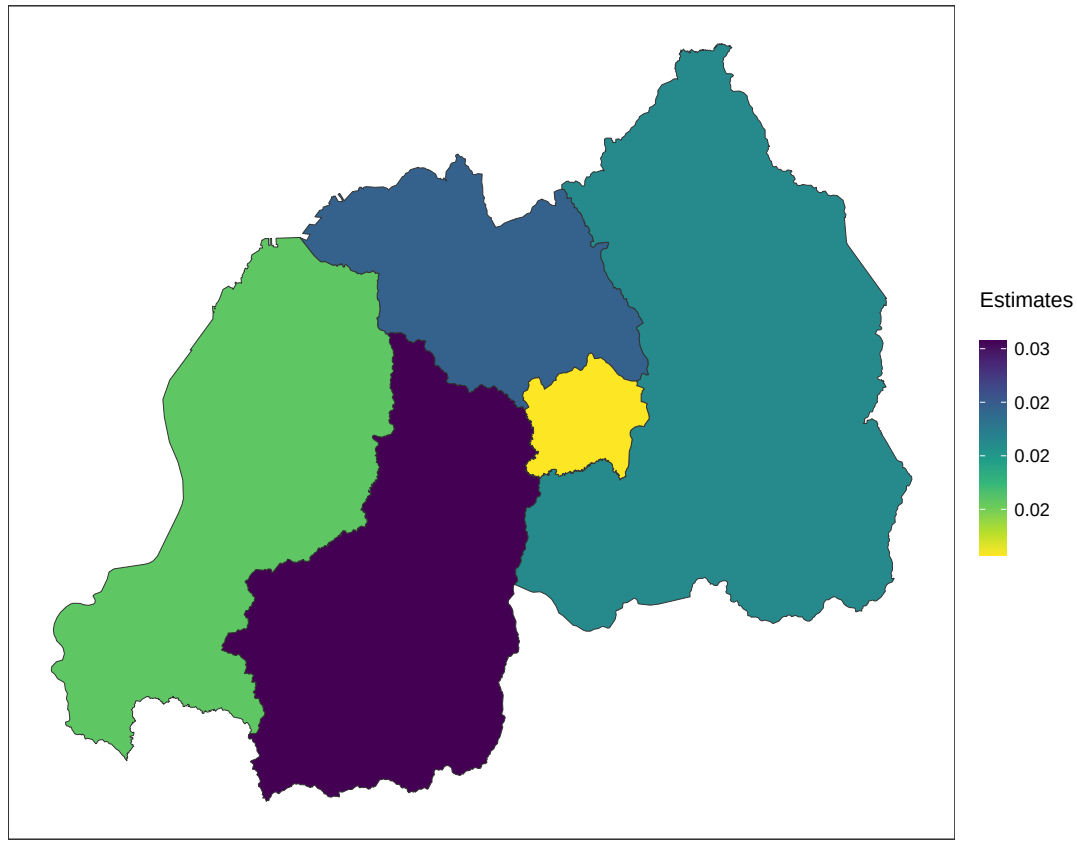


FIGURE 7: Weighted estimates of NMR prevalence for provinces in Rwanda.

- Range of estimates is 0.013 to 0.029 – large between-area variation.
- But uncertainty in each estimate is needed for interpretatation.

FIGURE 8: Rwanda Admin1 (province) map.

RW2019_CM_ECMR_C_NNR_Admin_1_Direct

Region_Name	Mean	Median	Standard_Error	Lower_CI	Upper_CI	Width_95_CI	Coefficient_of_Variation
Amajyaruguru	0.024	0.024	0.0035	0.018	0.032	0.014	0.149
Amajyepfo	0.029	0.029	0.0036	0.022	0.037	0.014	0.126
Iburasirazuba	0.021	0.021	0.0029	0.016	0.027	0.011	0.138
Iburengerazuba	0.017	0.017	0.0027	0.012	0.023	0.011	0.162
Umujyi wa Kigali	0.013	0.013	0.0030	0.008	0.020	0.012	0.241

FIGURE 9: Summaries from weighted estimates at Admin1 level for Rwanda.

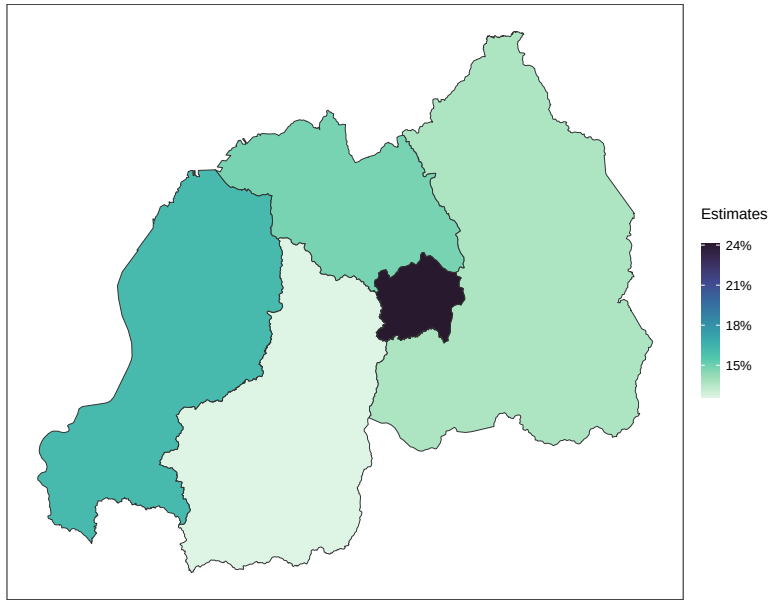


FIGURE 10: Coefficients of variation, for NMR at Admin1 level in Rwanda. Range is 13% to 24%.

- Greatest CV is in Kigali, which has the lowest estimate, which makes sense (standard errors are very similar across areas).

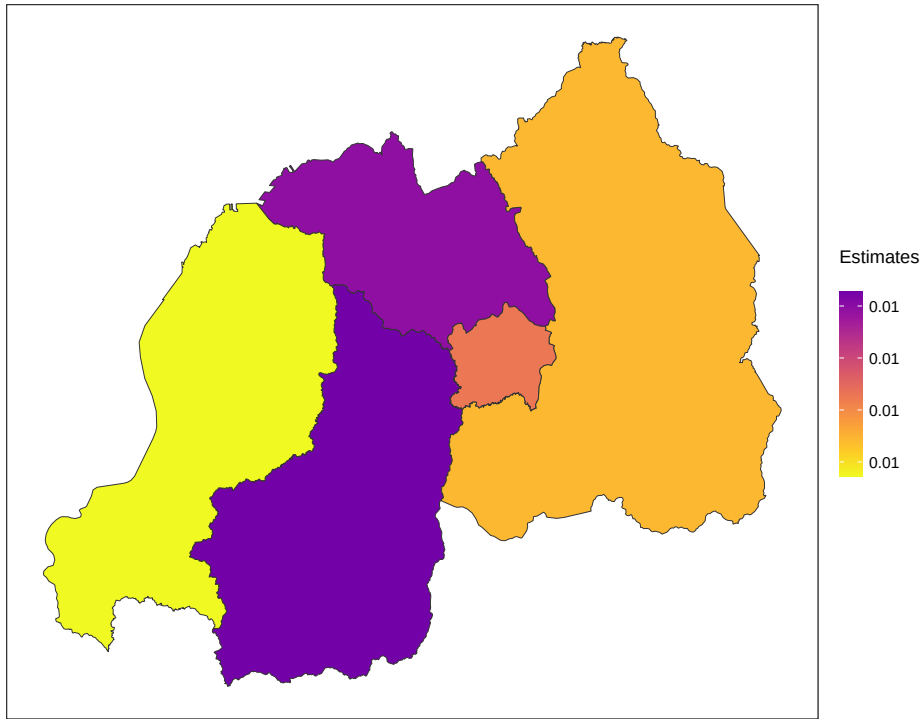


FIGURE 11: Width of 95% confidence interval, for NMR at Admin1 level in Rwanda.

- Confidence interval widths are not so different across areas.
- The CV map is more useful to think about precision, as it tells us the relative uncertainty.

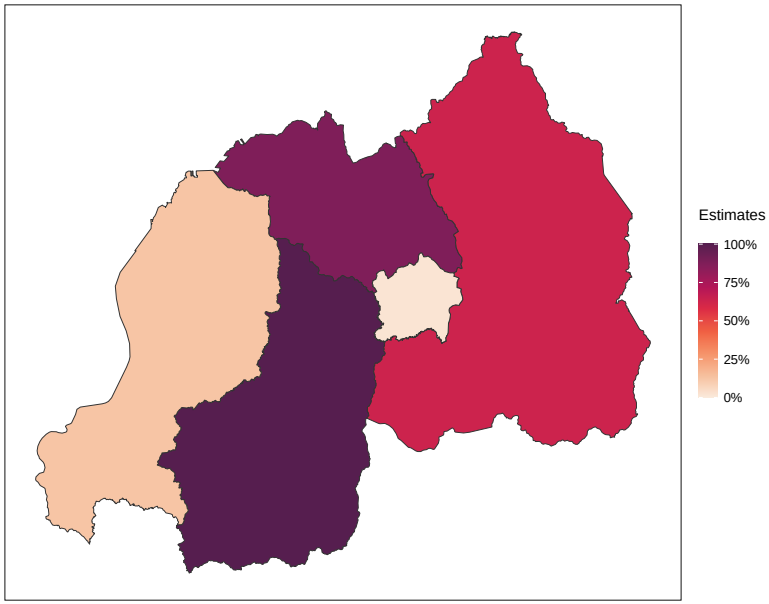


FIGURE 12: Probabilities of each area exceeding the national estimate of 21 deaths per thousand.

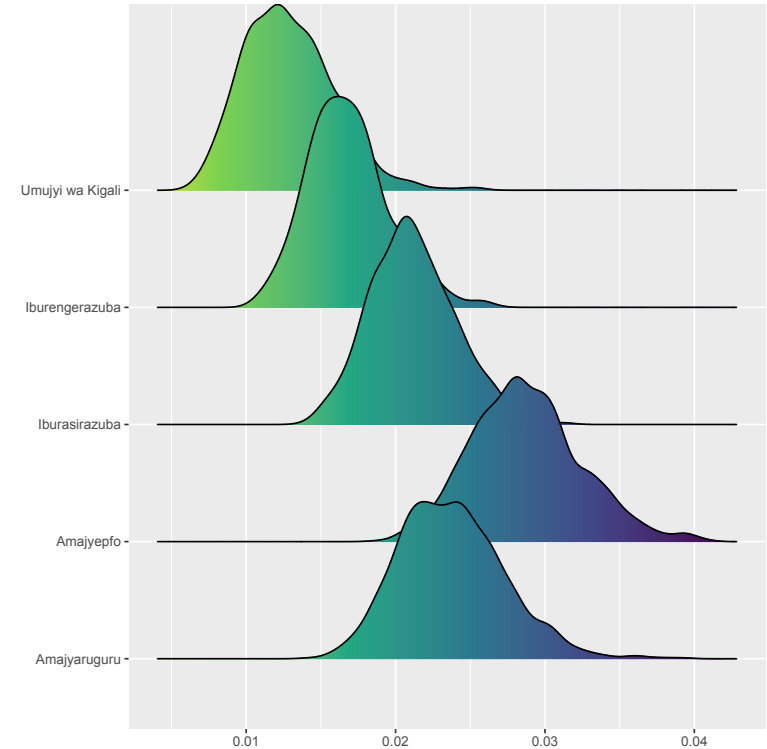
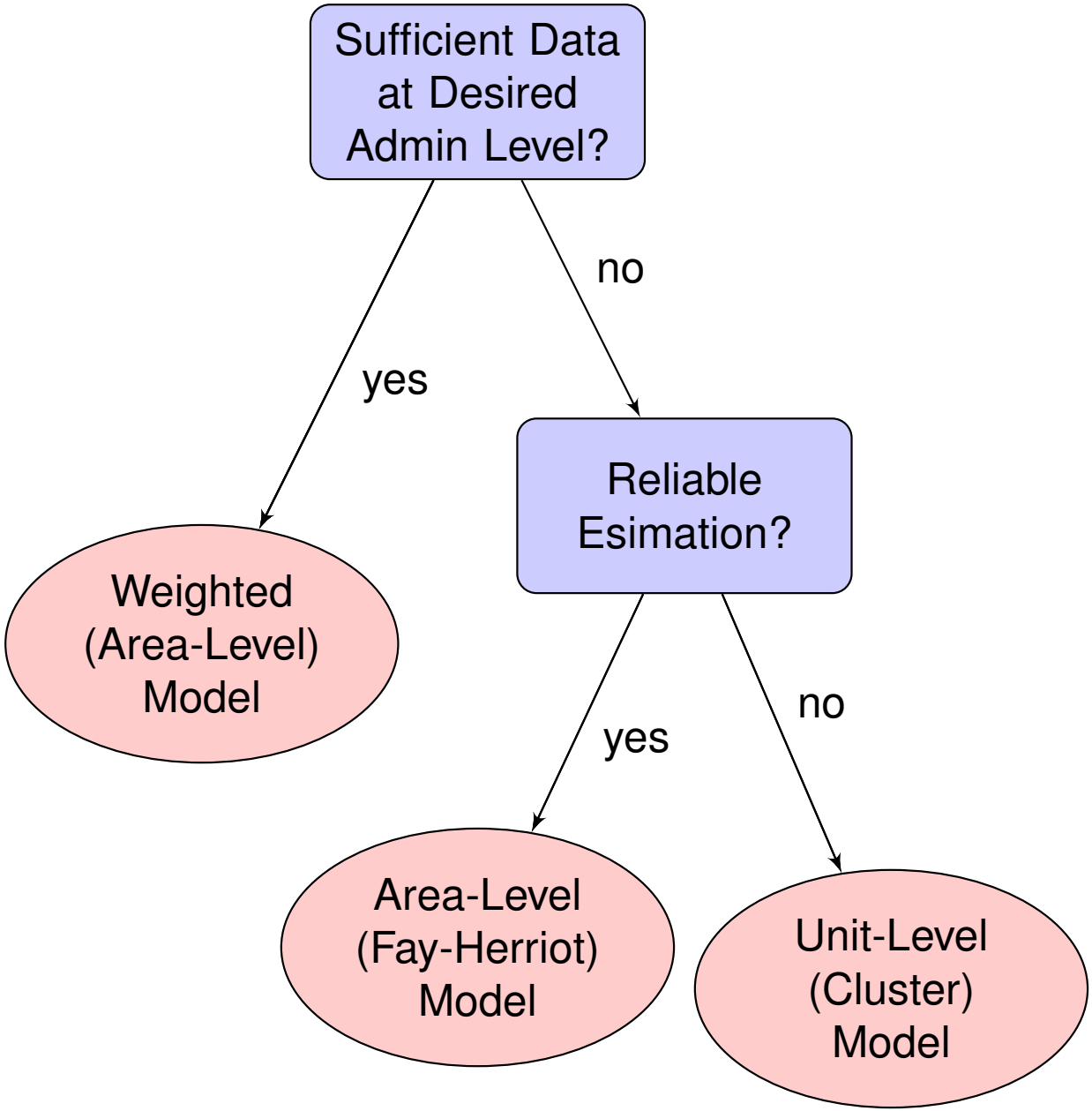


FIGURE 13: Ridgeplots for weighted estimates of NMR prevalence for provinces in Rwanda.

- Probabilities of exceeding 0.21: Umujyi wa Kigali 1.6%, Iburengerazuba 12.6%, Iburasirazuba 60.2%, Amajyepfo 99.8%, Amajyaruguru 88.6%.
- Huge variation across the 5 areas.

Subnational Modeling Big Picture



Discussion

Discussion

- SAE generally works with survey data, while disease mapping is concerned with data which are nominally a full enumeration.
- A key point is that methods must acknowledge the way the data were collected, i.e., the design.
- The way survey sampling estimation techniques have developed, via design-based inference, is quite different to standard frequentist model-based inference.
- Weighted estimates are great, when they (and a reliable standard error) are available.

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.

Lohr, S. (2010). *Sampling: Design and Analysis, Second Edition*. Brooks/Cole Cengage Learning, Boston.

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

Skinner, C. and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, **32**, 165–175.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. (2019). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, **28**, 2614–2634.

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