Small Area Estimation based on Household Survey Data

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington, Seattle

Outline

Motivation for Subnational Estimation

Complex Survey Sampling

Weighted Estimation

Small Area Estimation (SAE): Area-Level Models

- Methods
- Subnational NMR Estimation in Nigeria: Area-Level Models

Small Area Estimation (SAE): Unit-Level Models

- Methods
- Subnational NMR Estimation in Nigeria: Unit-Level Models

Software and Country Engagement

Models for Urban/Rural Stratification

Summary and Discussion

Motivation for Subnational Estimation

Small Area Estimation

- For purposes of programming and interventions, and to monitor progress toward targets such as the SDGs, subnational estimates of indicators of interest are desirable.
- Many indicators of interest:
  - **Simple prevalence:** NMR, vaccination coverage, stunting,...
  - **More complex variables:** U5MR, fertility, maternal mortality.
- For many health and demographic indicators there exists considerable within-country variation.
- The endeavor of small area estimation (SAE) is the task of estimating a variable of interest within geographic areas, based on potentially sparse data (this is the "small").
- Notation for nested geographic hierarchy: Admin0 is national, Admin1 is one below national, and Admin2 is two below national.

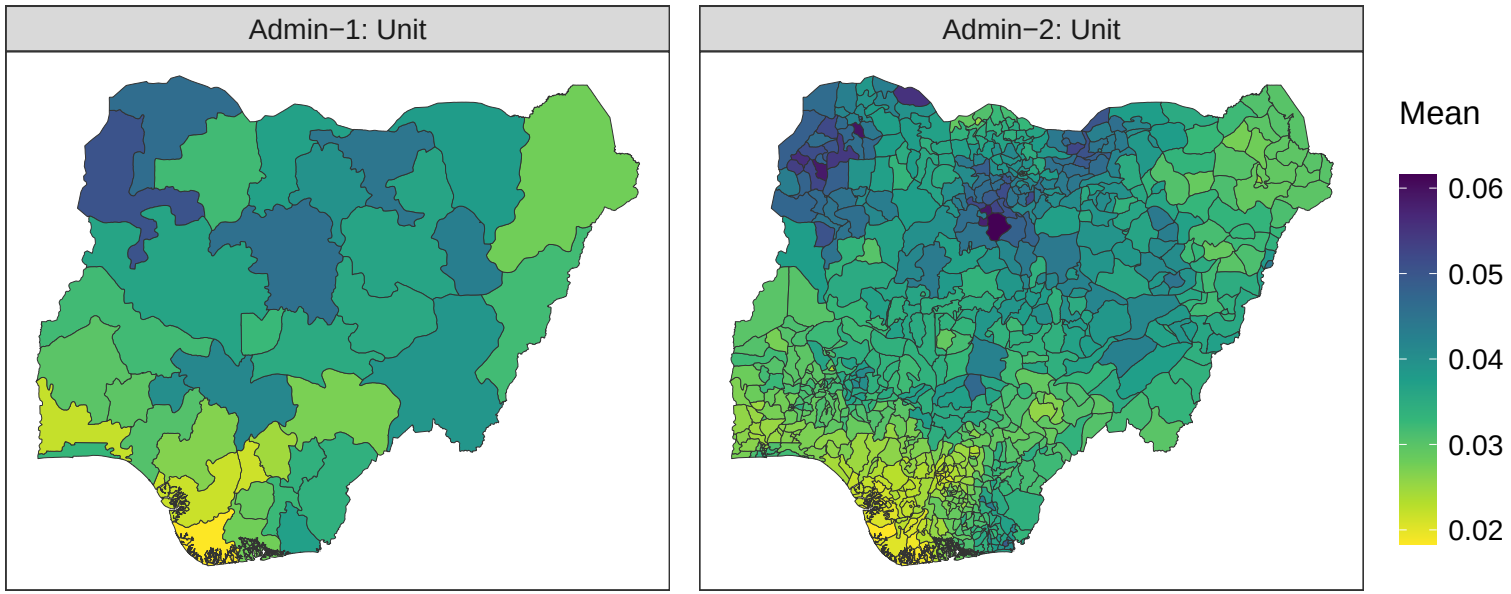


Figure 1: Estimated NMR in 10 years preceding the 2018 DHS across 37 Admin1 (states) and 775 Admin2 areas (LGAs) in Nigeria.

Small Area Estimation (SAE) in LMICs

- In LMICs, household surveys are often the most reliable data source.
- But not powered to produce weighted estimates at fine spatial scales.
- To overcome this data sparsity, SAE smoothing models in space (and time) can be used:
  - Area-level models: Relatively easy to implement and not too heavy on assumptions – often work for Admin1 and sometimes Admin2.
  - Unit-level models: Much trickier to implement as they must acknowledge the design, and more assumptions leant on – but far more powerful.
- Rao and Molina (2015) is the classic text on SAE.

A key point is that models must acknowledge the survey design – there are two fundamentally different approaches to analysis, design-based inference and model-based inference

Complex Survey Sampling

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 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.

Nigerian DHS Data

In the 2018 DHS:

- Sampling frame: 2006 with 664,999 census EAs (clusters); stratification: urban/rural crossed with 37 states.
- 1389 sampled clusters.
- Approximately 42,000 households were selected in total, 30 from each EA.

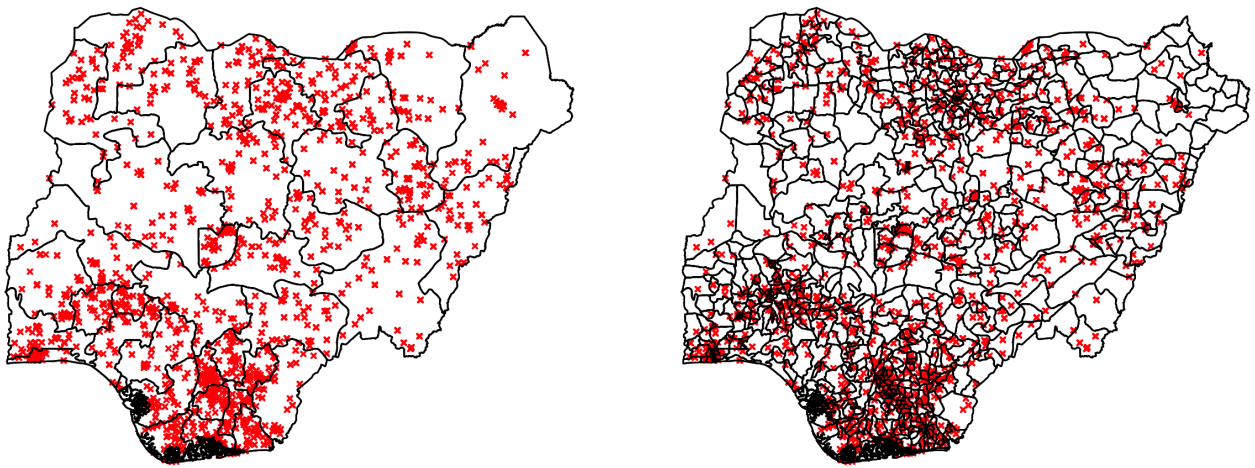


Figure 2: Maps of Nigeria showing the 37 Admin1 areas (left), and the 775 Admin2 areas (right). The crosses show the locations of the 1389 sampled clusters.

- The 2021 Nigerian MICS has the same design.

Nigeria 2021 DHS: Abundant Sampling at State Level

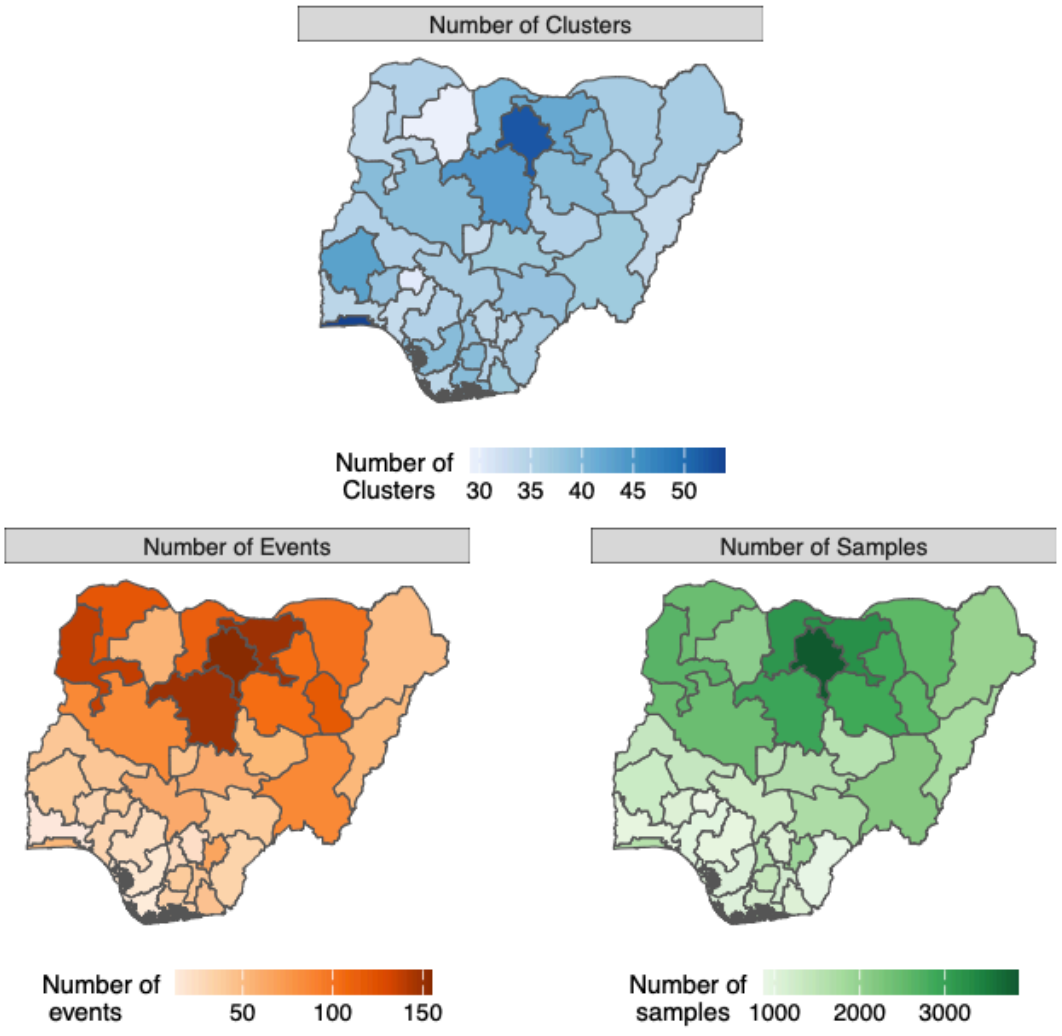


Figure 3: Cluster summary statistics, by state, from Nigeria 2018 DHS (NMR).

Nigeria 2021 DHS: Sparse Sampling at LGA Level

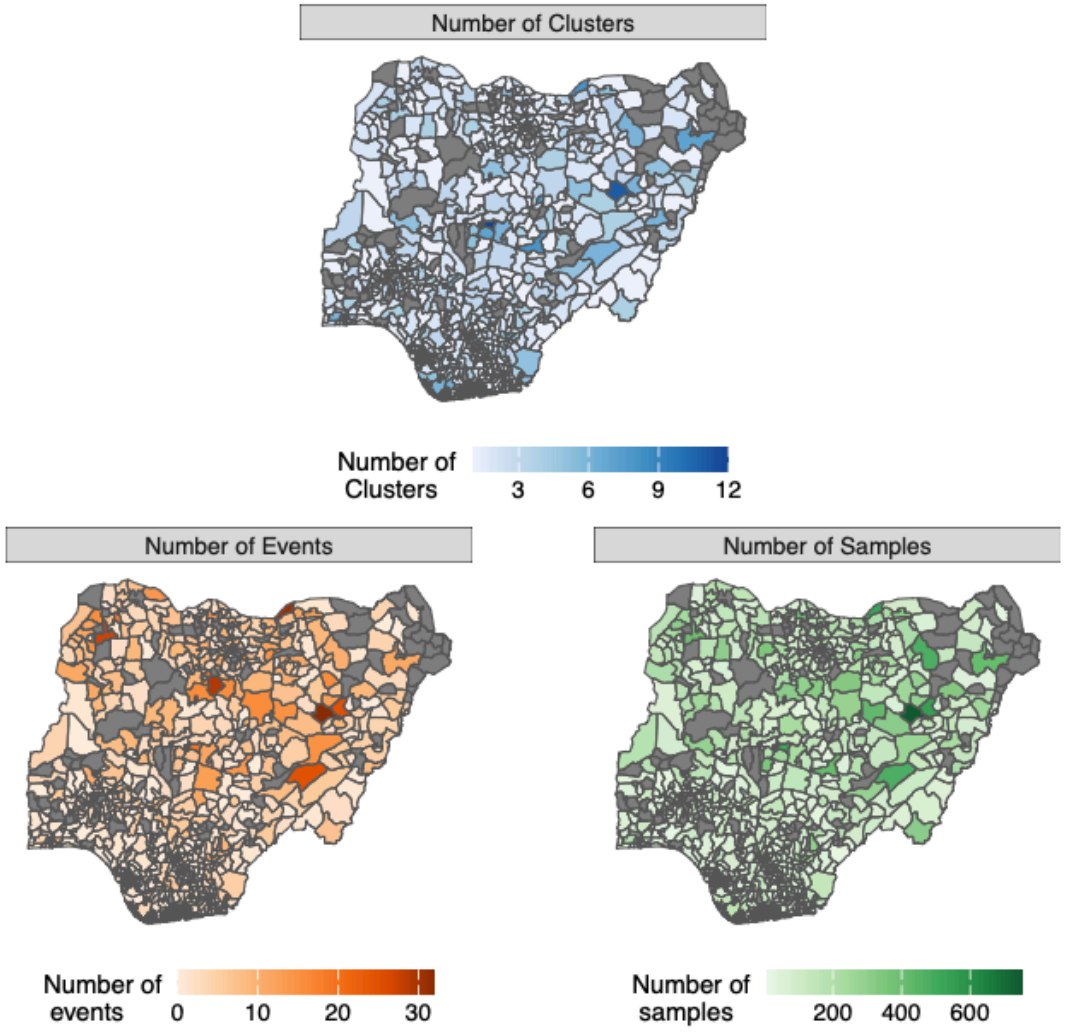


Figure 4: Cluster summary statistics, by LGA, from Nigeria 2018 DHS (NMR).

Weighted Estimation

Weighted Estimates

- In area i, if N_i is the population size and T_i is the number of events, then the prevalence is

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

- 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

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

  where S_i is the set of sampled births.

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

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 the 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 the data are sparse in an area, then the estimator will have uncertainty that is high, or not possible to estimate (because the variance formula breaks down).
- If no data in an area, then no estimate available.

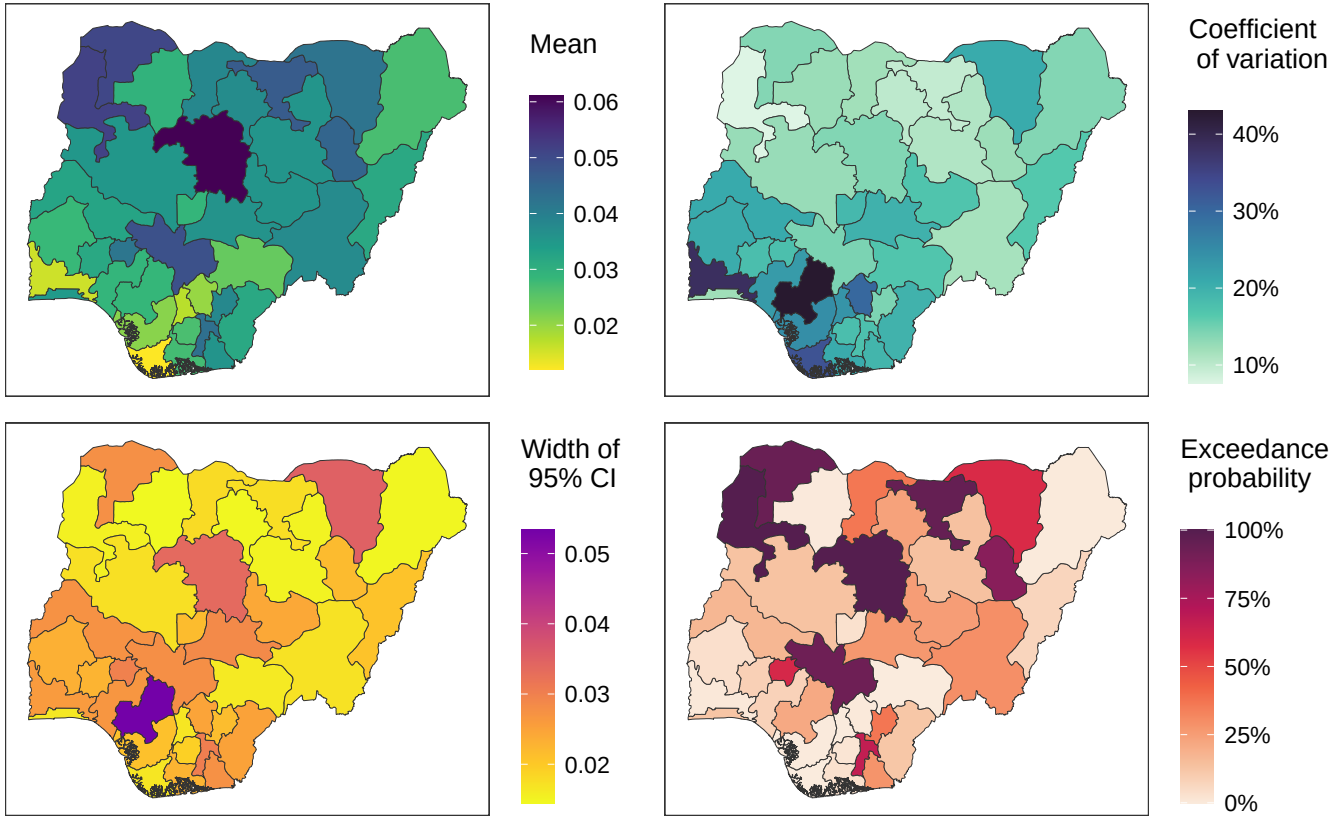


Figure 5: Summaries of state level weighted estimates for Nigeria for NMR.

- Weighted estimates account for the complex design and are reliable for states.
- Range of estimates is large – between-area variation.
- But uncertainty in each estimate is needed for interpretation.

Within-Area Estimation Variation of Prevalence

- 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
    - There may be many areas and then we may just stack the “top” and “bottom” areas, Figure 6 gives an example

- The ordering may be via the point estimate.

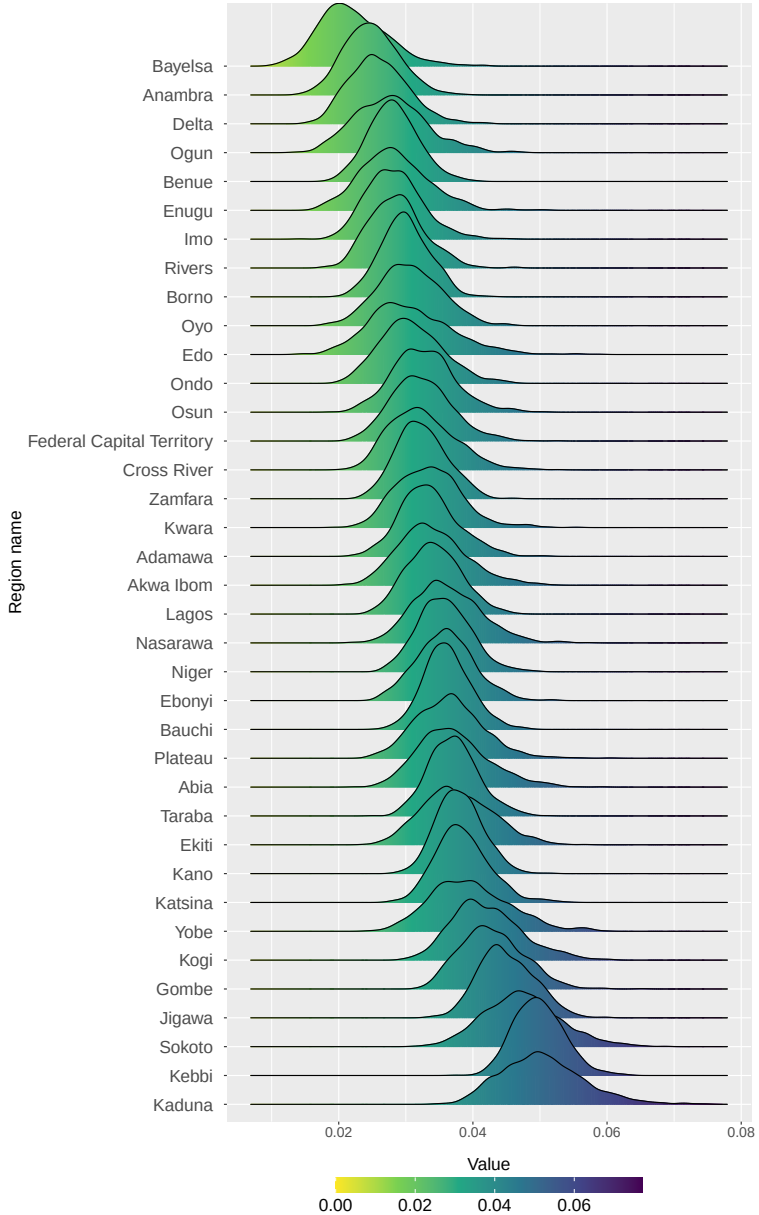
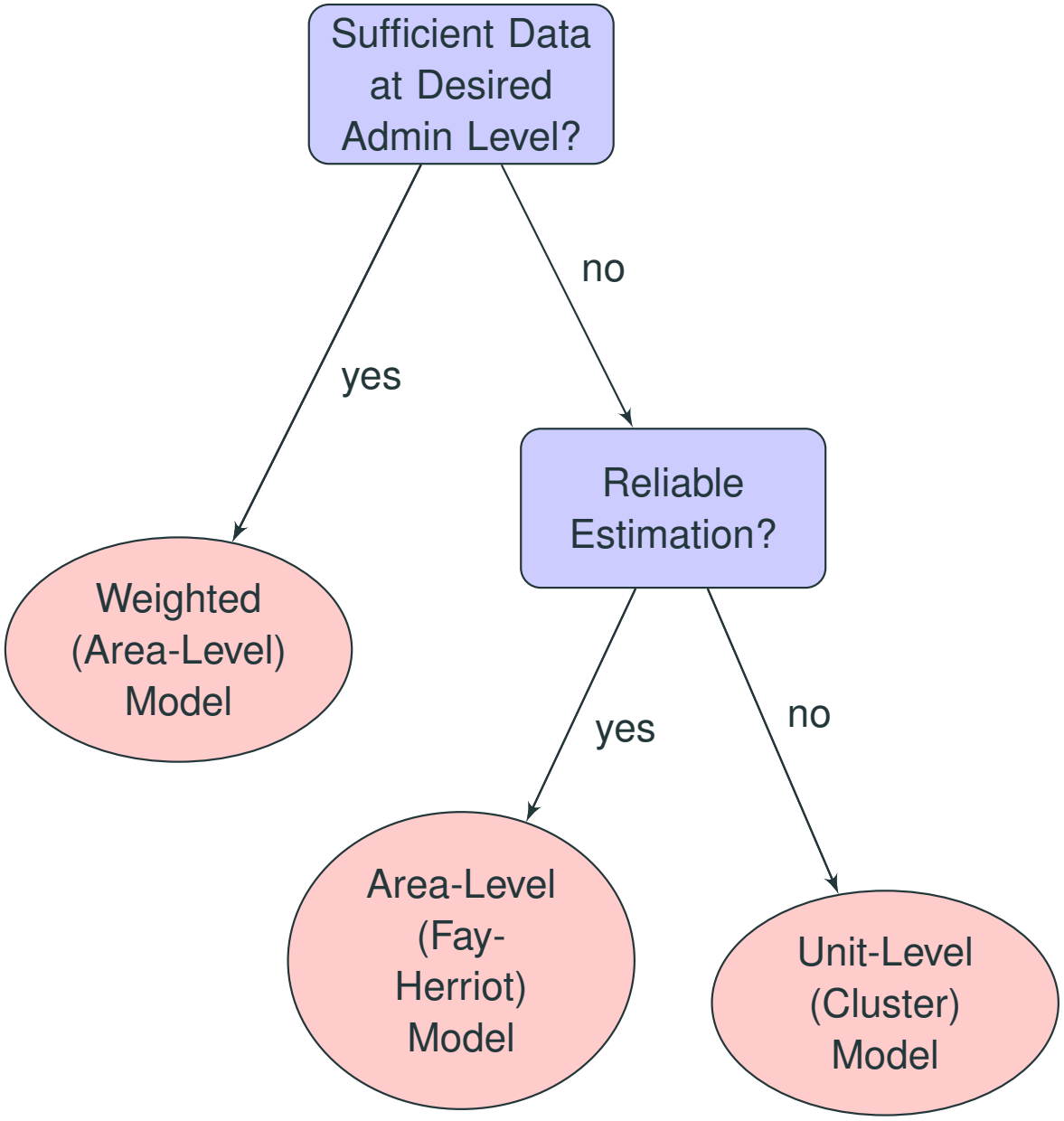


Figure 6: Ridgeplots for NMR across states in Nigeria.

Small Area Estimation (SAE): Area-Level Models

Methods we Employ Offer Two SAE Models



Area-Level (Fay-Herriot) Model

- The basic idea behind the area-level Fay-Herriot model, is to simultaneously model the collection of weighted estimates

$$\widehat{p}_i^{\,\mathrm{w}} \quad = \quad \frac{\widehat{T}_i}{\widehat{N}_i} \quad = \quad \frac{\sum_{k \in s_i} w_{ik} Y_{ik}}{\sum_{k \in s_i} w_{ik}}. \tag{2}$$

  from all areas.

- These estimates, $\widehat{p}_i^{\,\mathrm{w}}$, along with their standard errors, $\sqrt{V_i}$, constitute the data for $i = 1, \ldots, n$.

- The intuition is that we would expect some similarity of prevalences in areas that are "close" – this suggests we might benefit from using a model that encourages spatial smoothness in the estimates.

- We are effectively increasing the sample size in each area, by leveraging spatial similarity in prevalence.

Neighborhood Scheme in Nigeria

- Neighbors are conventionally defined through sharing a **common boundary** – this is a common choice but others are possible.

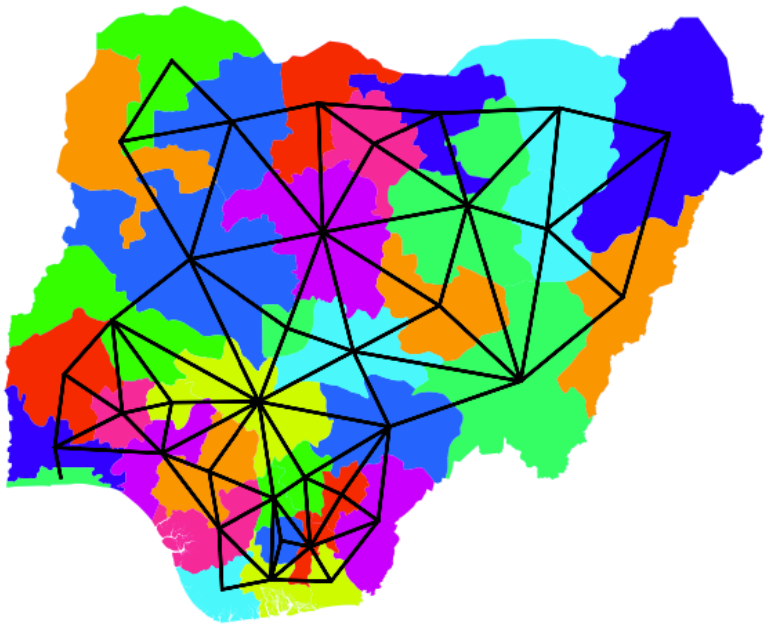


Figure 7: Neighborhood structure for 37 Admin1 states in Nigeria – this is used in the spatial smoothing models. Black lines link neighbors.

Area-Level (Fay-Herriot) Model

A wrinkle is that prevalences lie between 0 and 1, which makes modeling more tricky, so we take the logit transform.

Let,

$$\mu_i = \log\left(\frac{p_i}{1-p_i}\right),$$

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

The response variable is taken as,

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

The variance of these responses can be obtained from the variance of the prevalences, using the delta method, denote these as $V_i^\star$. So **data** is:

$$\underbrace{(Y_1, \sqrt{V_1^\star}), (Y_2, \sqrt{V_2^\star}), \quad \ldots \quad , (Y_{n-1}, \sqrt{V_{n-1}^\star}), (Y_n, \sqrt{V_n^\star})}_{\text{AREAS ARE LINKED THROUGH NEIGHBORHOOD STRUCTURE}}$$

Area-Level (Fay-Herriot) Model

In large samples,

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

with

- ϵ_i following a N(0, $V_i^\star$) distribution,
- $V_i^\star$ estimated using formulas relevant to the survey design.

In the Fay-Herriot area-level spatial model, it is assumed that

$$\underbrace{\mu_i}_{\text{TRUTH}} = \underbrace{\alpha}_{\text{OVERALL LEVEL}} + \underbrace{\delta_i}_{\text{SPATIAL RESIDUAL}}$$

The spatial terms δ_i are assumed to have spatial structure, i.e., they are spatially linked, strength of smoothness is controlled with a tuning parameter that is estimated from the data.

The spatial model we use corresponds to the celebrated BYM2 model (Riebler et al., 2016), with the name based on the original Besag, York, Mollié model (Besag et al., 1991).

Area-Level (Fay-Herriot) Model: Technical Version

The spatial residual is decomposed as:

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix} = \lambda \left(\sqrt{\phi} \begin{bmatrix} S_1 \\ \vdots \\ S_n \end{bmatrix} + \sqrt{1-\phi} \begin{bmatrix} E_1 \\ \vdots \\ E_n \end{bmatrix} \right)$$

where

- S_i is the spatial term and $E_i \sim \mathrm{N}(0,1)$ is the non-spatial term,
- 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.
- For the spatial terms:

$$S_i \mid S_j, \quad j \in \mathrm{ne}(i) \quad \text{follow a} \quad \mathrm{N}(\overline{S}_i, 1/m_i) \quad \text{distribution},$$

where $\mathrm{ne}(i)$ is the set of neighbors of i, $\overline{S}_i$ is the mean of the neighbors and m_i is the number of such neighbors – a local smoothing model.

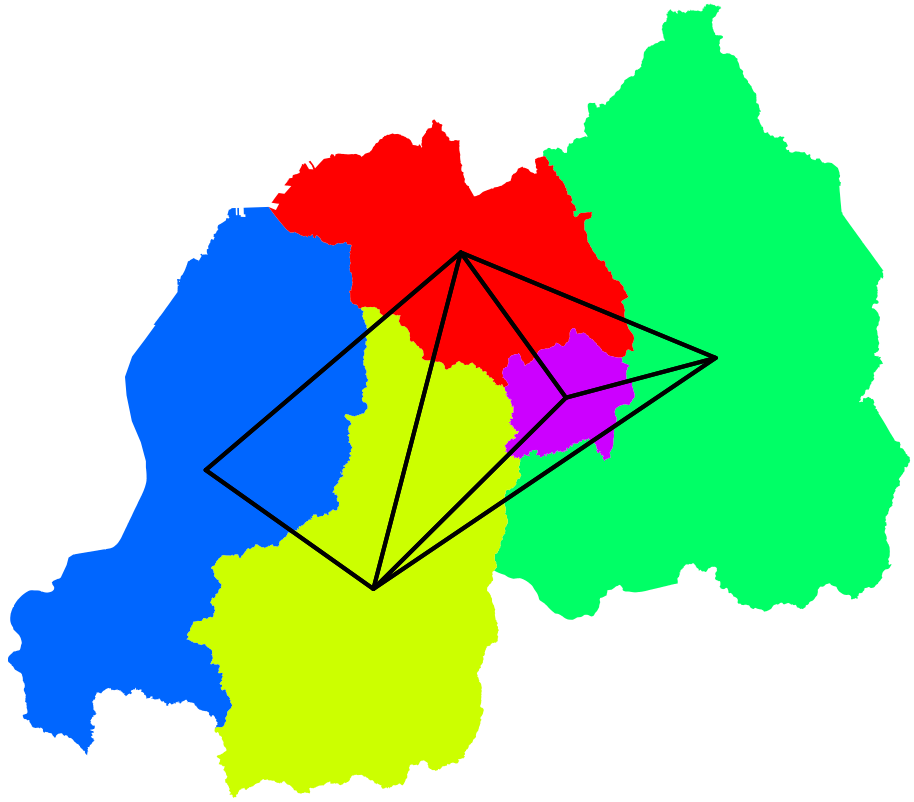


Figure 8: Neighborhood structure for 5 Admin1 areas in Rwanda. Black lines link neighbors.

- 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):

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

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).
- Inference is Bayesian with implementation via integrated nested Laplace approximation (INLA) (Rue et al., 2009) and is fast.
- Penalized complexoty (PC) priors (Simpson et al., 2017) are used for hyperparameters.

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).

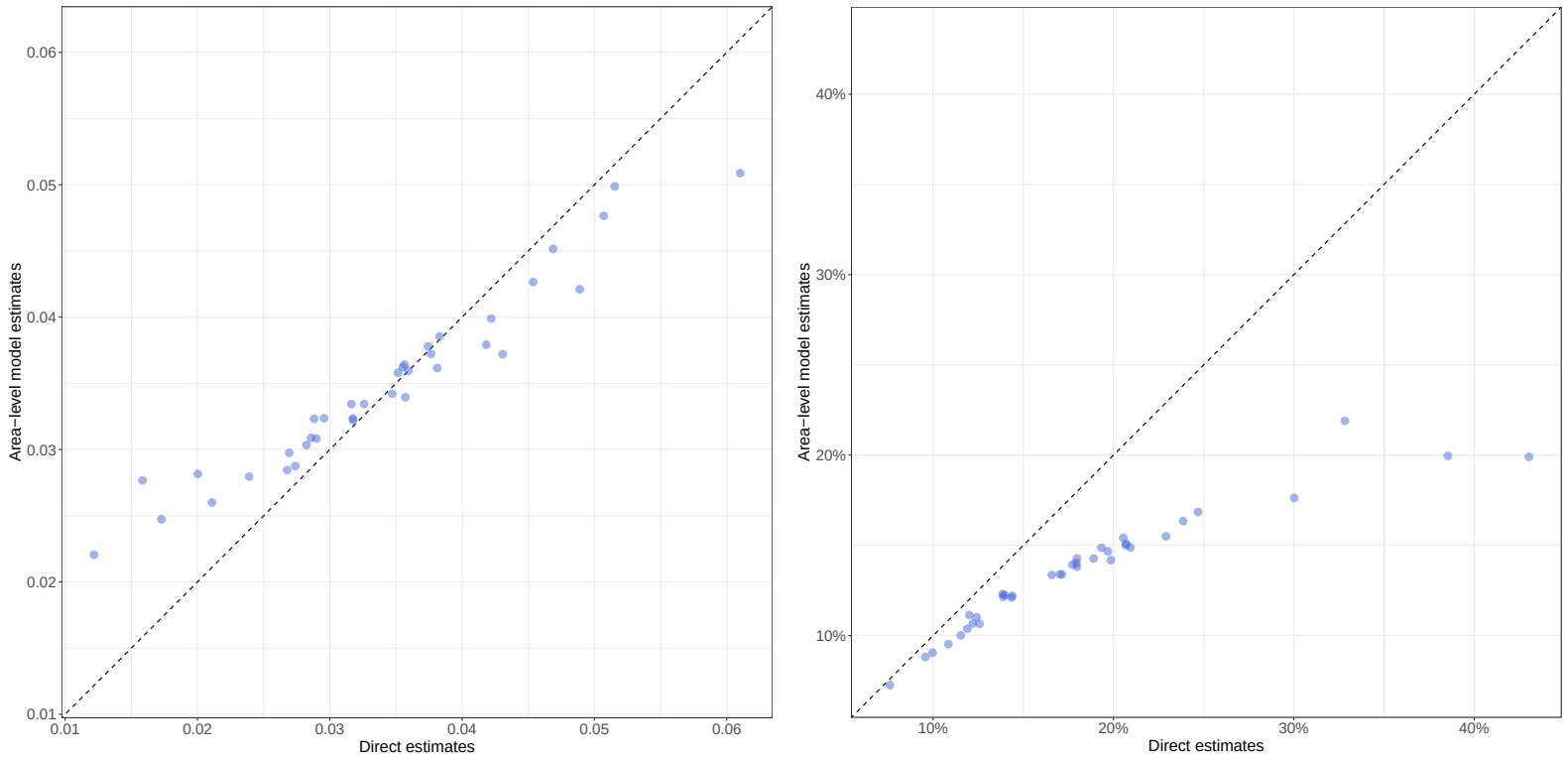


Figure 9: Area-level estimates versus direct estimates (left) and area-level CVs versus direct CVs (right), for states.

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

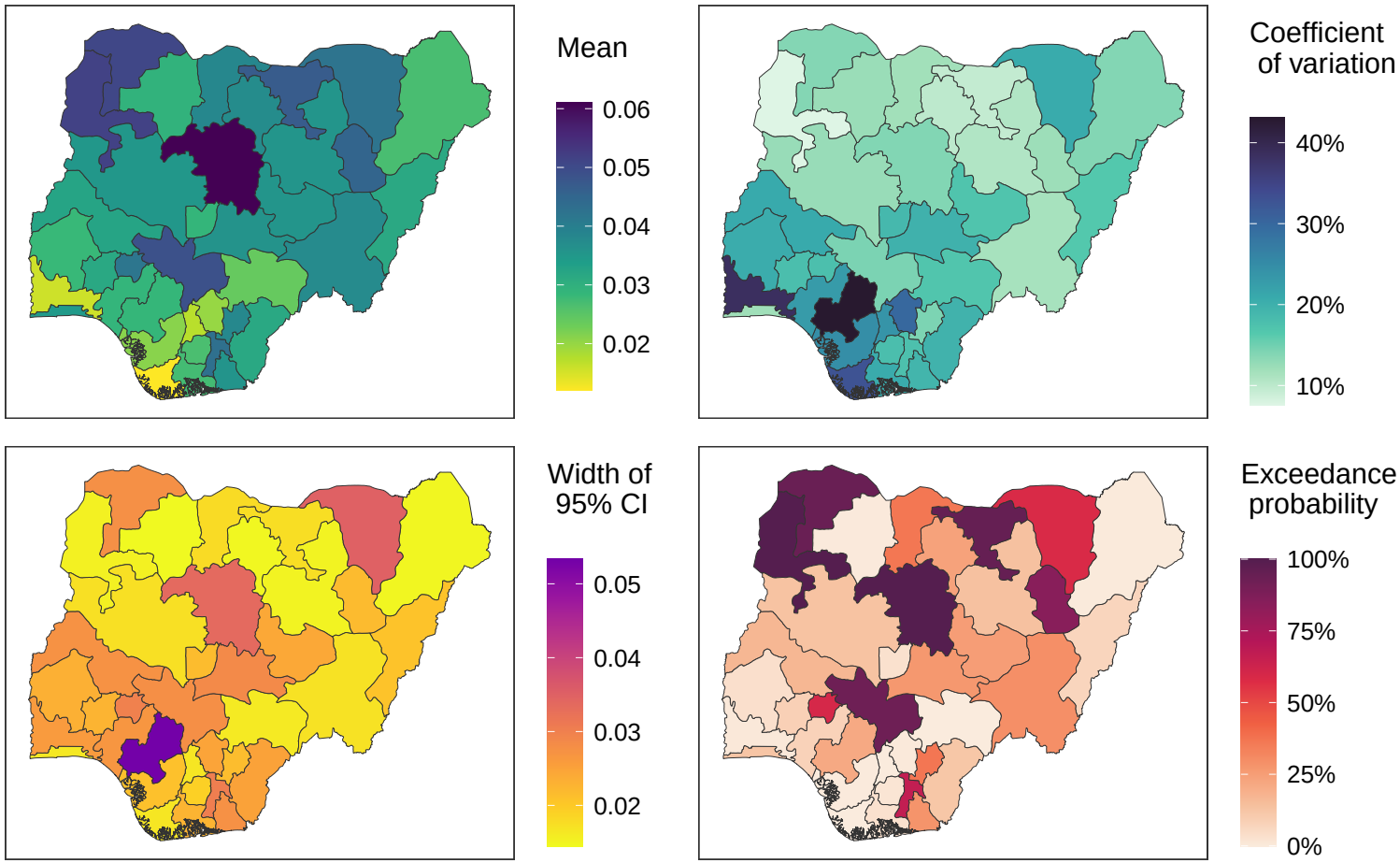


Figure 10: Summary measures for direct estimates at Admin1.

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

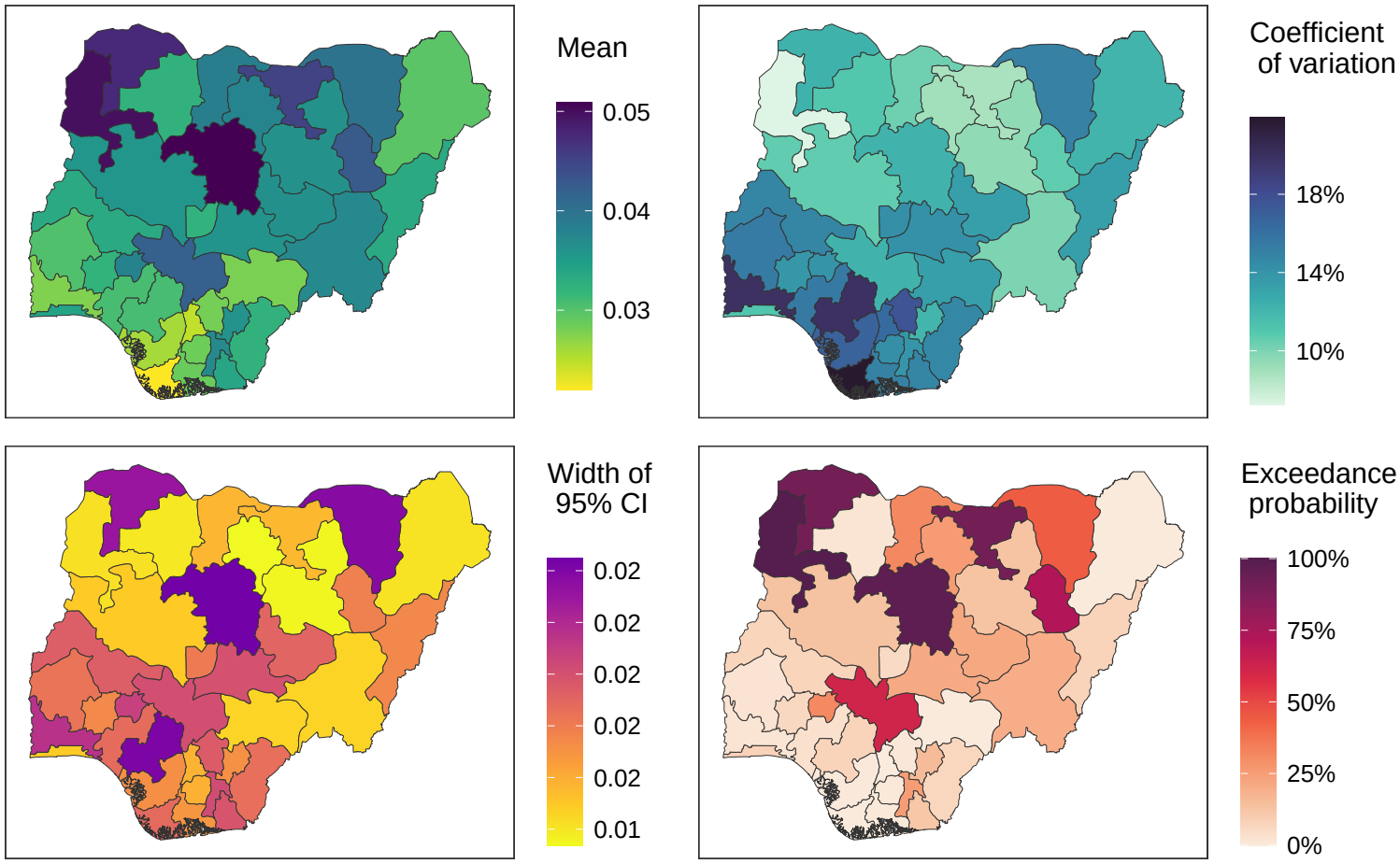


Figure 11: Summary measures for area-level (Fay-Herriot) estimates at Admin1.

- Compared to direct estimates (Figure 10) a narrower range in the point estimates.
- Uncertainty measure show less variation also, due to sharing of information.

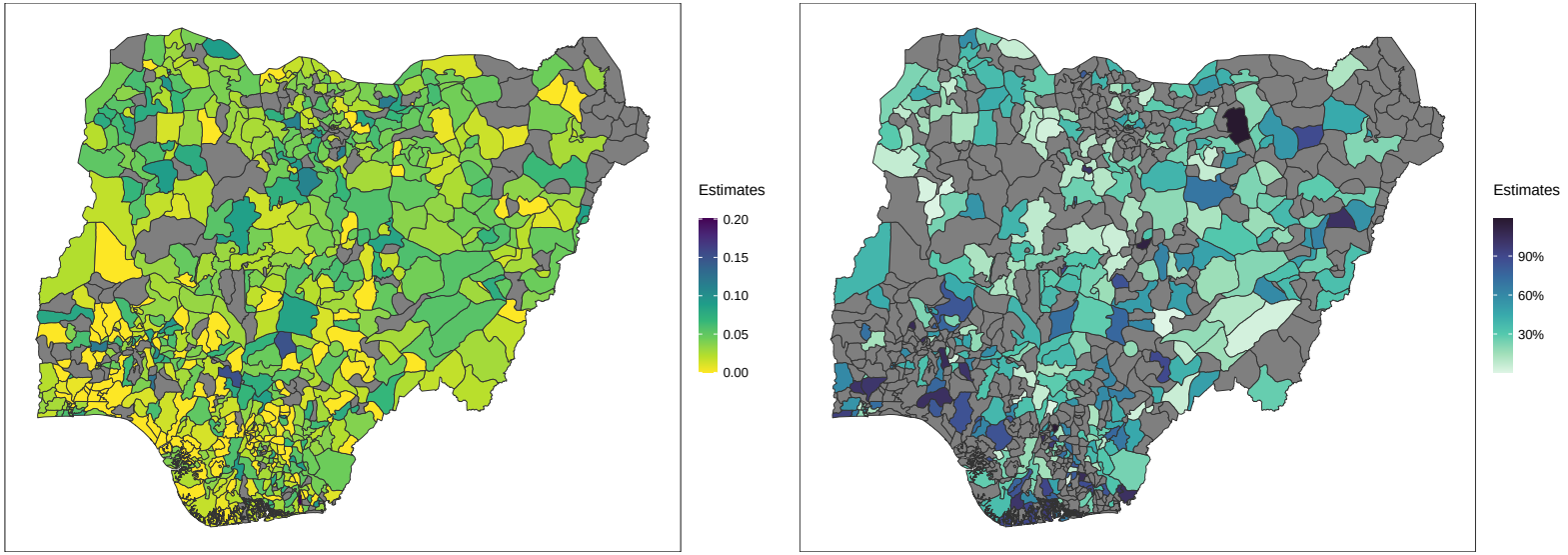


Figure 12: Direct prevalence and CV maps at Admin2.

- 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 Admin2 (LGA) level – unit-level models are the only possibility.
- The shinyApp gives warnings.

Small Area Estimation (SAE): Unit-Level Models

Unit-Level Models

- When it is not possible to fit area-level models, we must turn to unit-level models.
- These represent a conventional model-based (as opposed to design-based) approach.
- We model individual responses at the cluster-level.
- The obvious model for a set of Bernoulli (0/1) responses in each cluster is a binomial – we use an overdispersed version, namely the betabinomial – it has an extra parameter to allow for extra-binomial variation (dependent observations within each cluster imply this will be present).
- An important point: While we would like estimates with high precision, the first hurdle is getting a precision that is appropriate.

Unit-Level Model

- As a concrete example, consider neonatal mortality risk.
- Suppose Y_{ic} deaths out of n_{ic} births in sampled cluster c.
- We use an overdispersed binomial model:

$$Y_{ic}|p_{ic} \quad \sim \quad \text{BetaBinomial}(n_{ic}, p_{ic}, d), \tag{3}$$

  where
  - p_{ic} is the risk of neonatal death in cluster c of area i.
  - $d > 0$ is an overdispersion parameter that accounts for within-cluster correlation. Parameterizaton: $\text{var}(Y_{ic}) = n_{ic}p_{ic}(1-p_{ic})[1+(n_{ic}-1)d]$.
- Model for cluster risk p_{ic}:

$$\frac{p_{ic}}{1-p_{ic}} = \exp(\underbrace{\alpha}_{\text{OVERALL LEVEL}} + \underbrace{\delta_i}_{\text{SPATIAL RESIDUAL}})$$

- The δ_i again encourage spatial smoothing of risk, via the BYM2 model.

Unit-Level Model

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).

Unit-Level Model

Disadvantages:

- Since the weights are not used, one must adjust for the design within the model specification, for example, by modeling the association between prevalence and urban/rural.
- 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).
- 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 variance too small? This is a very difficult question to answer.

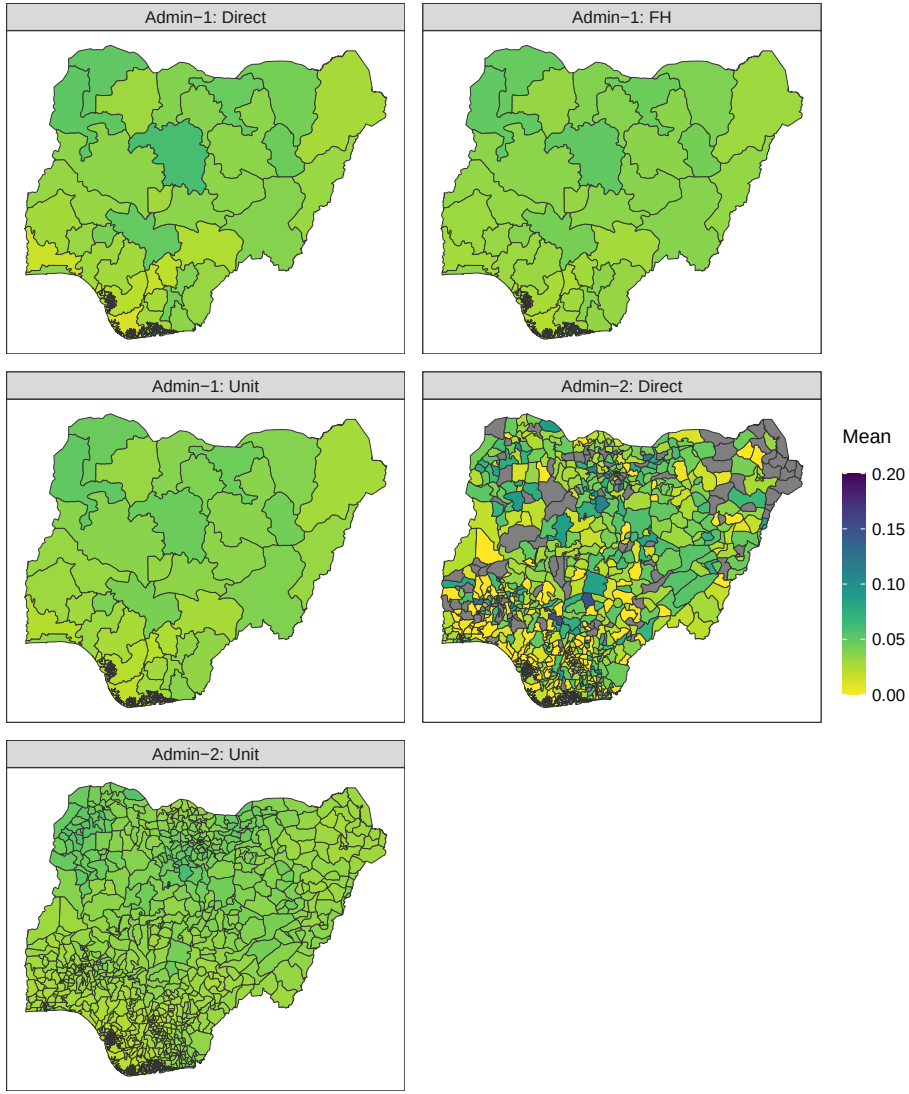


Figure 13: Direct, area-level and unit-level point estimates, at Admin1 and Admin2.

- At Admin1, all 3 methods give very similar results.
- At Admin2, there is more variation, with unit-level and area-level being similar and displaying shrinkage.

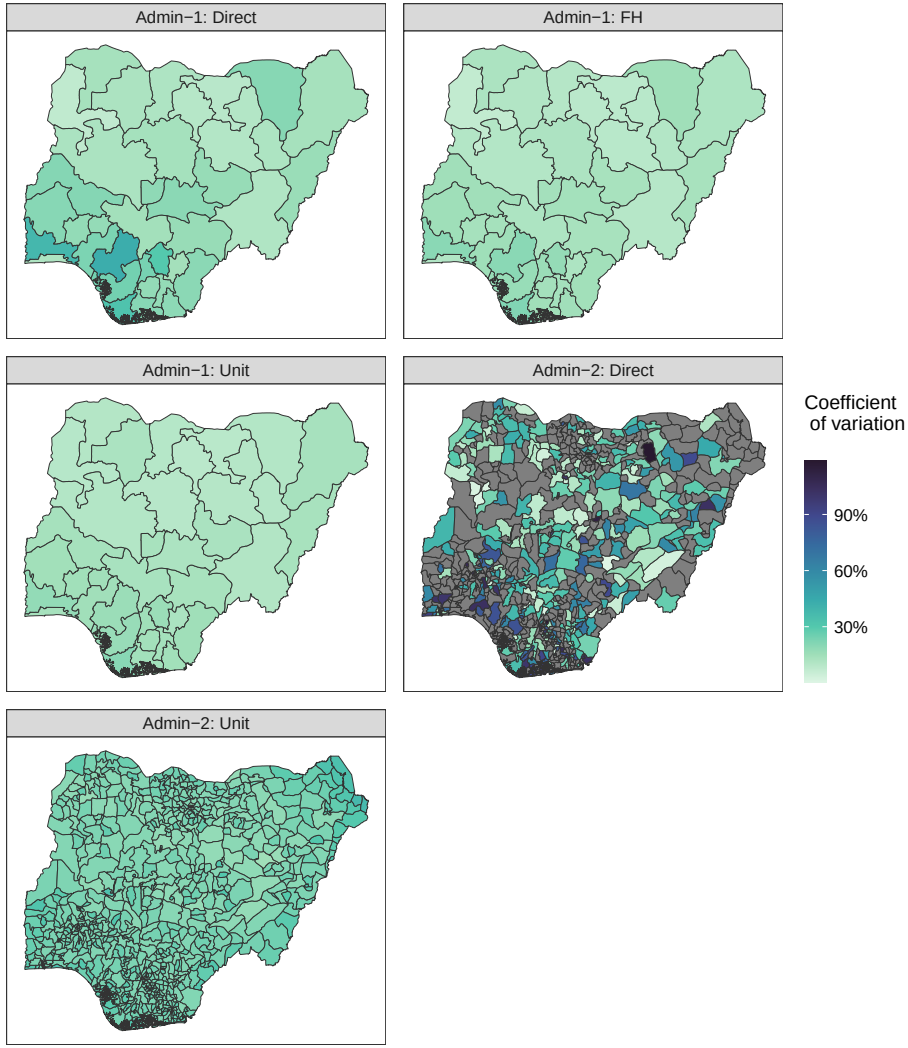


Figure 14: Direct, area-level and unit-level CVs, at Admin1 and Admin2.

- At both Admin1 and Admin2, there is an increase in precision in moving from direct to area-level to unit-level.

Ranking Plots

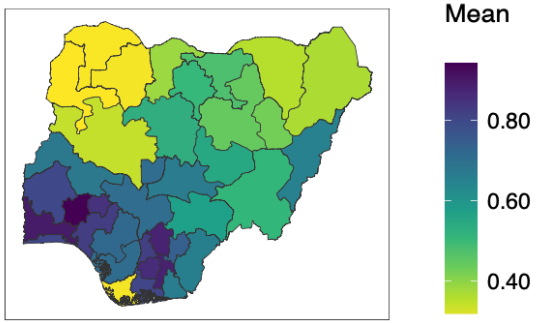


Figure 15: ANC4+ visits from Nigeria DHS 2018.

Low prevalence areas are mostly in the north-west (Sokoto, Kebbi, Zamfara) but also Bayelsa in the south.

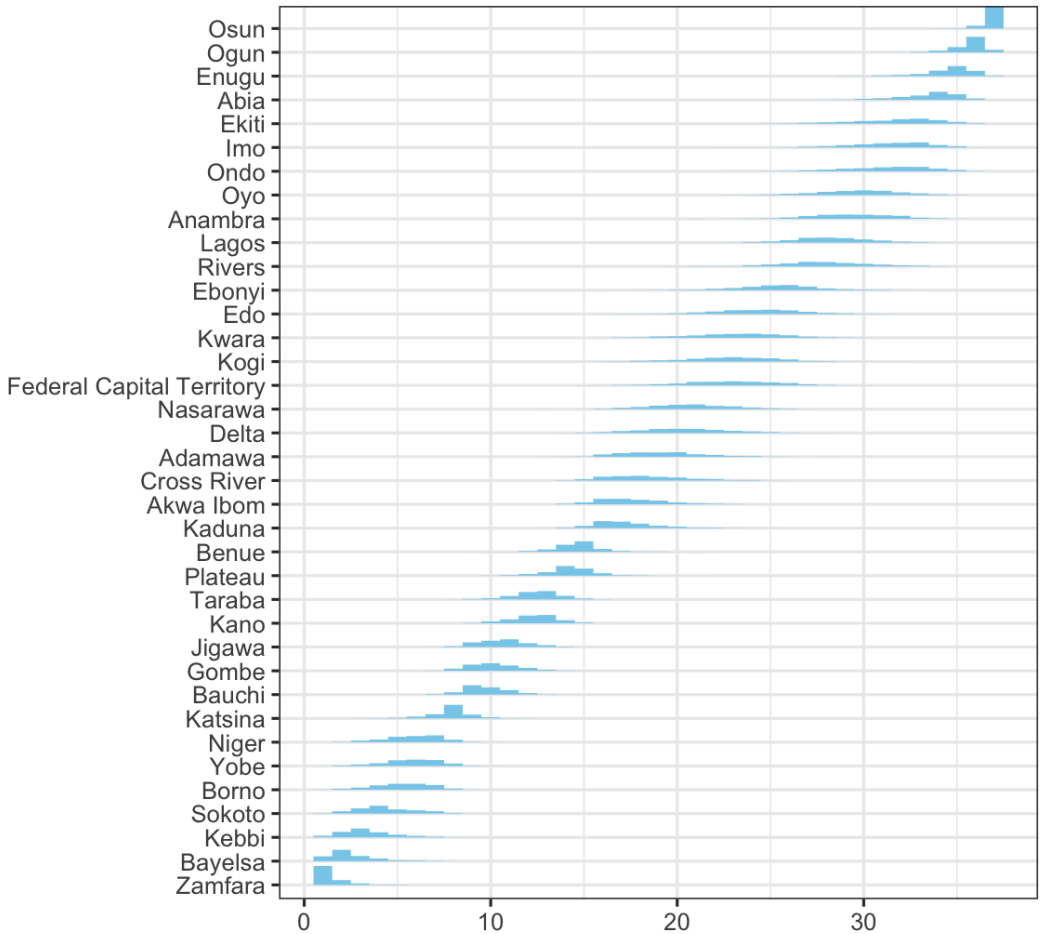


Figure 16: State-level ranks for ANC4+.

Software and Country Engagement

Software Evolution

- We have been working on SUMMER package in R since 2017 – focusses mostly on mortality estimation.
- Experience from UN and WHO workshops in 2019 in Ecuador, South Africa, Malawi is that researchers need a more user-friendly interface.
- surveyPrev package in R allows SAE for binary indicators – introduced in 2023.
- U5MR and fertility are hard because data model is complex – can be done in SUMMER.
- Created a pipeline for U5MR using DHS data and SUMMER (Wu et al., 2021).
- survey package now has a generic sae module.
- Most recently, developed a shinyApp that has successfully used at WHO and UNICEF remote and online workshops – internet based app with all computation done in the cloud; $>$ 150 indicators in the app, and more being added all the time.

Country Engagement

Countries who have already been exposed to the shinyApp, either through collaborations or workshops, in-person or remote:

- Benin
- Burkina Faso
- Central African Republic
- Côte d'Ivoire
- Democratic Republic of the Congo
- Guinea-Bissau
- Kenya
- Nigeria
- Rwanda
- Senegal
- Sierra Leone
- United Republic of Tanzania
- Zambia

Methodology: Selected Publications

- Wu, Y., and Wakefield, J. (2024). Modeling urban/rural fractions in low- and middle-income countries. Journal of the Royal Statistical Society, Series A. Published online: 19 January, 2024.
- Gao, P.A., and Wakefield, J. (2024). Smoothed model-assisted small area estimation of proportions. The Canadian Journal of Statistics, 52, 337-358.
- 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.
- Paige J, Fuglstad G-A, Riebler A and Wakefield J (2022). Spatial aggregation with respect to a population distribution: impact on inference. Spatial Statistics, 52, 100714.
- Wu, Y., Li, Z.R., Mayala, B.K., Wang, H., Gao, P., Paige, J., Fuglstad, G.-A., Moe, C., Godwin, J., Donohue, R.E., Croft, T.N., and Wakefield, J. (2021). Spatial modeling for subnational administrative level 2 small-area estimation. DHS Spatial Analysis Reports No. 21. Rockville, Maryland, USA.
- 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.
- Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. International Statistical Review, 88, 398-418.
- Marquez, N. and Wakefield, J. (2021). Harmonizing child mortality data at disparate geographic levels. Statistical Methods in Medical Research, 30, 1187-1210.
- Godwin, J. and Wakefield, J. (2021). Space-time modeling of child mortality at the admin-2 level in a low and middle income countries context. Statistics in Medicine, 40, 1593-1638.
- Paige, J., Fuglstad, G.-A., Riebler, A. and Wakefield, J. (2021). Design- and model-based approaches to small-area estimation in a low and middle income country context: comparisons and recommendations. Journal of Survey Statistics and Methodology, 10, 50-80.
- Li, Z.R., Hsiao, Y., Godwin, J., Martin, B.D., Wakefield, J., and Clark, S.J. (2019). Changes in the spatial distribution of the under-five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. PLoS One, 14(1), e0210645.
- Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S.J. (2018). Estimating under five mortality in space and time in a developing world context. Statistical Methods in Medical Research, 9, 2614-2634.
- Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Mosanja, H., and Clark, S. (2015). Small area estimation of childhood mortality in the absence of vital registration. Annals of Applied Statistics, 9, 1889-1905.

Models for Urban/Rural Stratification

Acknowledging the Design: Stratification

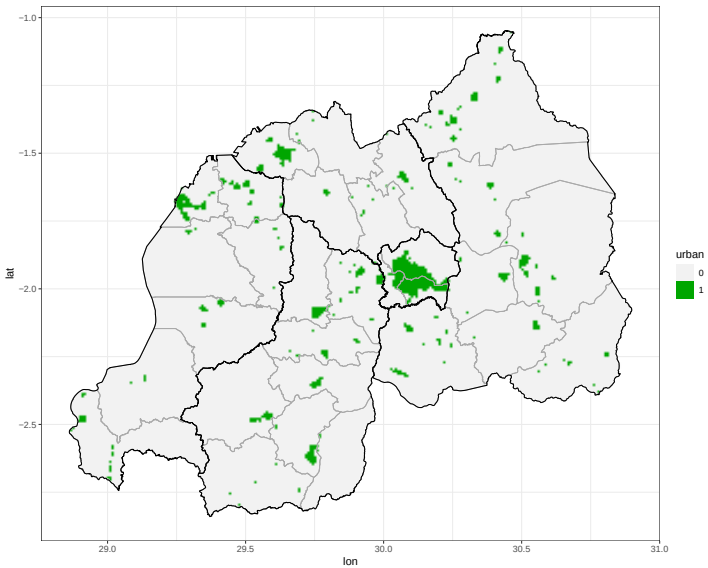


Figure 17: In the DHS in Rwanda, stratification is based on provinces and urban/rural.

- 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 of 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.

Sampling Frame Summaries

Table B.2 Enumeration areas and households

Distribution of enumeration areas (EAs) and average number of households in a EA by district, according to residence, Malawi DHS 2015-16

	Number of EAs			Average EA size		
District	Urban	Rural	Total	Urban	Rural	Total
Chitipa	11	205	216	266	170	175
Karonga	37	370	407	232	133	142
Nkhatabay	12	229	241	190	175	175
Rumphi	12	156	168	321	206	215
Mzimba	122	825	947	255	168	179
Likoma	2	9	11	150	191	184
Kasungu	29	486	515	309	243	247
Nkhota kota	16	177	193	313	325	324
Ntchisi	6	204	210	259	225	226
Dowa	18	450	468	249	261	260
Salima	22	416	438	277	172	177
Mchinji	12	374	386	298	250	252
Dedza	15	486	501	299	291	291
Ntcheu	11	468	479	301	236	238
Lilongwe	458	1,173	1,631	336	235	263
Mangochi	25	614	639	339	289	291
Machinga	19	436	455	279	252	253
Chiradzulu	2	334	336	296	212	213
Mwanza	9	80	89	383	232	247
Thyolo	12	674	686	200	207	207
Mulanje	17	658	675	191	189	189
Phalombe	3	316	319	372	239	240
Chikwawa	16	380	396	177	251	248
Nsanje	14	241	255	302	201	206
Balaka	17	275	292	296	257	259
Neno	3	157	160	122	160	159
Zomba	79	584	663	241	244	243
Blantyre	412	381	793	373	212	296
Malawi	**1,411**	**11,158**	**12,569**	**319**	**225**	**235**

Source: The 2008 Malawi Population and Housing Census (MPHC) Sampling frame provided by the Malawi National Statistical Office (NSO).

Figure 18: Number of EAs (clusters) in the sampling frame (2008 census) used in the 2015 DHS.

Sampled Clusters Summaries

Table B.3 Sample allocation of clusters and households

Sample allocation of clusters and households by district, according to residence, Malawi DHS 2015-16

	Number of clusters allocated			Number of households allocated		
District	Urban	Rural	Total	Urban	Rural	Total
Chitipa	5	20	25	150	660	810
Karonga	8	20	28	240	660	900
Nkhatabay	5	22	27	150	726	876
Rumphi	6	20	26	180	660	840
Mzimba	11	24	35	330	792	1,122
Likoma*	4	18	22	120	594	714
Kasungu	7	26	33	210	858	1,068
Nkhota kota	6	22	28	180	726	906
Ntchisi	4	23	27	120	759	879
Dowa	5	28	33	150	924	1,074
Salima	6	23	29	180	759	939
Mchinji	5	26	31	150	858	1,008
Dedza	5	29	34	150	957	1,107
Ntcheu	4	28	32	120	924	1,044
Lilongwe	14	23	37	420	759	1,179
Mangochi	6	29	35	180	957	1,137
Machinga	5	27	32	150	891	1,041
Chiradzulu	2	27	29	60	891	951
Mwanza	7	19	26	210	627	837
Thyolo	4	30	34	120	990	1,110
Mulanje	4	29	33	120	957	1,077
Phalombe	3	27	30	90	891	981
Chikwawa	4	27	31	120	891	1,011
Nsanje	6	21	27	180	693	873
Balaka	6	24	30	180	792	972
Neno	3	23	26	90	759	849
Zomba	9	26	35	270	858	1,128
Blantyre	19	16	35	570	528	1,098
Malawi	**173**	**677**	**850**	**5,190**	**22,341**	**27,531**

Note: Due to the small number of EAs in Likoma, 11 EAs, all the EAs were included in the DHS, where each EA was segmented in to two segments, yielding a total number of 22 clusters; 30 and 33 households were selected per urban and rural cluster, respectively.

Figure 19: Number of EAs (clusters) in the sample, for the DHS 2015.

Comparison of Urban Cluster Fractions in Sampling Frame and Sample

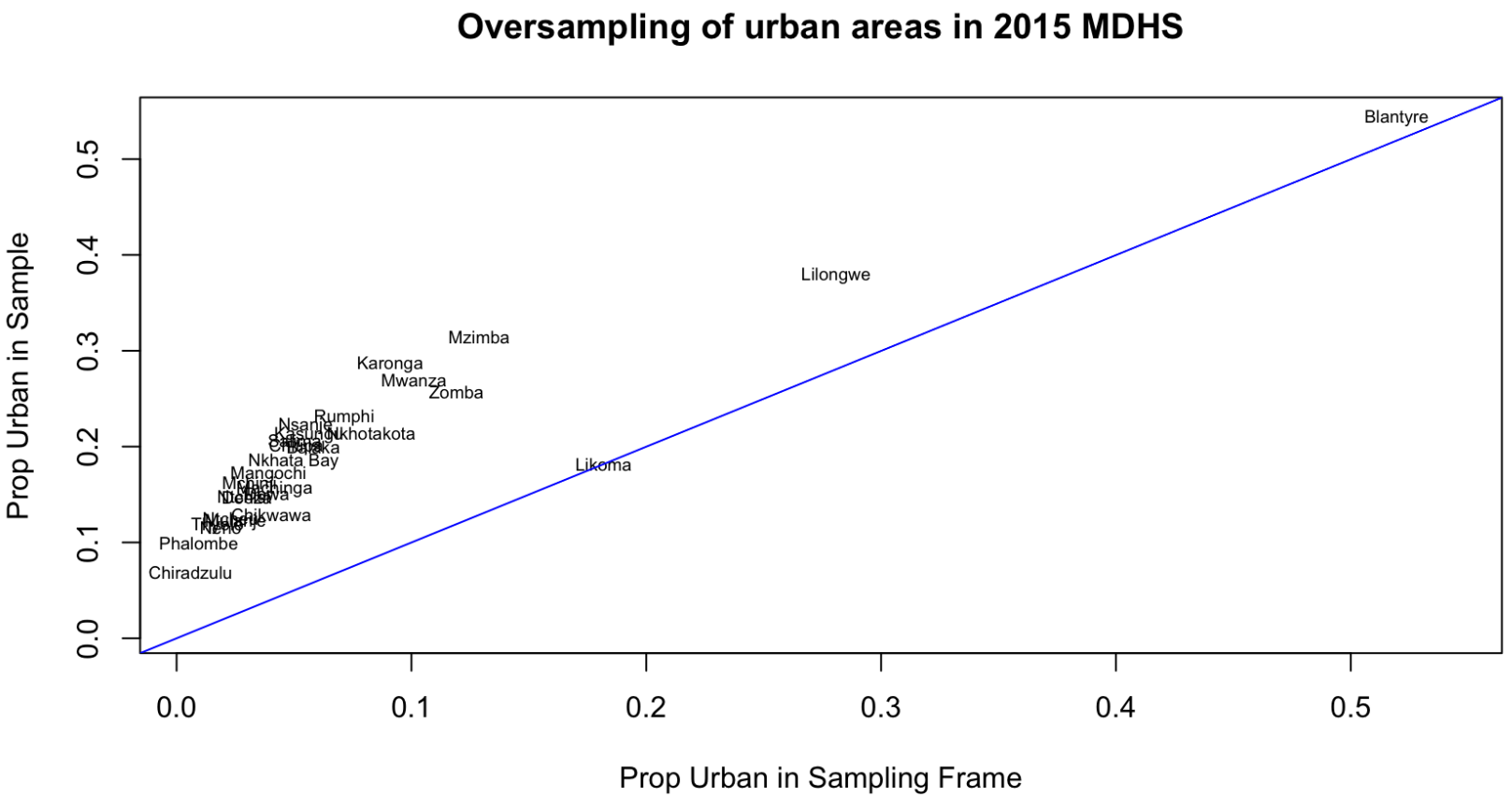


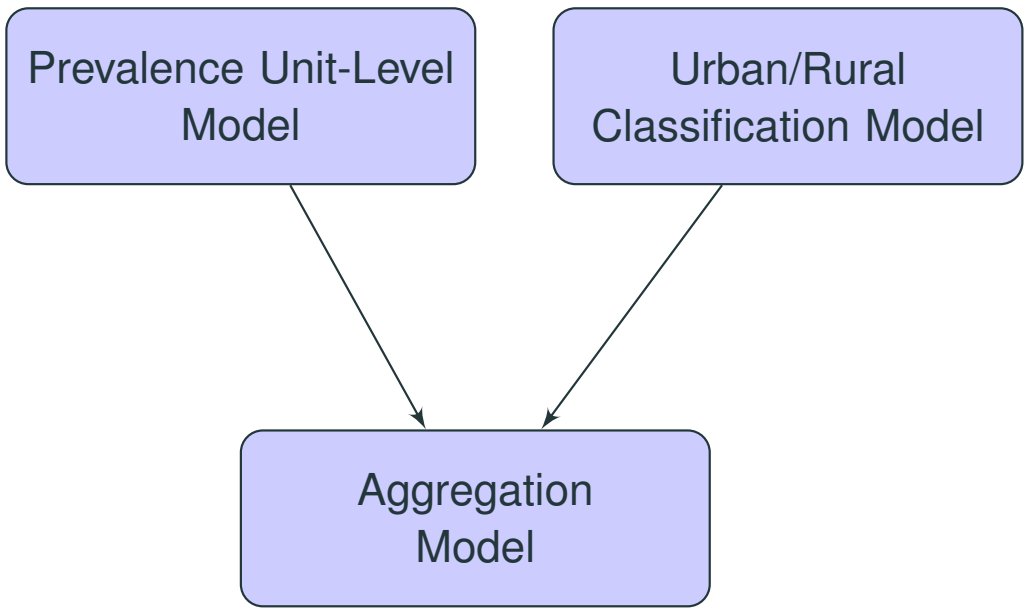
Figure 20: Proportions of clusters urban in the sample versus in the sampling frame.

Overview of Modeling

Bias will result under the following two conditions:

- The sampling of clusters is such that the urban/rural ratio differs from the sampling frame.
- The response is associated with urban/rural.

If this holds we need to model the association, and weight the urban/rural prevalances appropriately.



Unit-Level Prevalence Model with Stratification

Unit-level model:

$$\begin{aligned} Y_{ic}|p_{ic} &\sim \text{BetaBinomial}(n_{ic}, p_{ic}, d) \\ p_{ic} &= \text{expit}\left(\alpha^{\text{rural}} I(c \in \text{rural}) + \alpha^{\text{urban}} I(c \in \text{urban}) + \delta_i\right) \end{aligned}$$

where

- d is the scale parameter.
- $\exp(\alpha^{\text{rural}})$ and $\exp(\alpha^{\text{urban}})$ are the associations with rural and urban (odds parameters).
- δ_i is an area-level spatial (BYM2) random effect.

The area-level prevalence in area i is:

$$\begin{aligned} p_i &= \Pr(\text{response}|\text{rural}, i) \times \Pr(\text{rural}|i) + \Pr(\text{response}|\text{urban}, i) \times \Pr(\text{urban}|i) \\ &= \text{expit}(\alpha^{\text{rural}} + \delta_i) \times r_i + \text{expit}(\alpha^{\text{urban}} + \delta_i) \times (1 - r_i) \end{aligned}$$

where r_i is the fraction rural of the relevant population and needs estimating.

The Sampling Frame: How to Estimate the Fraction Urban

- In the DHS, we have 2 urban/rural strata which we label as A and B, so as to remove the complex connection with "true" urban and rural classification over time – conceptually it is easiest to think of this as a geographical partition of clusters.
- In a model-based approach, we need to adjust for A and B when the outcome depends on A/B and the survey did unequal sampling with respect to A and B.
- The definition and allocation of A and B clusters occurs at the time of the census (which is the sampling frame), say at $t = 0$.
- The A and B cluster map is formed whenever the sampling frame is formed, and is constant over time (just as the weights are).
- However, the populations in those areas do change over time.
- In a perfect world, we would have access to all clusters in the sampling frame, along with the relevant populations in these clusters.

Methods for Estimating the Urban/Rural Fractions

1. From the DHS manual, use the proportion of households in each area (usually not available at Admin-2).

2. From the DHS manual, use the proportion of clusters in each area (usually not available at Admin-2).

3. Estimate of urban/rural relevant population using DHS.

4. Classification algorithms:

    - logistic regression,
    - machine learning prediction algorithms (regression trees, super learner,...).
    - Adjust for jittering?

5. Census proportions.

Details in Wu and Wakefield (2024).

Covariates

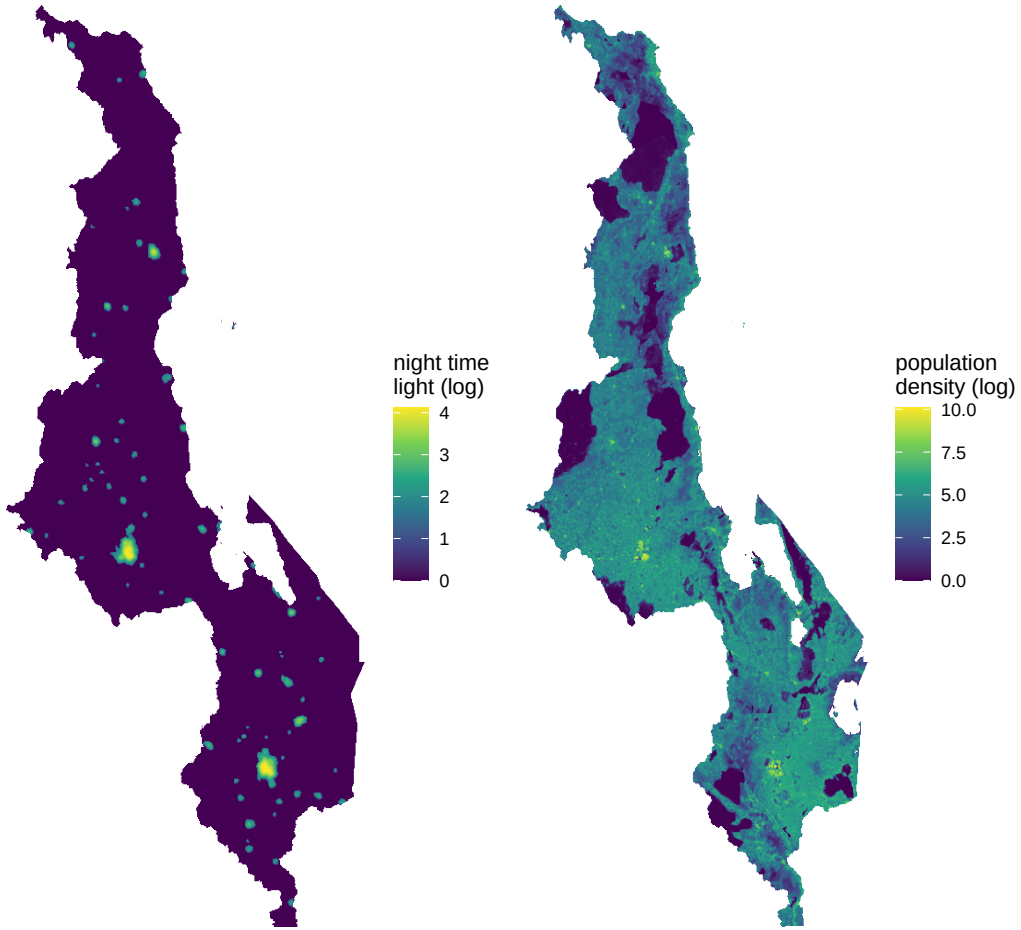


Figure 21: Log nighttime lights and log population density.

Classification Surfaces

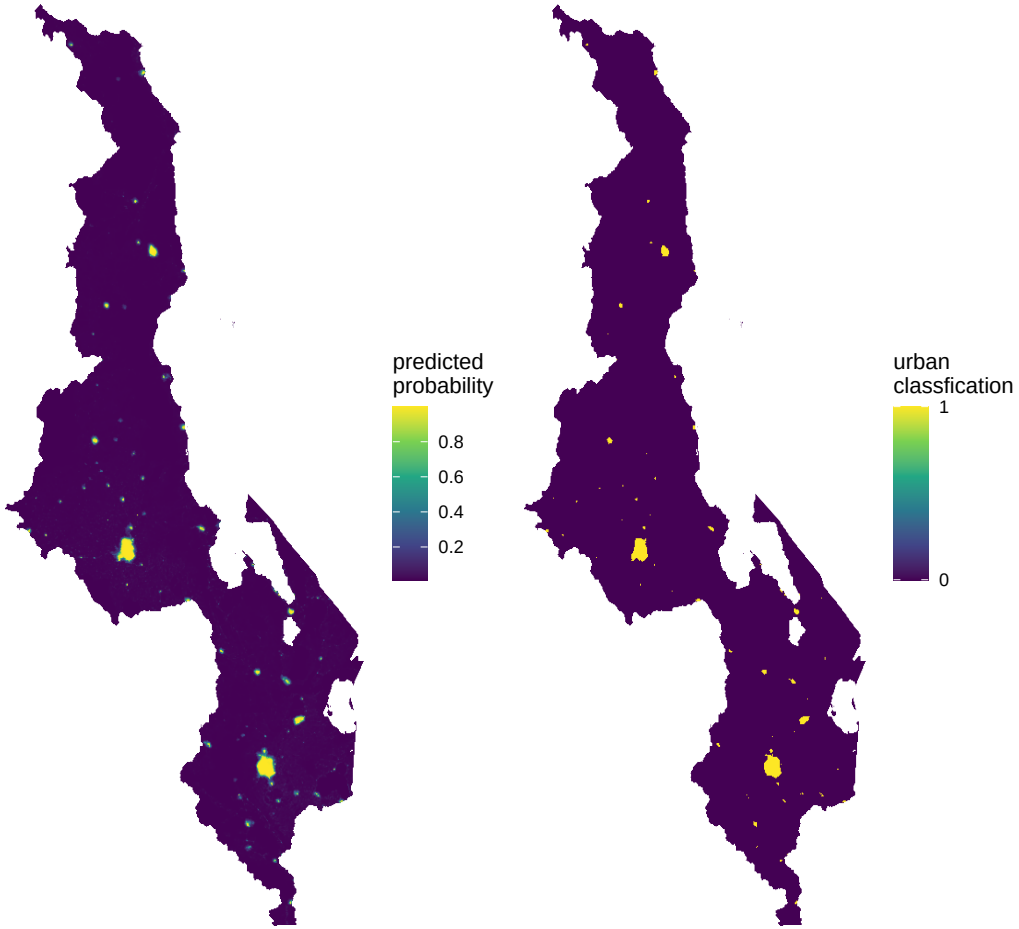


Figure 22: Probabilistic classification (left) and binary classification (right). Classification via logistic regression.

Admin-1 Urban Fractions

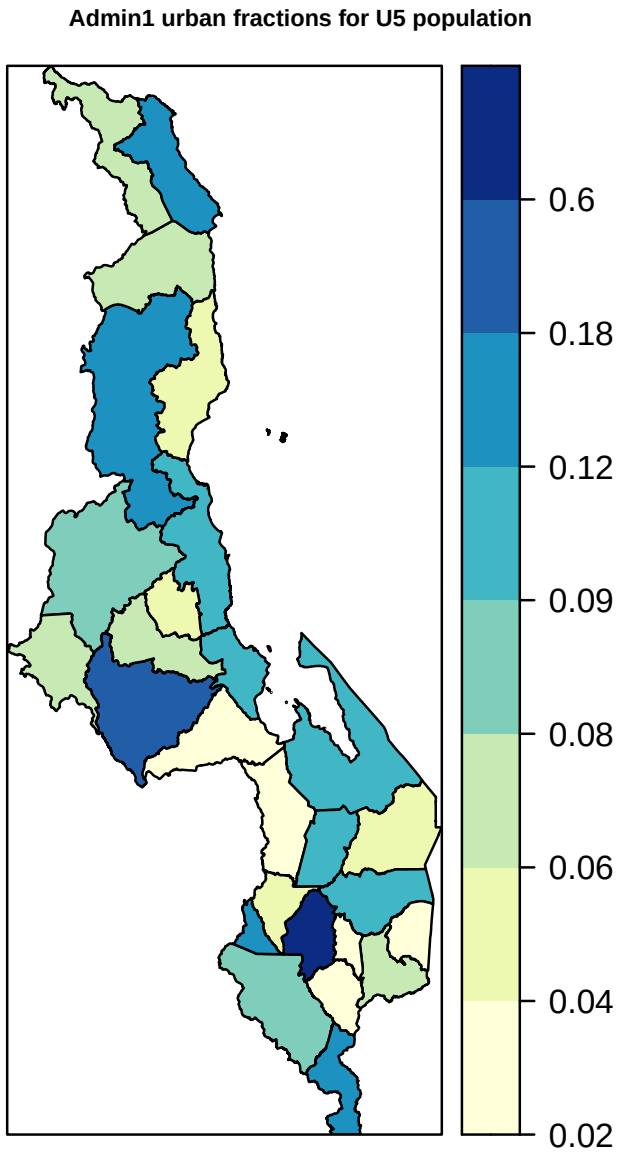


Figure 23: Admin-1 urban fractions from GBT corrected classification in 2015.

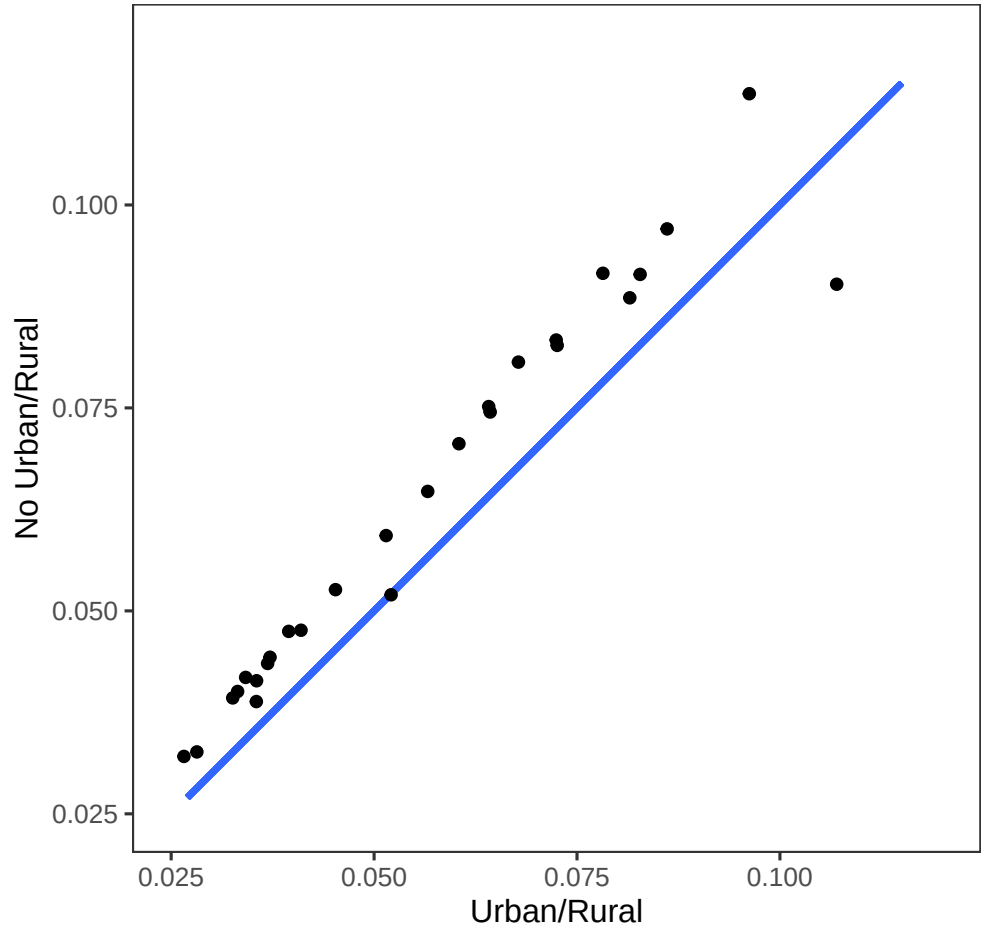


Figure 24: Malawi district-level prevalence estimates from two unit-level models. On the y-axis, the prevalence estimates are from a model with no urban/rural adjustment, while on the x-axis the model has an adjustment.

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

Summary and Discussion

Summary

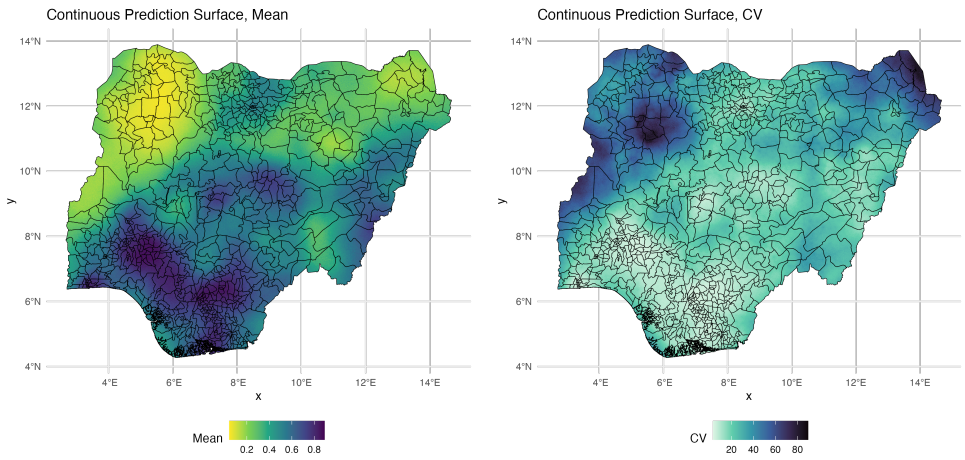
- If direct estimates have acceptable precision, then they are highly recommended as reliable!
- If we can calculate direct estimates and their CVs, but the latter are deemed too large, then we should turn to area-level (Fay-Herriot) models.
- If area-level models result in insufficient precision, or the data are too sparse too produce the required inputs to an area-level model (point estimate and standard error), then we can turn to unit-level models.
- Unit-level models should be used with caution.

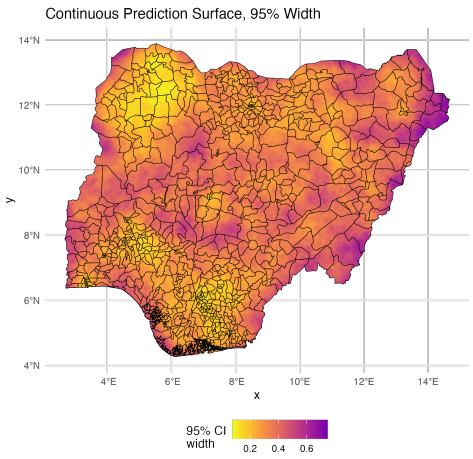
Details on methods, videos, latest updates, and much more can be found at:

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

Discussion: Two Approaches to Spatial Smoothing

- Traditional SAE approaches introduce random effects at the area level. The discrete spatial model (BYM2) is implemented in the shinyApp.
- An alternative is to model at the point level using a continuous spatial model (as right).
- Beware of spurious geographic detail and precision.





If using continuous models, in addition to acknowledging the design, one must also perform aggregation with respect to population density, which can introduce extra bias and noise – not yet ready for prime time in the shinyApp.

Discussion

- Overshrinkage a worry: nested spatial models being investigated.
- Model checking still in infancy: we use cross-validation to predict weighted estimates.
- Machine learning approaches are very intoxicating, but can the uncertainty be quantified?
- Benchmarking to known totals is less popular in LMICs because the population information may be inaccurate of out of data, see (Okonek and Wakefield, 2022) for references to the benchmarking literature.
- Fay-Herriot variance modeling can be carried out to increase the utlity of this method (Gao and Wakefield, 2023).

Next Steps for the App

- Have started a collaboration to allow MICS data to be analyzed.
- Covariates/auxiliary data to improve prediction.
- Nested spatial models.
- Expand the indicator list, including U5MR and IMR.
- Model checking diagnostics.
- Guidelines for which of weighted, area-level, unit-level, all fail, needed.
- Adding more graphics (including ranking plots).
- Continuous spatial models.
- Prevalences to Counts using WorldPop estimates.
- More training materials.

- Feedback and suggestions essential!

The Team

Space Time Analysis Bayes (STAB) Lab: Cory Arrouzet, Orvalho Augusto, Miaolei Bao, David Coomes, Qianyu Dong, Ameer Dharamshi, Geir-Arne Fuglstad, Peter Gao, Jessica Godwin, Jitong Jiang, Ziyu Jiang, Victoria Knutson, Zehang Richard Li, Alana McGovern, Taylor Okonek, Johnny Paige, Katie Paulson, Andrea Riebler, Austin Schumacher, Aja Sutton, Jon Wakefield, Quinn White, Yunhan Wu, Jieyi Xu, Joshua Yang, Zihang Yu.

DHS: Ben Mayala, Trevor Croft.

MICS: Liliana Carvajal, Yadigar Coskun, Nazim Gashi, Ivana Bjelic, Attila Hancioglu, Stephanie Kauv, Bo Pedersen.

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

WHO: Charlton Callendar, Haidong Wang, Luhua Zhao.

WorldPop: Andy Tatem, Edson Utazi.

Gates: Ash Shah.

References i

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

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

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

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

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

References ii

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B 71*, 319–392.

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

Wu, Y., Z. R. Li, B. Mayala, H. Wang, P. Gao, J. Paige, G.-A. Fuglstad, C. Moe, J. Godwin, R. Donohue, T. Croft, and J. Wakefield (2021). Spatio-temporal modeling for admin-2 small-area estimation. Technical report, ICF International. DHS Spatial Analysis Reports No. 21.

Wu, Y. and J. Wakefield (2024). Modeling urban/rural fractions in low- and middle income countries. *Journal of the Royal Statistical Society, Series A*.