Bayesian SAE with an Emphasis on Low- and Middle-Income Countries: Lecture 5: Further Topics

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Motivation

In these slides we discuss:

- Nested spatial models.
- Model-based geostatistics.
- Variance modeling in area-level models.
- Unmatched linking modeling in area-level models.
- The SAR model.
- Space-time modeling.
- Spatial machine learning.

Nested Spatial Models

Nested Spatial Models

The BYM2 spatial models we use smooth the raw data both locally, through the ICAR term, and nationally, through the IID term.

Oversmoothing is a concern when data are sparse, particularly for rare outcomes.

Nigeria has 774 Local Government Areas (LGAs) at Admin2, and SAE at this level using DHS or MICS is often treacherous.

Recall that surveys are powered to produce reliable estimates for some indicators at Admin1 and, hence, we would like to "recover" the weighted estimates, rather than distort away from those.

To try to reduce overshrinkage at the Admin1 level, we may use a nested model.

Nested Area-Level Model

Let i denote Admin2 areas, $a[i]$ represent the Admin1 area within which Admin2 area i is contained, with $a = 1, \ldots, A$, Admin1 areas.

A nested Fay-Herriot area-level model has the form:

$$\begin{aligned} \widehat{\theta}_i^{\mathsf{w}} \mid \theta_i &\sim_{ind} \mathsf{N}(\theta_i, \widehat{V}_i) \\ \theta_i &= \underbrace{\alpha_{a[i]}}_{\text{Admin1 LEVEL}} + \underbrace{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}_{\text{COVARIATE MODEL}} + \underbrace{u_i}_{\text{SPATIAL RESIDUAL}} . \end{aligned}$$

The terms α_a are Admin1 level parameters that are included to effectively recover the weighted estimates.

The u_i terms then spatially smooth within each Admin1 area – they are nested random effects.

This model can be viewed as being consistent with the sampling scheme in which the Admin1 areas correspond to sampling strata.

Nested Cluster-Level Model

- Y_{ic} deaths from n_{ic} births in sampled cluster c in Admin2 area i.
- Overdispersed binomial model:

$$Y_{ic} \mid r_{ic} \sim \text{BetaBinomial}(n_{ic}, r_{ic}, d), \quad (1)$$

where r_{ic} is the risk of neonatal death in cluster c of area i.

- Nested model for cluster risk p_{ic}:

$$\frac{r_{ic}}{1 - r_{ic}} = \exp(\underbrace{\alpha_{a[c]}}_{\text{Admin1 LEVEL}} + \underbrace{\boldsymbol{x}_{ic}^{\mathsf{T}}\beta}_{\text{COVARIATE MODEL}} + \underbrace{u_i}_{\text{SPATIAL RESIDUAL}}).$$

The terms α_a are Admin1 level parameters that are included to effectively recover the weighted estimates.

- The u_i terms then spatially smooth within each Admin1 area.
- This model can be viewed as being consistent with the sampling scheme in which the Admin1 areas correspond to sampling strata.

An Example with Nested Models

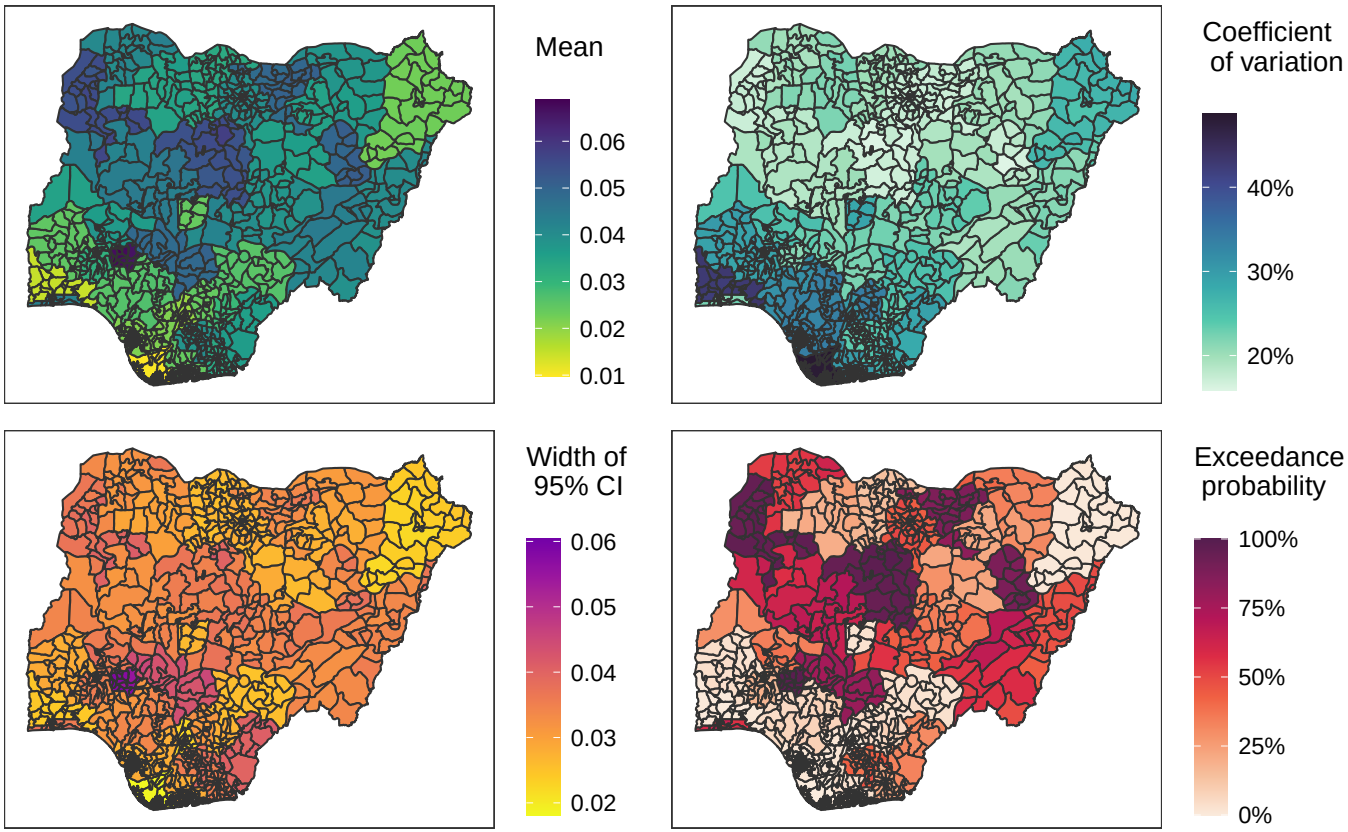


Figure 1: Fixed effects at Admin1, with nested BYM2 random effects at Admin2.

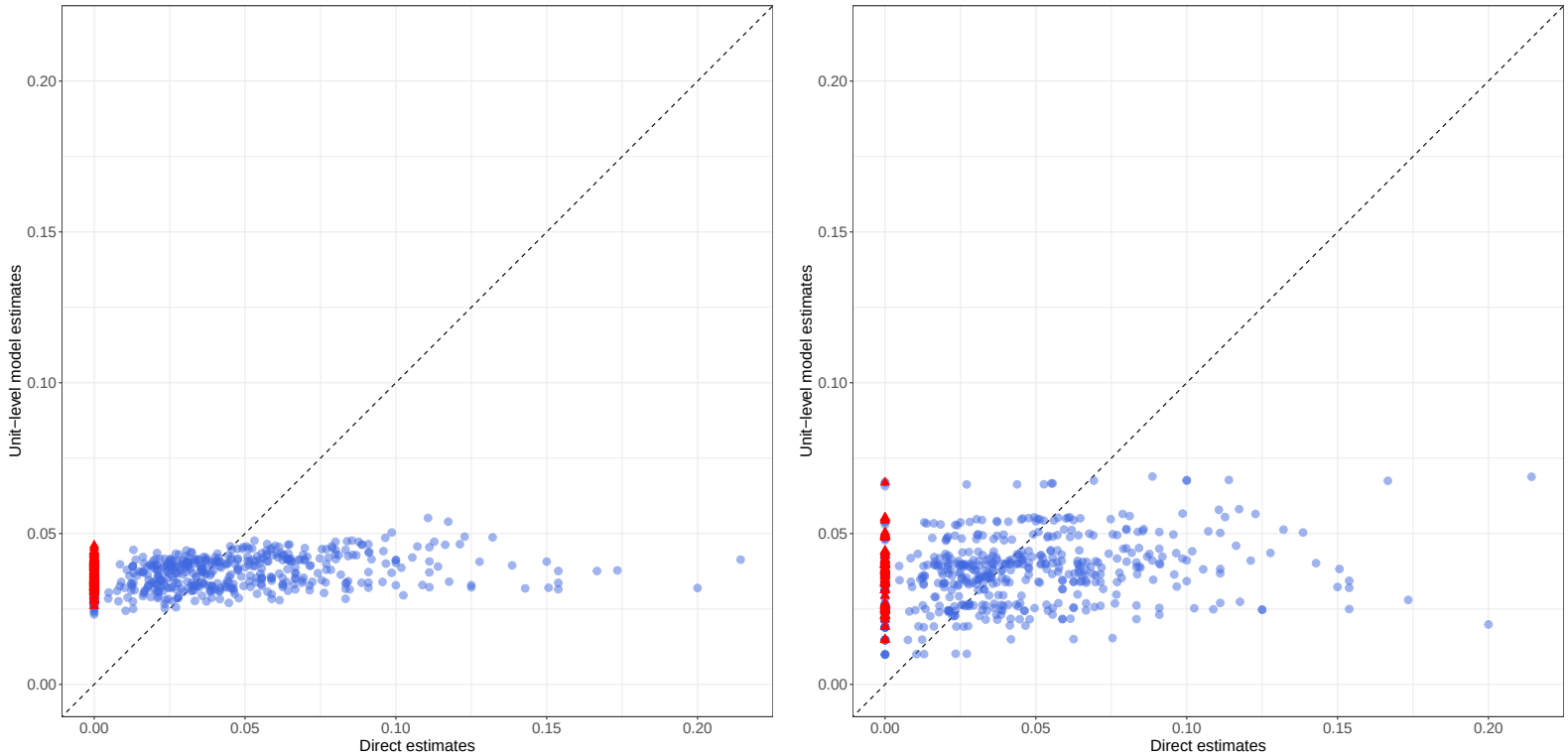


FIGURE 2: Estimates from non-nested (left) and nested (right) models, against weighted estimates. Notice the wider range in the right hand plot – the shinrkage has been reduced. But we don't pick up much variation within Admin1 areas, because the data are so sparse.

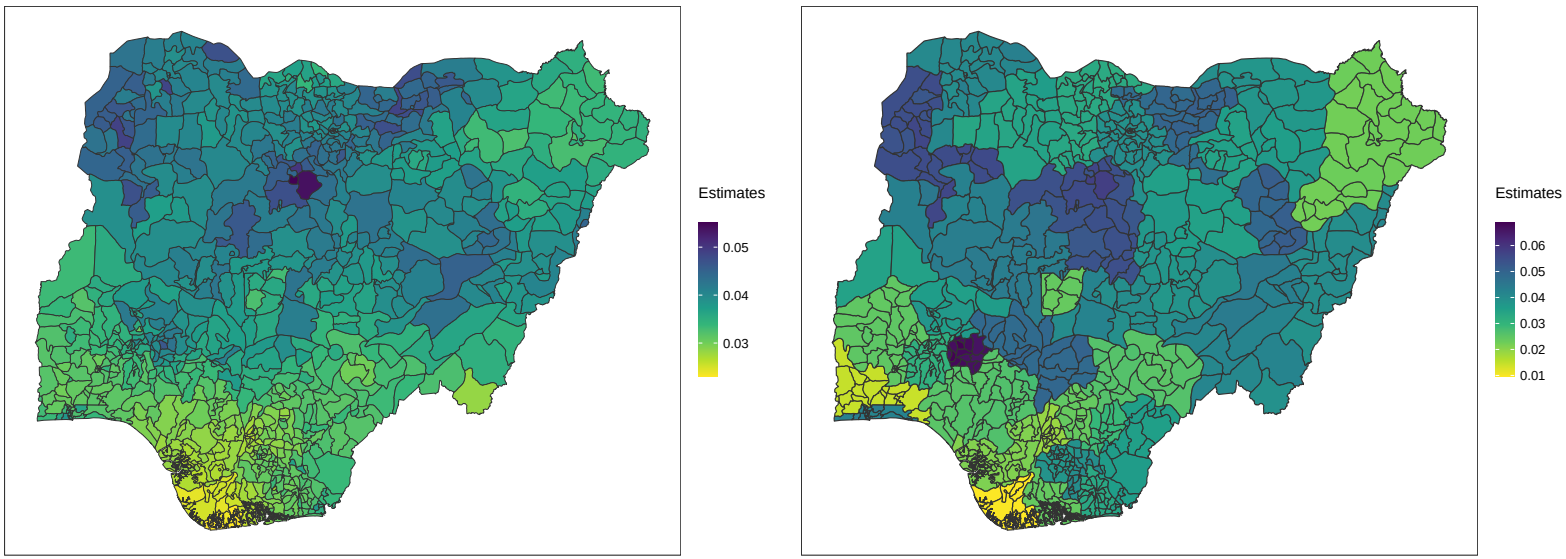


FIGURE 3: Mapped estimates from non-nested (left) and nested (right) models. Notice the larger scale in the right hand plot.

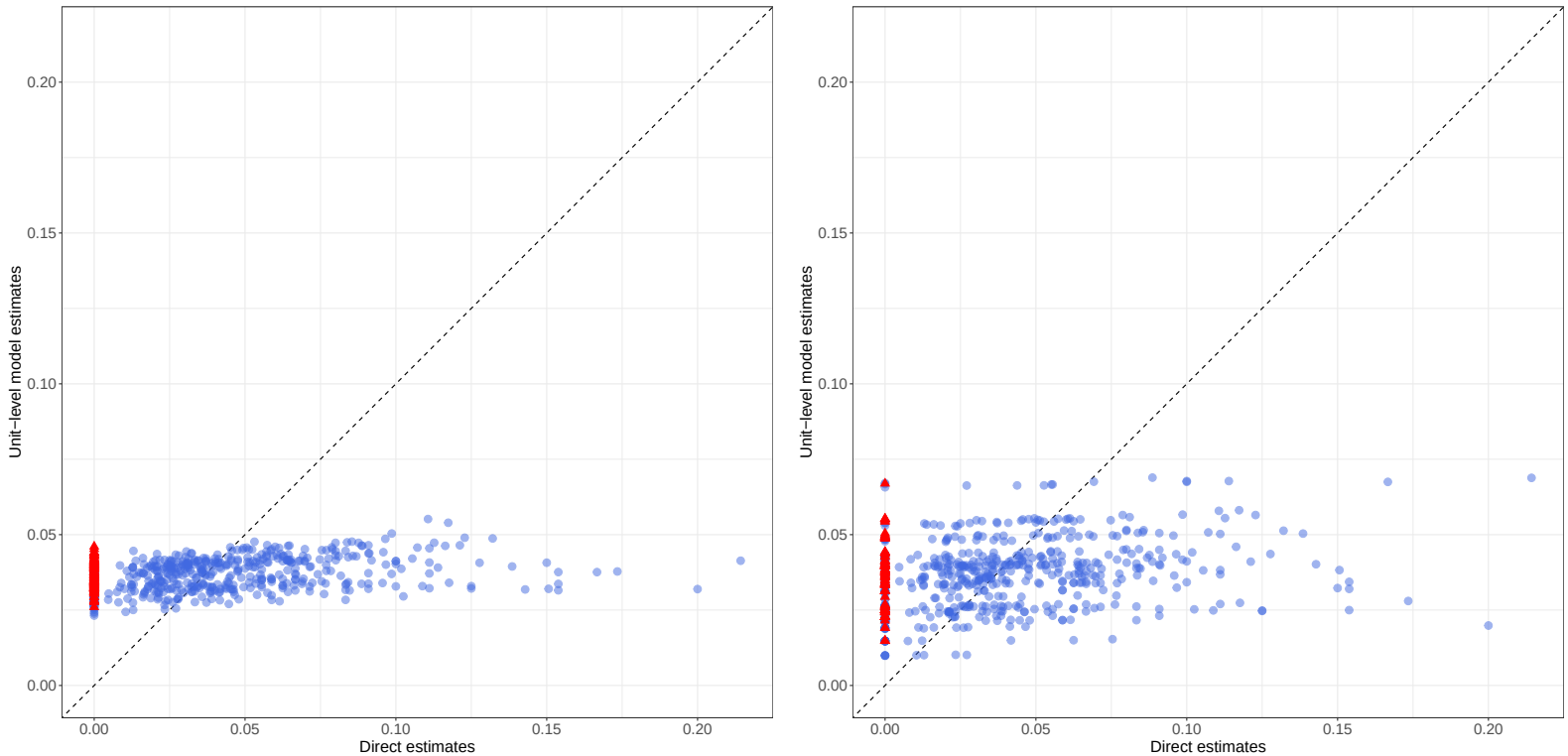


FIGURE 4: Estimates from non-nested (left) and nested (right) models, against weighted estimates. Notice the wider range in the right hand plot.

Model-Based Geostatistics

Model-Based Geostatistics (MBG)

- MBG models are (at first glance) unit-level models, with continuous spatial priors.
- The sampling model can be a binomial with a normal random effect (sometimes vaguely referred to as the nugget) on the logistic scale.
- Here we describe a Betabinomial model which is an example of an overdispersed binomial:

$$Y_{ic} \mid r_{ic}, d \quad \sim \quad \text{BetaBinomial}(n_{ic}, r_{ic}, d), \tag{2}$$

where $r_{ic} = r(\boldsymbol{s}_{ic})$ is the risk associated with location $\boldsymbol{s}_{ic}$.

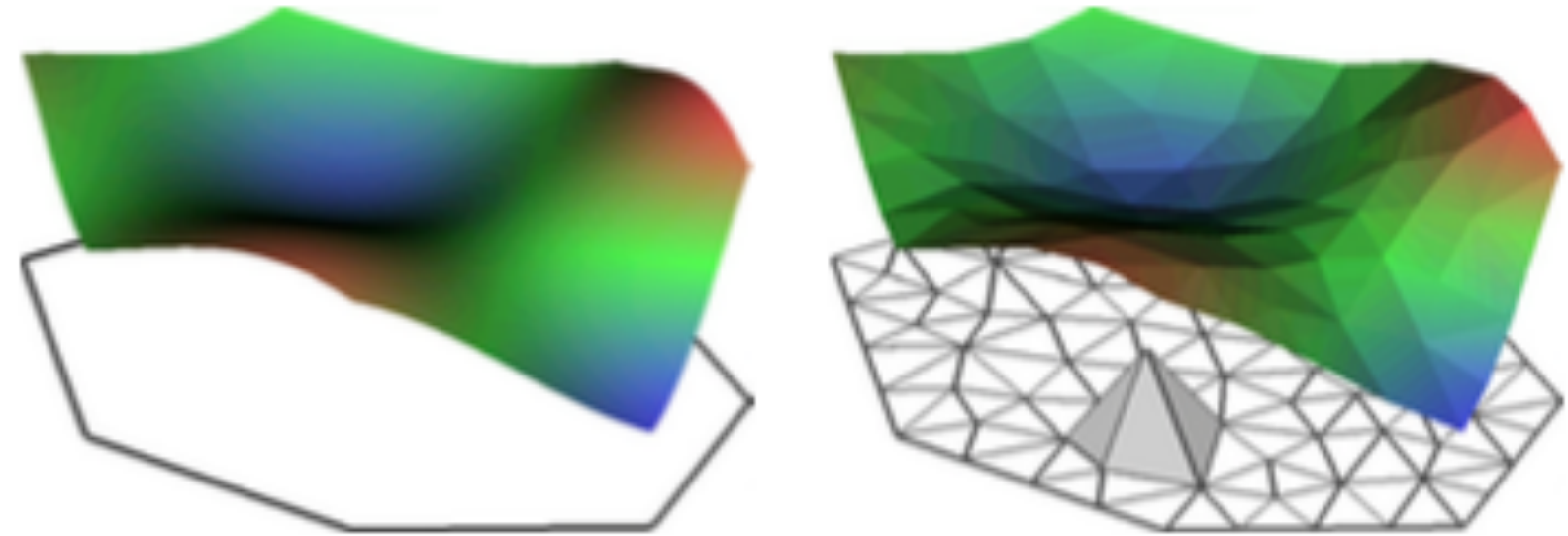
Model-Based Geostatistics (MBG)

- The latent model is:

$$\frac{r(\boldsymbol{s}_{ic})}{1-r(\boldsymbol{s}_{ic})} = \exp\Big(\underbrace{\alpha}_{\text{OVERALL LEVEL}} + \underbrace{\boldsymbol{x}(\boldsymbol{s}_{ic})^{\mathsf{T}}\boldsymbol{\beta}}_{\text{COVARIATE MODEL}} + \underbrace{\omega(\boldsymbol{s}_{ic})}_{\text{SPATIAL RESIDUAL}} \Big)$$

  where $\omega(\boldsymbol{s}_{ic})$ is:

  - Modeled as a realization from a Gaussian Random Field (GRF).
  - A Matérn correlation function is used, and implementation is via the Stochastic Partial Differential Equations (SPDE) approximation (Lindgren *et al.*, 2011) and implemented in `R-INLA`.
  - We can replace α with a fixed effect $\alpha_{a[i]}$ where a indexes Admin1 units – this may aid in preventing overshrinkage.
  - The Bayesian model is completed with priors on $\alpha, \boldsymbol{\beta}, d$ and the parameters of the GRF – for the latter penalized complexity (PC) priors are used – see the separate appendix on these.



(a) A continuous function. (b) A piecewise linear approximation.

Fig. 2. Piecewise linear approximation of a function over a triangulated mesh.

FIGURE 5: GMRF representation of a Markovian GRF, via triangulation, from Simpson *et al.* (2012b).

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

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

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

In practice, we can aggregate over a grid:

$$r_i \approx \sum_{g=1}^{G_i} F_{ig} \times \text{expit}(\alpha + \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{s}_{ig})\beta + \omega(\boldsymbol{s}_{ig})), \tag{3}$$

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

Constructing a predictive surface

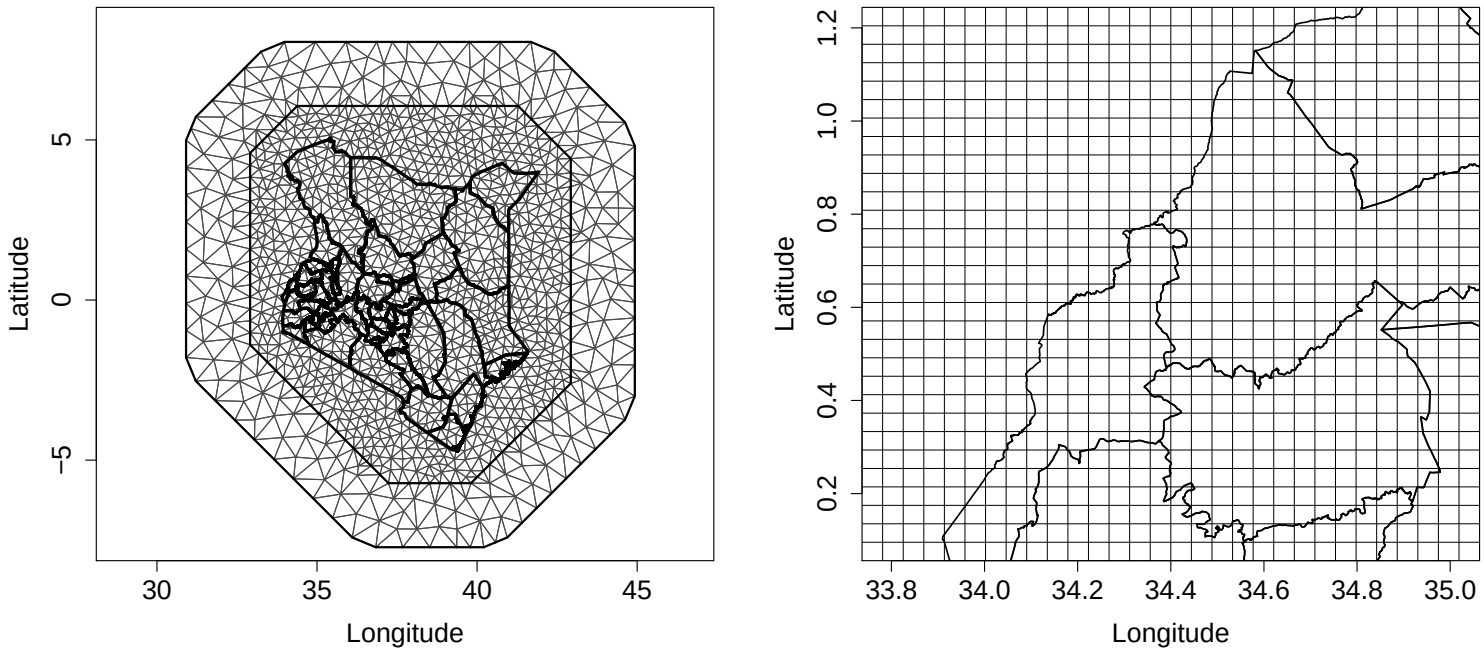


Figure 6: Inference via SPDE is carried out via the triangulation mesh (left), and then spatial predictions are calculated on a grid (right), as illustrated here for Kenya.

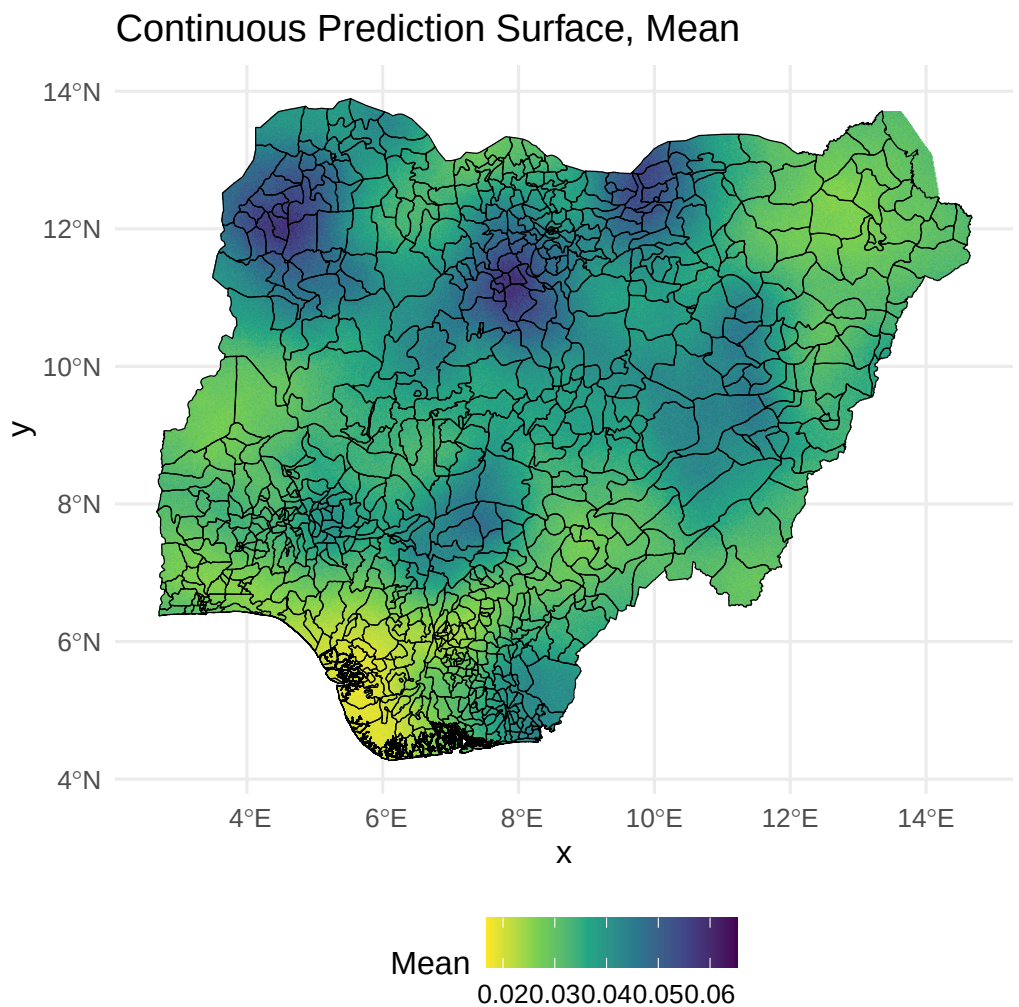


Figure 7: NMR continuous surface map for Nigeria, from 2018 DHS.

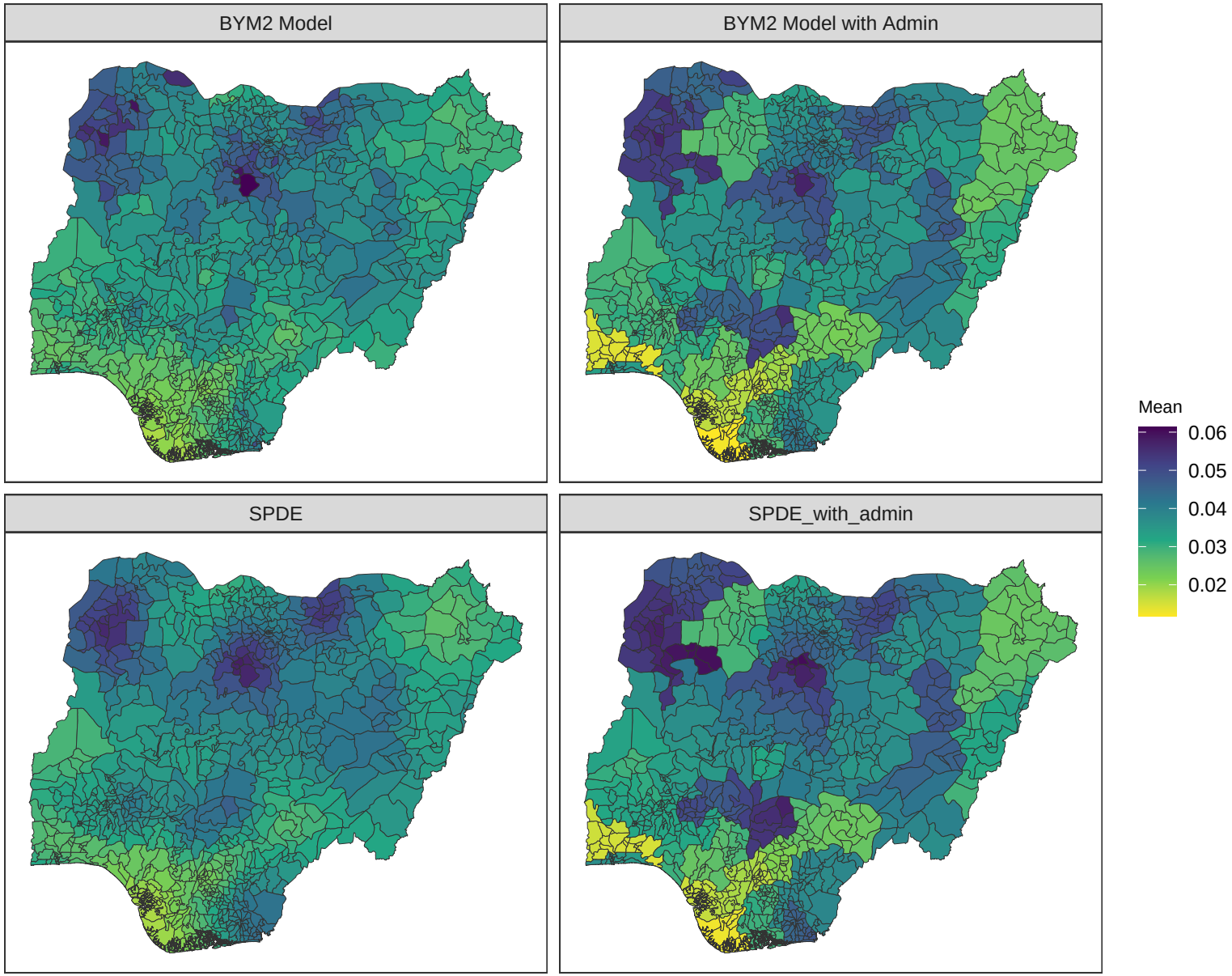


FIGURE 8: Comparison of aggregations from spatial models – BYM2 and GRF spatial models give quite similar area-level estimates.

Variance Fix for Fay-Herriot

Variance Instability

- The Fay-Herriot model is the most widely-used approach to SAE.
- But it requires as input the variance of the weighted estimate, $\widehat{V}_i$, and this may be unstable or unavailable for certain configurations of data.
- In general, to alleviate sampling variance difficulties, generalized variance functions (Wolter, 2007, Chapter 7) may be used.
- Such approaches leverage the mean-variance relationship between θ_i^{w} and $\widehat{V}_i^{1/2}$, and/or incorporate covariates (Otto and Bell, 1995; Mohadjer *et al.*, 2012; Franco and Bell, 2013; Liu *et al.*, 2014).
- We describe a simple approach that is implemented in our software.

Variance Modeling

- For the surveys in LMICs the variance formula is:

$$\widehat{V}(\widehat{p}_i^{\text{w}}) = \frac{C_i}{C_i - 1} \frac{1}{(t_i^{\text{w}})^2} \sum_{c \in S_i} \left[\sum_{k \in S_{ic}} w_{ick}(y_{ick} - \widehat{p}_{ic}^{\text{w}}) \right]^2 \tag{4}$$

- 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).

- Formula (4) breaks down if all outcomes in an area are all 0 or are all 1, or there is a single cluster in the area.

- We have a variance fix in `surveyPrev`, based on a weighted beta distribution.

- Alternatively, Gao and Wakefield (2023) jointly model weighted estimator and variance estimator.

In the Beginning...

Since the dawn of time, Bayesians have been fixing binomials with $y = 0/n$, e.g., Wilson (1927), combining binomial likelihood with Be(2,2) prior.

Be(2,2) = Phantom data with $z^{\text{PH}} = 2$ successes out of $m^{\text{PH}} = 4$ which gives the augmented estimator,

$$\widehat{p}^{\text{AUG}} = \frac{z + z^{\text{PH}}}{m + m^{\text{PH}}}.$$

How to transfer over to the survey setting?

Consider the pseudo-likelihood for a generic area,

$$L^{\text{PL}}(p) = \prod_{k \in S} \left[p^{y_k} (1-p)^{1-y_k} \right]^{w_k}.$$

Now define a pseudo prior, based on phantom observations and weights, $\{y_k, w_k, k \in S^{\text{PH}}\}$,

$$g^{\text{PH}}(p) = \prod_{k \in S^{\text{PH}}} \left[p^{y_k} (1-p)^{1-y_k} \right]^{w_k} [p(1-p)]^{-1}.$$

The Fix

This gives augmented posterior:

$$\pi^{\text{AUG}}(p) \propto \prod_{k \in S} \left[p^{y_k}(1-p)^{1-y_k} \right]^{w_k} \times \prod_{k \in S^{\text{PH}}} \left[p^{y_k}(1-p)^{1-y_k} \right]^{w_k}$$

Maximizing this augmented posterior gives estimator:

$$\widehat{p}^{\text{AUG}} = \frac{\sum_{k \in S} w_k y_k + \sum_{k \in S^{\text{PH}}} w_k y_k}{\sum_{k \in S} w_k + \sum_{k \in S^{\text{PH}}} w_k}.$$

The augmented set of observations, $k \in S^{\text{AUG}}$, can be used in estimators and variance formulas.

This is implemented as default for pre-modeled estimates at

https://sae4lmic.stat.uw.edu/

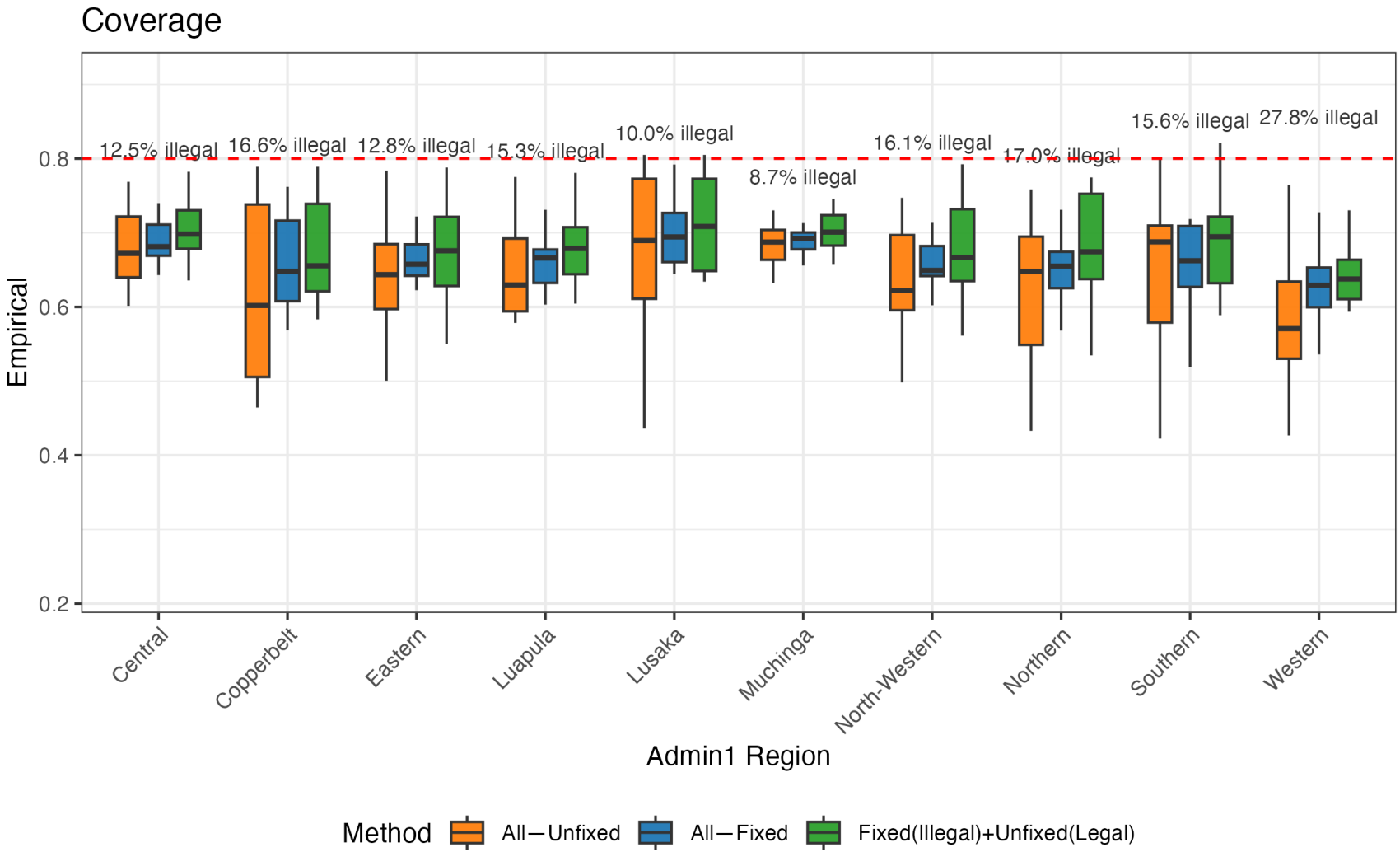


FIGURE 9: Empirical simulation coverage, mimicking DHS survey settings

Unmatched Linking Models

Unmatched Linking Models

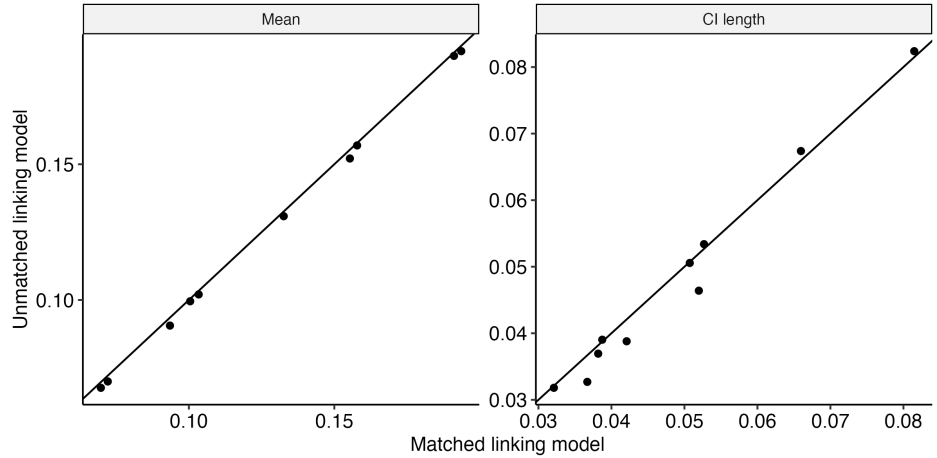
- In the context of modeling a prevalence, we previously modeled on the logit scale.
- But $\text{logit}(\hat{p}_i^{\text{w}})$ is not an unbiased estimator of the logit prevalence.
- To examine this, we consider the unmatched linking model:

$$\begin{aligned} \widehat{p}_i^{\,\text{w}} \mid p_i &\sim_{iid} \text{N}(p_i, \widehat{V}_i^{\star}), && (5) \\ \text{logit}(p_i) &= \alpha + u_i, \qquad i = 1, \ldots, m, && (6) \\ \boldsymbol{u} \mid \sigma_u^2, \phi &\sim \text{BYM2}(\sigma^2, \phi), && (7) \end{aligned}$$

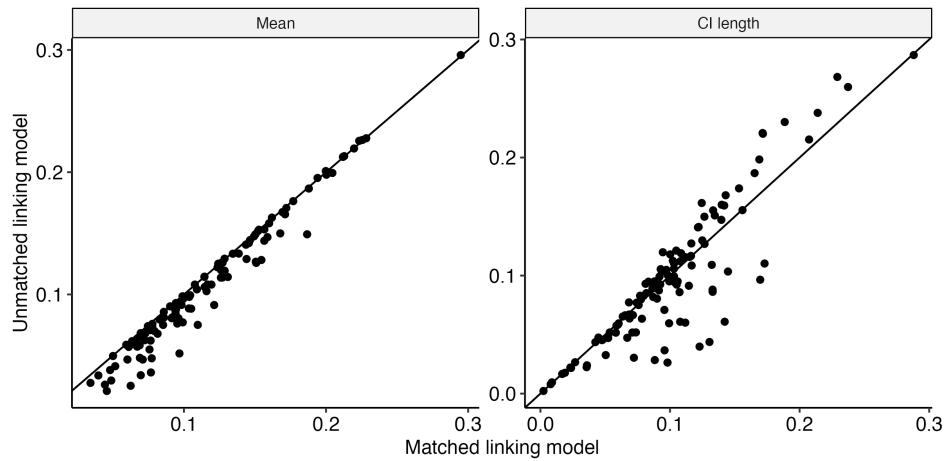
where $\widehat{V}_i^{\star}$ is the design-based variance estimate for $\widehat{p}_i^{\,\text{w}}$.

Unmatched Linking Models

- We fit this model to HIV prevalence data in adult women from the Zambia DHS, using the `surveyPrev` package.
- Figure 10 shows the resultant posterior mean and uncertainty estimates, plotted against estimates from the matched linking model.
- At the Admin1 level, we see strong agreement, which is reassuring, but there are systematic differences at the Admin2 level, as one might expect given the differences in the sampling models and the small sample sizes.



(a) Admin1 comparison.



(b) Admin2 comparison.

FIGURE 10: Comparison of posterior mean (left column) and 95% credible interval length (right column) under Fay-Herriot unmatched and matched linking models.

Unmatched Linking Models

- Via linearization, one can show that there is a positive bias when the logit is modeled for $p_i < 0.5$, and we see this in the bottom left panel of Figure 10.

- In Figure 11, we plot the weighted national estimate of female adult HIV prevalence, along with aggregated national estimates from the matched and unmatched BYM2 models (with aggregation fractions in both cases taken as normalized WorldPop population counts).

- We see that relative to the weighted estimate, the unmatched estimate is low, whereas the matched model shows greater agreement.

- The problem with the unmatched model is that there are many areas with low prevalence estimates and in these cases the true sampling distribution is skewed, which the logit matched model picks up more effectively.

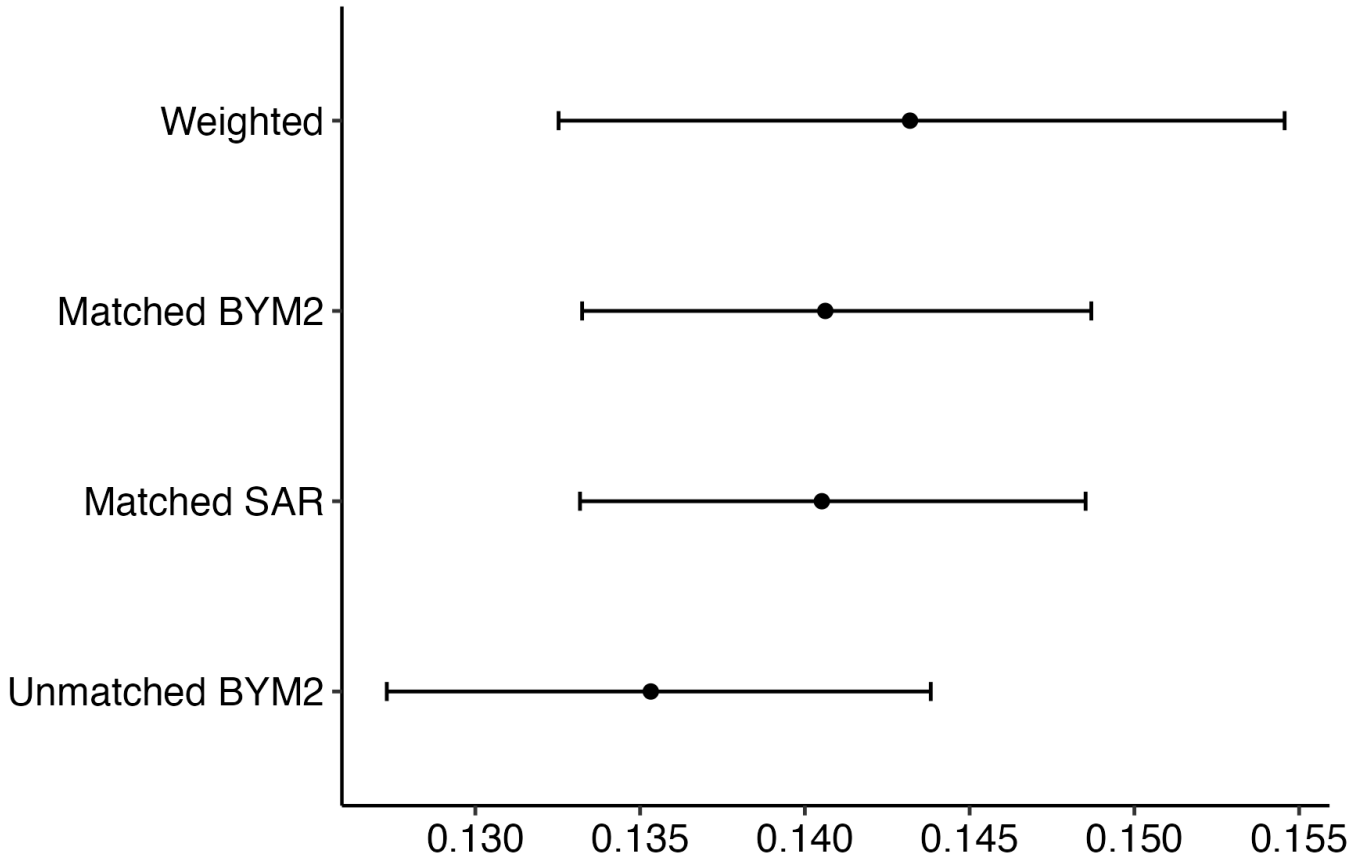


FIGURE 11: Comparison of weighted national estimate with estimates from various Fay-Herriot models, along with 95% uncertainty intervals.

Unmatched Linking Models

- Figure 12 shows the point estimates and asymptotic normal 95% confidence intervals for the prevalence, for the original and logit scales.
- On the original scale (corresponding to the unmatched model) there are many areas with intervals that include negative values (around 20% of the areas).
- The assumed sampling distribution places excess mass on values close to zero, which pulls down the resultant estimates and leads to the negative bias in the aggregated national estimate.
- We conclude that, in this example, the matched model is preferred, but further research on its general use is merited.

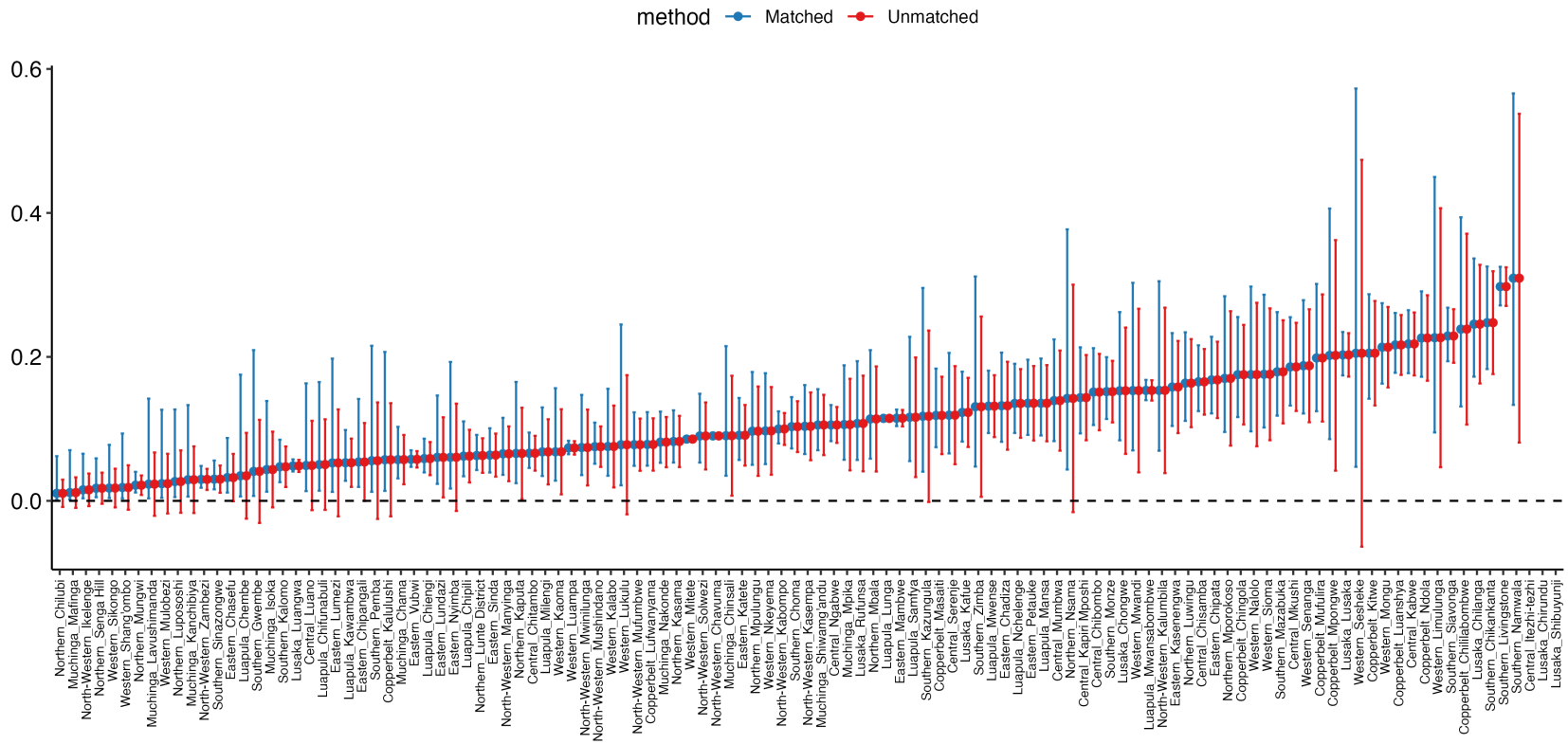


FIGURE 12: Assumed asymptotic sampling distributions, at Admin2, for unmatched and matched linking models.

The SAR Model

Simultaneous Autoregressive (SAR) Model

We have so far described the BYM2 model for spatial smoothing, which models spatial dependence via an intrinsic conditional autoregressive (ICAR) form.

We now describe, as an alternative, a simultaneous autoregressive (SAR) Fay-Herriot model – this model is also available in the `surveyPrev`) package.

It is also in the `sae` package (Marhuenda *et al.*, 2013), where inference is frequentist.

The SAR Model

The SAR model (Banerjee *et al.*, 2015) is given by,

$$y_i = \boldsymbol{x}_i^\top \beta + \sum_{j=1}^{n} B_{ij}(y_j - \boldsymbol{x}_j^\top \beta) + \epsilon_i,$$

for $i = 1, \ldots, m$, where

- the SAR weights B_{ij} have $B_{ii} = 0$,
- ϵ_i are independent, zero mean errors terms with variance $\widehat{V}_i$,
- Let $\text{var}(\epsilon) = \widehat{\boldsymbol{V}}$ be the $m \times m$ diagonal matrix with i-th element $\widehat{V}_i$.

Conditional autoregressive (CAR) models (including the ICAR) proceed by conditioning on neighbors while the SAR model, as it says on the tin, considers simultaneous modeling.

Wall (2004) contains an interesting comparison of CAR and SAR models.

In vector form:

$$\boldsymbol{y} - \boldsymbol{x}^{\mathsf{T}}\beta = \boldsymbol{B}(\boldsymbol{y} - \boldsymbol{x}^{\mathsf{T}}\beta) + \boldsymbol{\epsilon},$$

or

$$(\boldsymbol{I}_m - \boldsymbol{B})(\boldsymbol{y} - \boldsymbol{x}^{\mathsf{T}}\beta) = \boldsymbol{\epsilon},$$

or

$$\boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}}\beta + (\boldsymbol{I}_m - \boldsymbol{B})^{-1}\boldsymbol{\epsilon},$$

where the latter assumes that $\boldsymbol{B}$ has been chosen so that $\boldsymbol{I}_m - \boldsymbol{B}$ is invertible.

If $\boldsymbol{B} = \boldsymbol{0}$ we have an iid model.

So spatial dependence is induced through $(\boldsymbol{I}_m - \boldsymbol{B})^{-1}$.

Under normality of the errors

$$\boldsymbol{Y} \mid \beta, \boldsymbol{B} \sim \mathrm{N}\left(\boldsymbol{x}^{\mathsf{T}}\beta, [(\boldsymbol{I}_m - \boldsymbol{B})^{\mathsf{T}} \widehat{\boldsymbol{V}} (\boldsymbol{I}_m - \boldsymbol{B})^{-1}]^{\mathsf{T}} \right).$$

For a well-defined model we require $(\boldsymbol{I}_m - \boldsymbol{B})$ to be non-singular, which puts conditions on $\boldsymbol{B}$.

A common choice is

$$\boldsymbol{B} = \rho_s \boldsymbol{W},$$

for a spatial proximity matrix $\boldsymbol{W}$ with elements 1 for neighbors and 0 otherwise.

The SAR Model: Some Technical Details

Let $\lambda_{(1)}, \ldots, \lambda_{(n)}$ be the ordered eigenvalues of $\boldsymbol{W}$.

Then, $(\boldsymbol{I}_m - \boldsymbol{B})$ will be invertible if

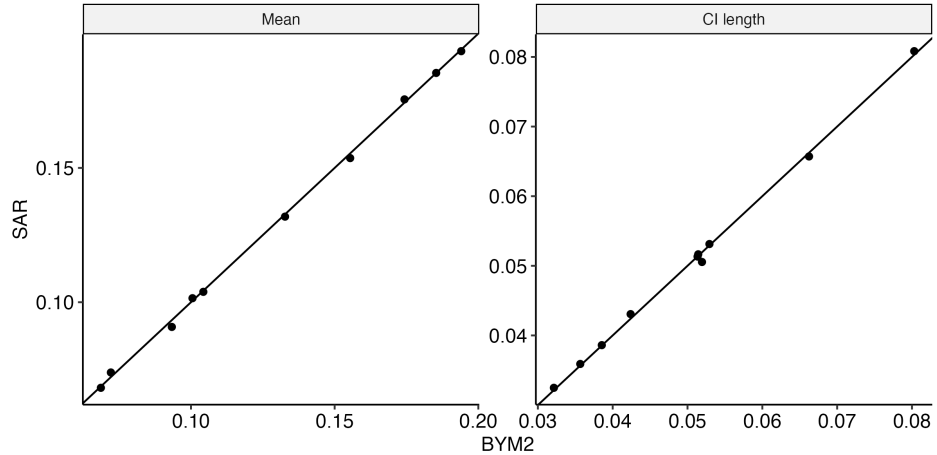
$$\frac{1}{\lambda_{(1)}} < \rho_s < \frac{1}{\lambda_{(n)}}.$$

If the row sums of $\boldsymbol{W}$ are standardized to 1 then $\lambda_{(n)} = 1$ and $\lambda_{(1)} \leq -1$ so $\rho_s < 1$ but may be less than -1.

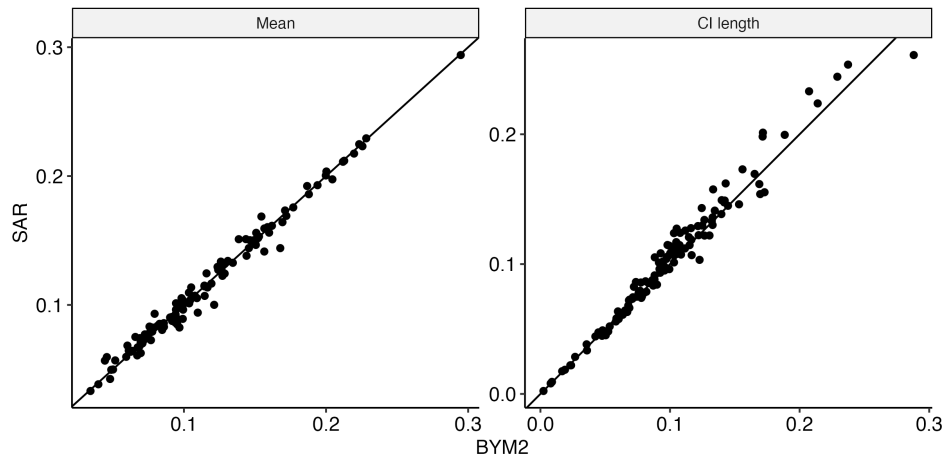
For more discussion, see Banerjee et al. (2015, Section 4.4) and Schabenberger and Gotway (2005, Section 6.2.2.1).

The SAR Model

- We compare simultaneous autoregressive (SAR) and BYM2 Fay-Herriot models using the Zambia HIV DHS data.
- Specifically, we fit models at the Admin1 and Admin2 levels, using INLA (Gómez-Rubio *et al.*, 2020), obtaining almost identical point and interval estimates for the two models, as we see in Figure 13.
- Figure 11 illustrates that the aggregate national estimates under SAR and BYM2 models are also very similar.
- The SAR model and the unmatched model described by (5)–(7), are included in the `surveyPrev` package (Dong *et al.*, 2025).



(a) Admin1 comparison.



(b) Admin2 comparison.

FIGURE 13: Comparison of posterior mean (left column) and 95% credible interval length (right column) under Fay-Herriot models with SAR and BYM2 spatial random effects.

Spatio-Temporal Modeling

Main Effects and Interactions

- To motivate space-time models, when space is modeled discretely, we consider simple two-way factor models.
- Suppose we have a univariate continuous response Y.
- Now consider modeling as a function of two factors, A and B say, with $i = 1, \ldots, m$ and $j = 1, \ldots, T$ indexing the levels.
- A main effects only model takes the form

$$\mathrm{E}[Y_{ij} \mid \alpha, \eta_i, \delta_j] = \alpha + \eta_i + \tau_j.$$

- Interpretation: η_i is the effect of being at level i for factor A, regardless of the level assumed by B, and τ_j is the effect of being at level j for factor B, regardless of the level assumed by A, i.e. there is no interaction.

Main Effects and Interactions

- An interaction model adds a set of interaction parameters

$$\mathsf{E}[Y_{ij} \mid \alpha, \eta_i, \tau_j, \delta_{ij}] = \alpha + \eta_i + \tau_j + \delta_{ij}.$$

- Interpretation: δ_{ij} is the additional effect, beyond $\eta_i + \tau_j$ of being simultaneously at levels *i* and *j* of factors A and B.

- If the factor correspond to nominal levels (e.g., a factor for color with 2 levels: ”red”, ”blue”) then we would not expect similarity between adjacent levels.

- In a space-time context the “factors” space and time have an “ordering” and we might expect similarity in the response at neighboring levels.

Main Effects Model

- First, consider the space-time model,

$$\begin{aligned} Y_{it} \mid \theta_{it} &\sim \mathrm{N}(\theta_{it}, \widehat{V}_{it}) \\ \theta_{it} &= \mathrm{expit}(\alpha + \eta_i + S_i + \omega_t + \tau_t) \end{aligned}$$

- Components:
  - Y_{it} is the weighted estimate with design-based variance $\widehat{V}_{it}$.
  - Unstructured spatial term $\eta_i \sim_{iid} \mathrm{N}(0, \sigma_e^2), i = 1, \ldots, m$.
  - Smooth spatial term $[S_1, \ldots, S_m]$, e.g., from an ICAR model.
  - Unstructured temporal term $\omega_t \sim_{iid} \mathrm{N}(0, \sigma_\omega^2), t = 1, \ldots, T$.
  - Smooth temporal term $[\tau_1, \ldots, \tau_T]$ smooth in time, e.g. follows an RW1 or RW2 model.
- Notice there is no interaction between space and time – the spatial effects are constant across time and temporal trends are constant across space.

Space-Time Interaction Models

- Knorr-Held (2000) considered the model:

$$\theta_{it} = \alpha + \eta_i + S_i + \omega_t + \tau_t + \delta_{it},$$

with η_i, S_i, ω_t, δ_t are as in the main effects only model.

- Four different models for the interaction δ_{it}:
  - Type I: Independent interaction.
  - Type II: Temporal trends differ between areas but don't have spatial structure.
  - Type III: Spatial patterns differ between time points but don't have temporal structure.
  - Type IV: Temporal trends differ between areas but more likely to be similar for adjacent areas.

We describe the Type IV model only, since it is the most appealing in a prevalence mapping context.

Inseparable Space-Time Interaction Models

- Type IV: Temporal trends differ between areas but more likely to be similar for neighboring areas.
- This will often be the most realistic model if interactions are present.
- In the case of a RW2 temporal model and an ICAR spatial model, the joint distribution can be written:

$$p(\boldsymbol{\delta} \mid \sigma_\delta^2) \propto \exp\left(-\frac{1}{2\sigma_\delta^2}\sum_{t=3}^{T}\sum_{i\sim j}(\delta_{it} - \delta_{jt} - 2\delta_{i,t-1} + 2\delta_{j,t-1} + \delta_{i,t-2} - \delta_{j,t-2})^2\right)$$

- The Knorr-Held (2000) models are implemented in the `SUMMER` package.

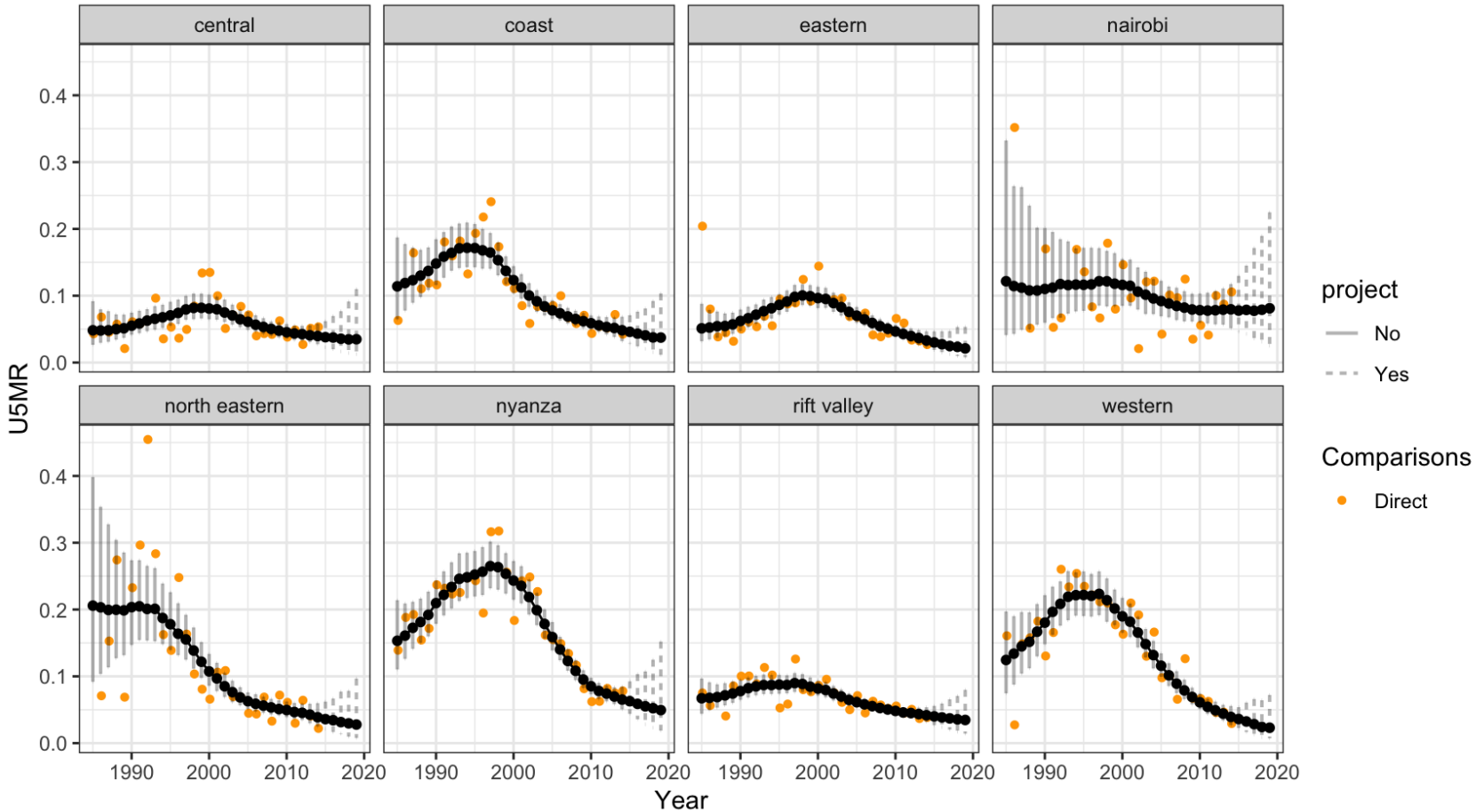


FIGURE 14: Weighted estimates and smoothed fits over time for 8 provinces of Kenya.

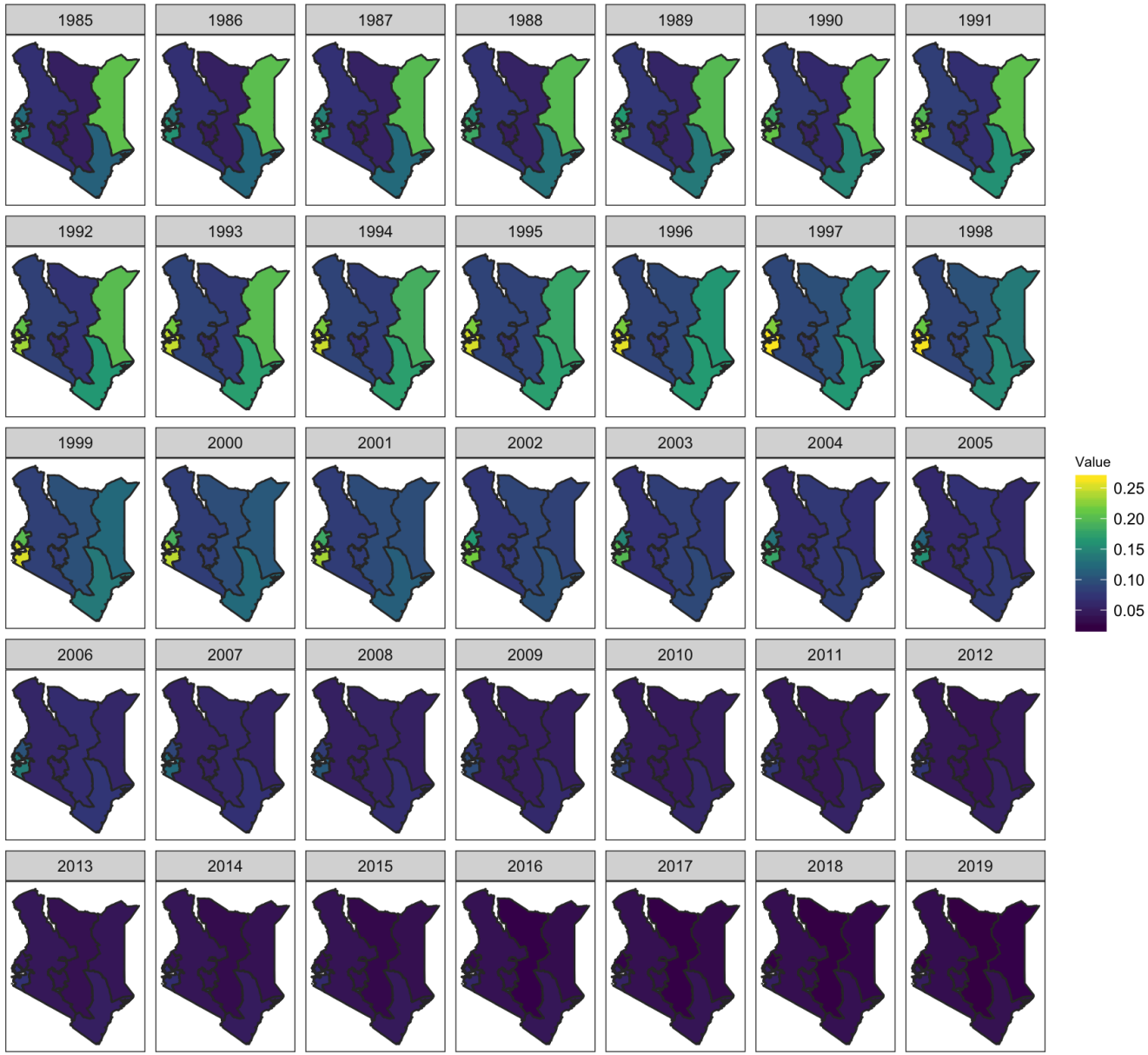


FIGURE 15: Maps of smoothed estimates over time for 8 provinces of Kenya.

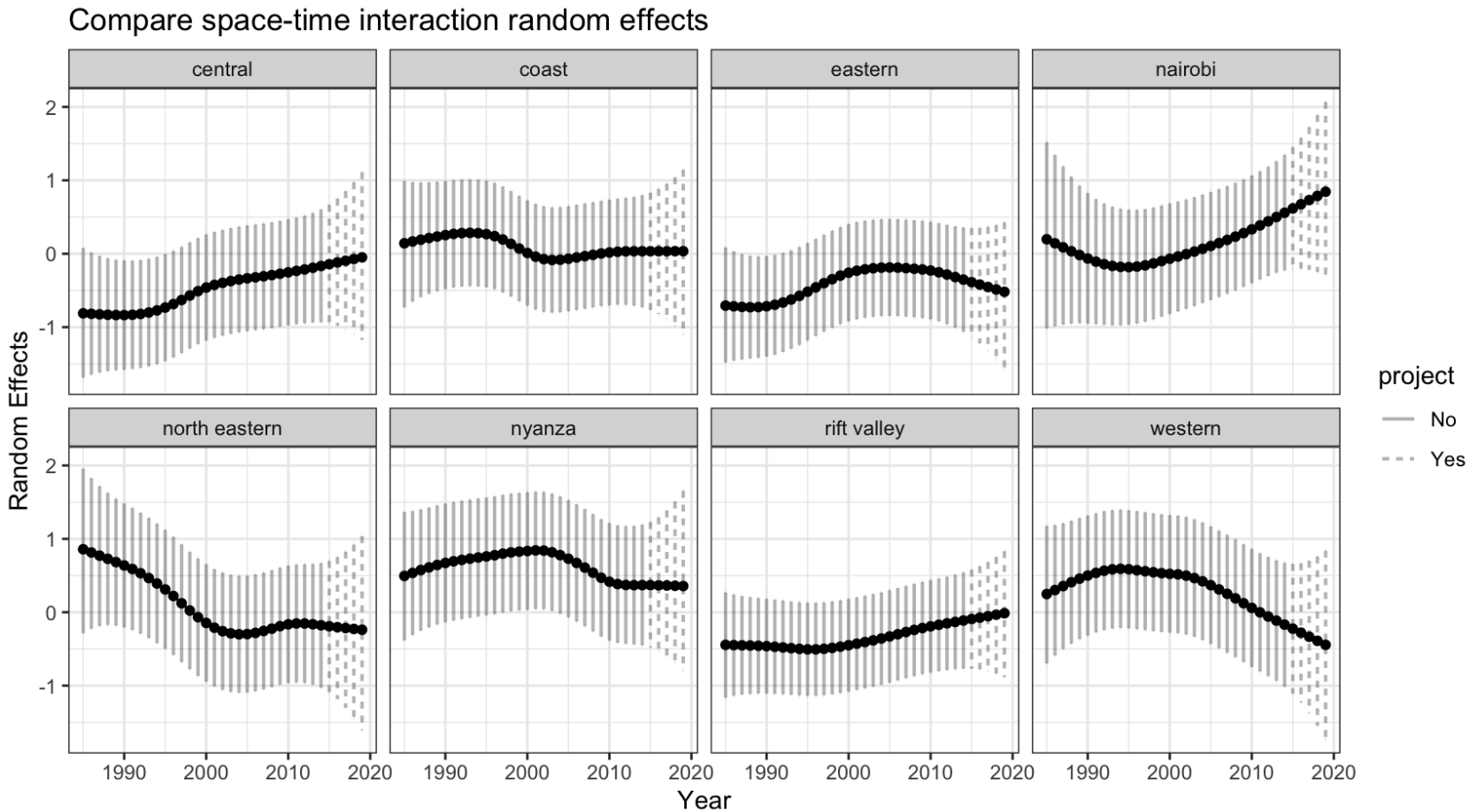


FIGURE 16: Space-time interactions δ_{it} for 8 provinces of Kenya.

Space-Time Fay-Herriot Model (Li *et al.*, 2019)

- The space-time Fay-Herriot model has been used for 35 African countries to estimate U5MR in Admin1 regions, by year.
- Data enter at the 5-year level (to give stable variances, $\widehat{V}_{it}$), but the RWs are defined on the 1-year scale.
- Data:
  - 121 DHS in 35 countries
  - 1.2 million children
  - 192 million child-months
- It took around 2.5 hours to obtain estimates for all countries – separate models for each country.
- This area-level Fay Herriot model is very reliable for examining Admin1 subnational variation, but the direct estimates are often unreliable for Admin2 estimation.

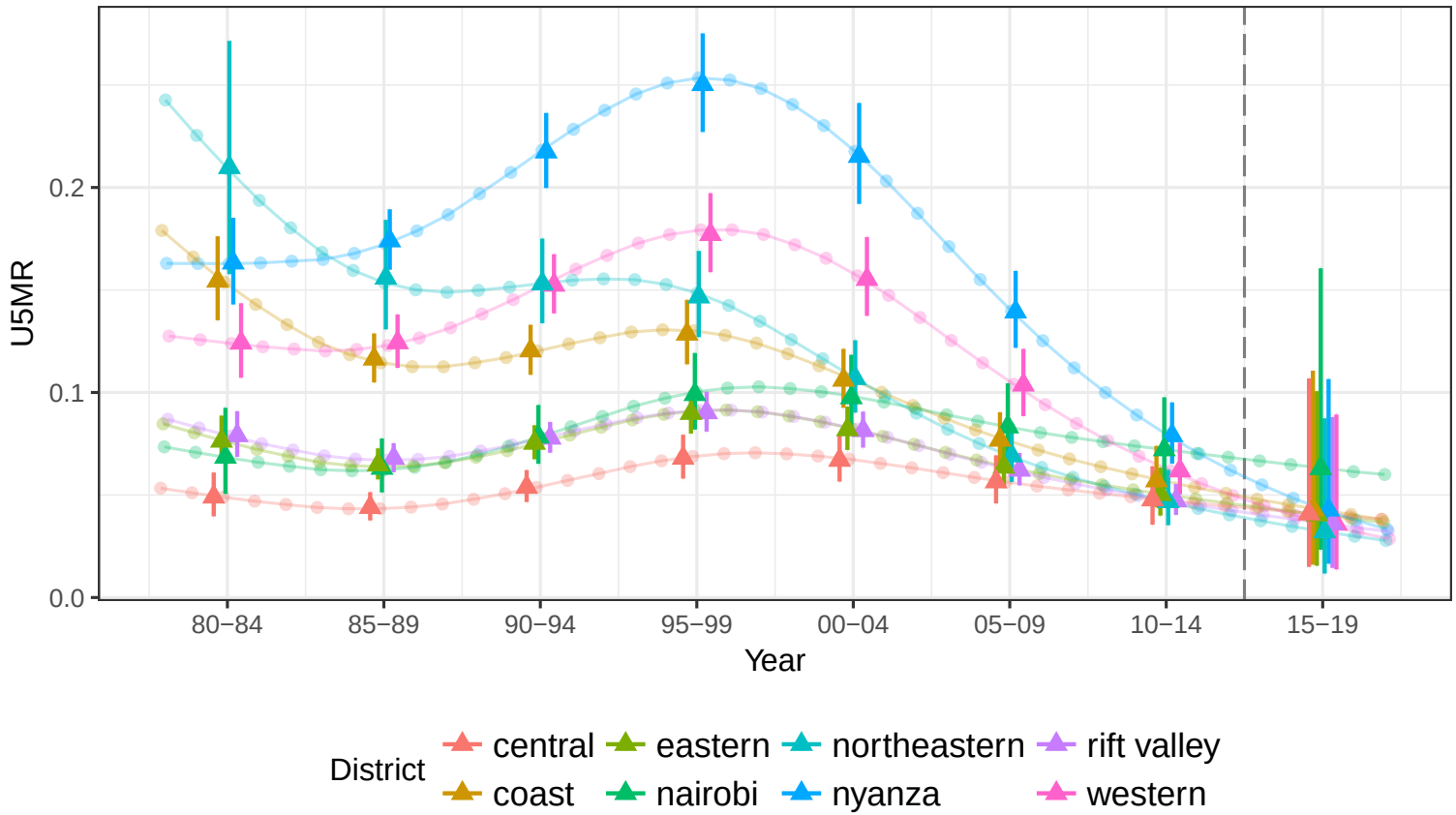


FIGURE 17: Posterior median estimates for Kenya districts.

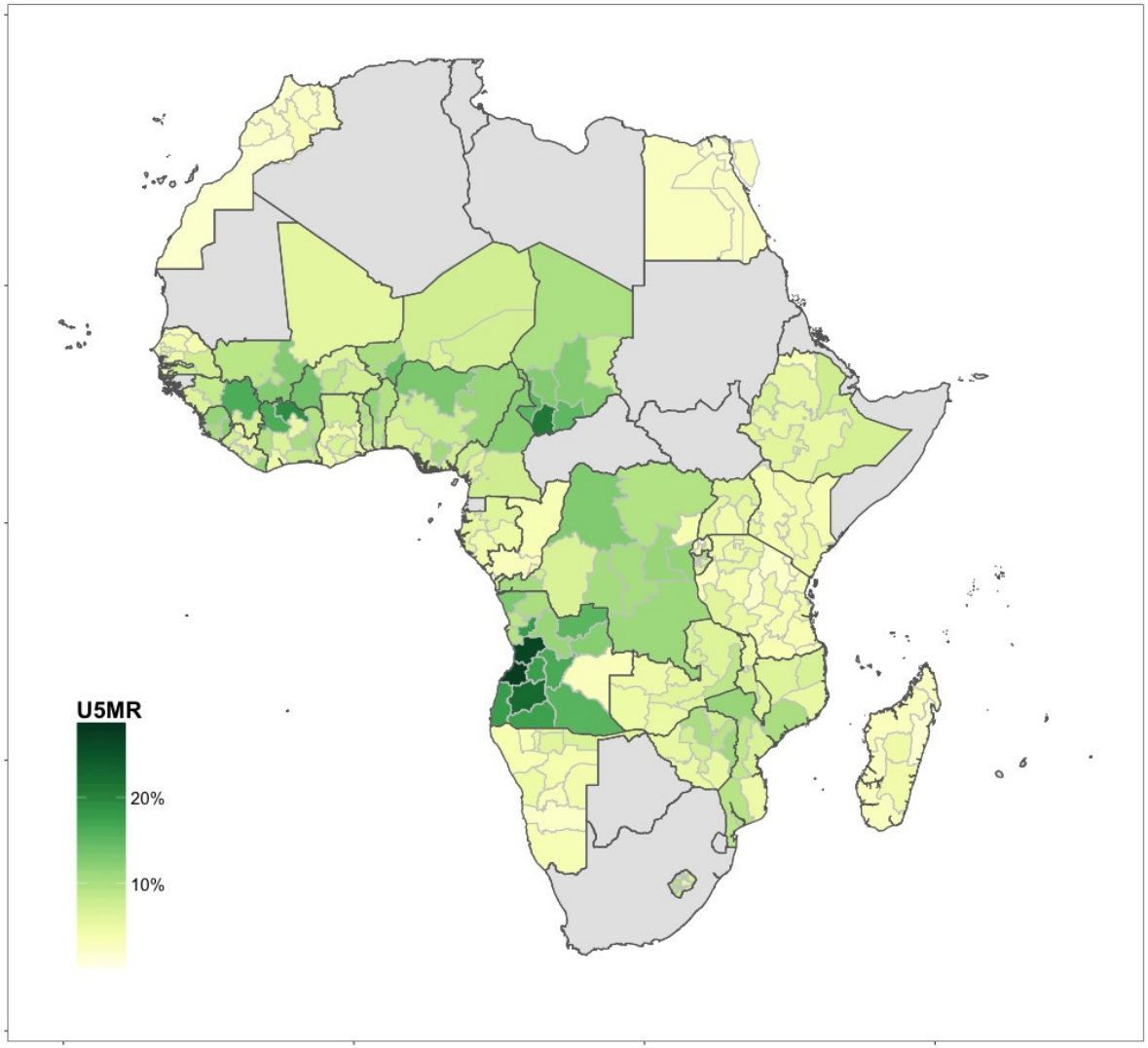


FIGURE 18: Predictions of U5MR for 2015, in 35 countries of Africa.

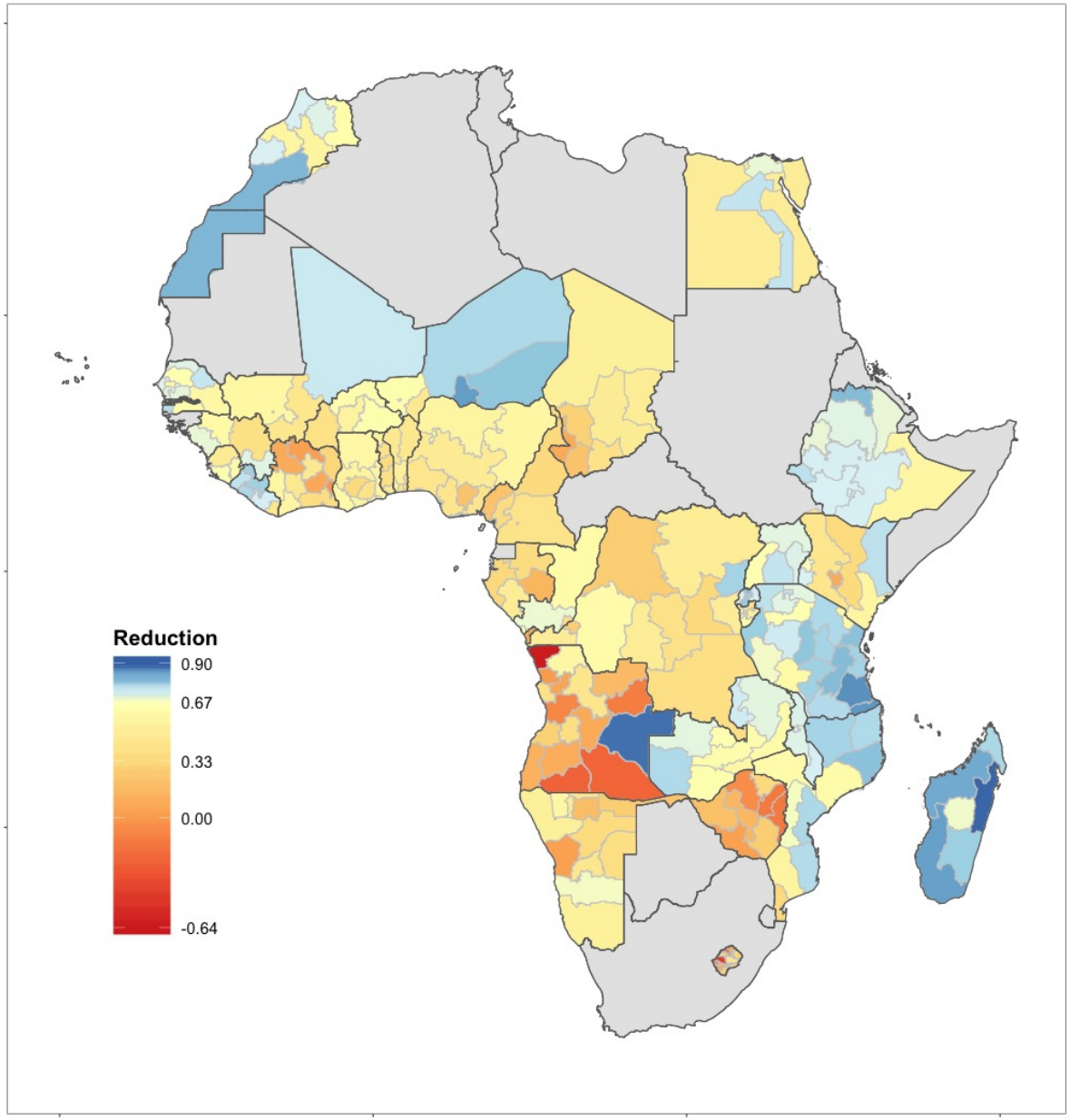


FIGURE 19: Percent reduction from 1990 to 2015, in 35 countries of Africa.

Spatial Machine Learning

The Pleasure and the Pain

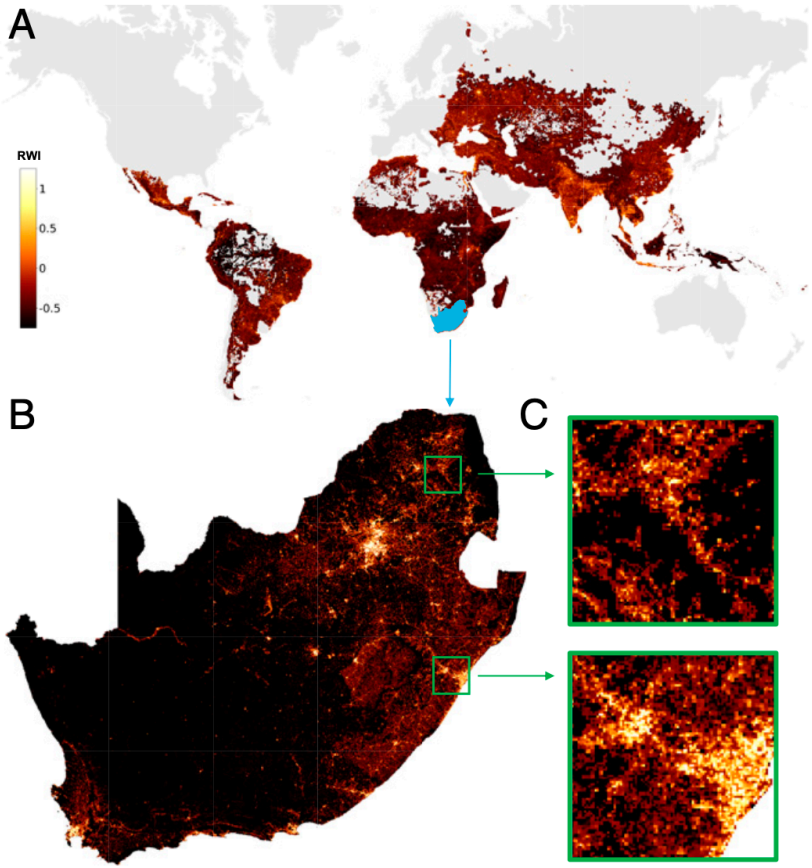


Fig. 1. Microestimates of wealth for all low- and middle-income countries. (*A*) Estimates of the relative wealth of each populated 2.4-km gridded region of all 135 LMICs. Interactive version is available at http://www.povertymaps.net. (*B* and *C*) Enlargements show (*B*) the countries of South Africa and Lesotho and (*C*) the regions around Durban and Polokwane.

Figure 20: From Chi et al. (2022).

ML approaches are very appealing:

- Flexible mean function to allow interactions and nonlinearities.
- Can be used with a large number of covariates (census, facilities, climatic, satellite imagery, social media, cell phone,...).
- Automatic model selection (hmmm).

But:

- How to tune algorithm?
- How to account for weights, spatial dependence.
- How to obtain measures of uncertainty?
- Lots of suggestions, but where is the theory?

Uncertainty Estimation in Spatial ML

- Many different resampling algorithms have been used: the bootstrap, jackknife, conformal prediction, prediction powered inference (PPI),...
- Model-assisted estimator:

$$\widehat{T}_i = \sum_{k \in U_i} \widehat{\mu}(\boldsymbol{x}_{ik}) + \sum_{k \in S_i} \frac{y_{ik} - \widehat{\mu}(\boldsymbol{x}_{ik})}{\pi_{ik}}$$

- Leverages population-level information.
- Links with many literatures including AIPW and casual inference, popular in the survey literature since the 1970s (Cassel *et al.*, 1976; Särndal *et al.*, 1992).
- Rediscovered to form the basis for PPI (Angeloulous et al, 2023).
- Very large sample size are needed, and the same is true for conformal prediction.
- Gao and Wakefield (2024) use model-assisted models in SAE.
- Dharamshi *et al.* (2025) use U/V statistic representations of model-assisted to obtain exact variance estimates that account for uncertainty in $\widehat{\mu}$.

Discussion

Discussion

- There are many discrete spatial smoothing models, and it may not make much difference which is used, in terms of predictions and uncertainty – so pick the one you're comfortable with.
- Continuous spatial modeling, as used in MBG, is appealing, but aggregation is much more difficult, and computation is greater.
- Continuous spatial models may be combined with temporal models but it gets quite complex and computation is complex, but models can be fitted in INLA, see Wakefield *et al.* (2019).
- Beware machine-learning approaches that are magical to behold, but most approaches ignore the survey aspects, and do not give valid uncertainty estimates (the vanilla bootstrap is not valid for many ML techniques, especially in a spatial setting).
- Smoothing over space and time can be done within INLA, but computational expense increases steeply if a third variable (such as age) is added.

Final Comments

- Extensive materials can be found at

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

- For space only, stat with `surveyPrev`, for space-time `SUMMER`.

- If you can't find something you need, write to me at `jonno@uw.edu`

References

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

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Dharamshi, A., Gao, P., and Wakefield, J. (2025). Exact variance estimation for model-assisted survey estimators using u-and v-statistics. *arXiv preprint arXiv:2502.11032*.

Dong, Q., Li, Z. R., Wu, Y., Boskovic, A., and Wakefield, J. (2025). *surveyPrev: Mapping the Prevalence of Binary Indicators using Survey Data in Small Areas*. R package version 1.0.0.

Franco, C. and Bell, W. R. (2013). Applying bivariate binomial/logit normal models to small area estimation. In *Proceedings of the American Statistical Association, Survey Research Section*, pages 690–702.

Franco-Villoria, M., Ventrucci, M., Rue, H., *et al.* (2019). A unified view on bayesian varying coefficient models. *Electronic Journal of Statistics*, **13**, 5334–5359.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, **114**, 445–452.

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.

Gao, P. A. and Wakefield, J. (2024). Smoothed model-assisted small area estimation of proportions. *Canadian Journal of Statistics*, **52**(2), 337–358.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2020). Bayesian model averaging with the integrated nested Laplace approximation. *Econometrics*, **8**, 23.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555–2567.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**, 79–86.

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**, e0210645.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, **73**, 423–498.

Liu, B., Lahiri, P., and Kalton, G. (2014). Hierarchical Bayes modeling of survey-weighted small area proportions. *Survey Methodology*, **40**, 1–13.

Marhuenda, Y., Molina, I., and Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics and Data Analysis*, **58**, 308–325.

Miller, D. L., Glennie, R., and Seaton, A. E. (2020). Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, **25**, 1–16.

Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., and de Kerckhove, W. V. (2012). Hierarchical Bayes small area estimates of adult

literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, pages 55–63.

Otto, M. C. and Bell, W. R. (1995). Sampling error modelling of poverty and income statistics for states. In *American Statistical Association, Proceedings of the Section on Government Statistics*, pages 160–165.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysi*. CRC press.

Simpson, D., Lindgren, F., and Rue, H. (2012a). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, **23**, 65–74.

Simpson, D., Lindgren, F., and Rue, H. (2012b). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, **1**, 16–29.

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

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

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

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.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, **121**, 311–324.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**, 434–449.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

Wolter, K. (2007). *Introduction to Variance Estimation, Second Edition*. Springer Science and Business Media.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92**, 921–936.

Appendix: Stochastic Partial Differential Equations

Overview of SPDE

The rationale behind the SPDE approach is:

1. GPs with Matérn covariance functions are the solutions to a particular stochastic partial differential equation.

2. These SPDEs can be numerically solved using an approximation, based on the finite element method, that produces a sparse precision matrix. Because of the sparseness, this approach is implemented in `INLA`.

In the following, we draw on Lindgren *et al.* (2011); Simpson *et al.* (2012a,b); Krainski *et al.* (2018); see also Miller *et al.* (2020) for a reasonably readable account – in particular, the .

Details of SPDE

There are two main parts to the SPDE approach to implementing GP models: represent the GP using an SPDE and solving using the finite element method.

The first part is based on a result of Whittle (1954) in which he showed that Matérn GPs can be represented as the solution to the SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2} \tau f(\boldsymbol{s}) = W(\boldsymbol{s}), \tag{8}$$

where

- $\Delta = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}$ is the Laplacian.
- $\alpha = \upsilon + d/2 > 0$ where υ is the smoothing parameter and d is the dimension of $\boldsymbol{s}$.
- $W(\boldsymbol{s})$ is white noise with unit variance.
- τ^2 is related to the variance.

Details of SPDE

The solution to (8) corresponds to a stationary GP for f with a Matérn covariance function:

$$\text{cov}[f(\boldsymbol{s}_i), f(\boldsymbol{s}_j)] = \frac{\sigma^2}{\Gamma(\upsilon + 1/2)\sqrt{4\pi}\kappa^{2\upsilon}2^{\upsilon-1}}(\kappa \mid\mid \boldsymbol{s}_i - \boldsymbol{s}_j \mid\mid)^{\upsilon} K_{\upsilon}(\kappa \mid\mid \boldsymbol{s}_i - \boldsymbol{s}_j \mid\mid)$$

where $K_{\upsilon}(\cdot)$ is a modified Bessel function of the second kind, $\sigma^2 > 0$ is a variance parameter, $\kappa > 0$ is a scale parameter and $\upsilon > 0$ is a smoothness parameter.

The variance is related to τ^2 via,

$$\sigma^2 = \frac{\Gamma(\upsilon)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\upsilon}\tau^2}$$

Markovian in Continuous Space

If α is an integer then the GRF has the continuous Markov property (Whittle, 1954), which is the key for implementation.

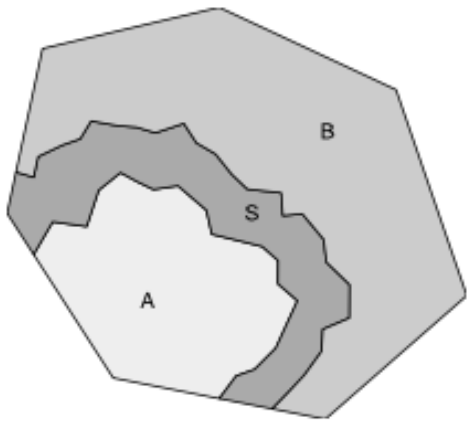


Figure 21: From Simpson *et al.* (2012b) who present this figure as an illustration of the spatial Markov property. If $\{x(\boldsymbol{s}) : \boldsymbol{s} \in A\}$ is independent of $\{x(\boldsymbol{s}) : \boldsymbol{s} \in B\}$ given $\{x(\boldsymbol{s}) : \boldsymbol{s} \in S\}$, then the field $x(\boldsymbol{s})$ has the spatial Markov property.

A bit of intuition on the SPDE

For $d = 1$ imagine we discretize time with equally intervals h:

$$-\frac{d^2 f}{ds^2} \approx \frac{1}{h^2}[2f(s) - f(s-h) - f(s+h)],$$

which is a SAR model

$$2f(s_i) - f(s_{i-1}) - f(s_{i+1}) \sim \mathrm{N}(0, h^2\sigma^2).$$

So if we let the spacing $h \to 0$ we get

$$-\Delta f(s) = -\frac{d^2 f}{ds^2} \sim W(s).$$

Similarly, we can discretize space on a grid with a mesh that has side length h, then

$$\begin{aligned} -\frac{d^2 f}{d\boldsymbol{s}^2} &\approx \frac{1}{h^2}[2f(s_1, s_2) - f(s_1 - h, s_2 - h) - f(s_1 - h, s_2 + h) \\ &- f(s_1 + h, s_2 - h) - f(s_1 + h, s_2 + h)] \end{aligned}$$

So SPDEs are the limit of a certain type of SAR model.

Matérn members that satisfy the continuous Markov property

In 1D (and other odd dimensions) $\nu = 1/2, 3/2, \ldots$, satisfy the continuous Markov property:

- Thin plate splines correspond to $\nu = 1$.
- The exponential covariance model corresponds to $\nu = 1/2$.

Matérn members that satisfy the continuous Markov property

For 2D (and other even dimensions), the continuous Markov property is satisfied for $\nu = 1, 2, 3, \ldots$:

- Thin plate splines correspond to $\nu = 1$.
- The exponential covariance model corresponds is not Markovian in 2D.

Discretizations of space lead to straightforward definitions of being Markovian.

Not so easy in continuous space, but intuitively as below, how can we use this for computation?

The SPDE Method

So we know that certain Matérn fields have the continuous Markov property, but unfortunately we cannot immediately write down a sparse precision matrix and use the Rue/Held tricks for an efficient implementation.

In the lingo, not every approximation to the precision operator will be sparse; in particular the approximation you'd get by evaluating the covariance matrix and inverting it won't usually be sparse.

Similarly the approximation you get using the Karhunen-Loéve expansion[1] won't be sparse (Simpson *et al.*, 2012b).

Another way of seeing this: if you have a GMRF on a grid and you make a new multivariate Gaussian by locally averaging the GRMF, the resulting Gaussian doesn't have a sparse precision matrix

[1] a representation of a stochastic process as an infinite linear combination of orthogonal functions, analogous to a Fourier series representation of a function on a bounded interval

Finite Element Approach

We need to construct a solution to the SPDE, in the case where the observations are not observed on a regular grid, but are irregularly spaced.

We want to find an approximation,

$$f(\boldsymbol{s}) \approx \sum_{k=1}^{m} w_k \phi_k(\boldsymbol{s})$$

that best approximates the SPDE.

In the approximation the w_k are random (multivariate Gaussian) and $\phi_k(s)$ are a set of basis functions, $k = 1, \ldots, m$.

Lindgren *et al.* (2011) describe how to carefully choose the basis functions to preserve the sparse structure of the resulting precision matrix – this is done using the finite element method, a common technique for solving sets of partial differential equations.

Stochastic Partial Differential Equations (SPDEs)

- For simplicity take $\alpha = 2$ and $d = 2$.
- As a reminder, we are trying to find the $f(\boldsymbol{s})$ that is the solution to the SPDE

$$(\kappa^2 - \Delta)\tau f(\boldsymbol{s}) = W(\boldsymbol{s}). \tag{9}$$

- Any solution to (9) also satisfies, for any suitable $\phi(\boldsymbol{s})$,

$$\int_{\mathbb{R}^2} (\kappa^2 - \Delta) f(\boldsymbol{s}) \tau \phi(\boldsymbol{s}) \, d\boldsymbol{s} = \int_{\mathbb{R}^2} \phi(\boldsymbol{s}) \, dW(\boldsymbol{s}). \tag{10}$$

Stochastic Partial Differential Equations (SPDEs)

- In the FEM method the smooth test function $\phi(\boldsymbol{s})$ is replaced by a set of m well-chosen bases, the finite elements).
- This is the Delauney triangulation which gives functions $\phi_k(\boldsymbol{s})$ that are piecewise linear functions that are equal to 1 on vertex k and 0 on all other vertices – this is the triangular mesh.
- The limited support is key to producing an approximation that has a sparse precision matrix.
- The finite element test functions lead to the linear equation

$$(\kappa^2 \widetilde{\boldsymbol{C}} + \boldsymbol{G})\widetilde{\boldsymbol{f}} \stackrel{d}{=} \mathrm{N}(\boldsymbol{0}, \widetilde{\boldsymbol{C}})$$

  where the matrices are given by

  - $[\widetilde{\boldsymbol{C}}]_{i,j} = \int \phi_i(\boldsymbol{s})\phi_j(\boldsymbol{s})\, d\boldsymbol{s}$
  - $[\widetilde{\boldsymbol{G}}]_{i,j} = \int \bigtriangledown\phi_i(\boldsymbol{s}) \cdot \bigtriangledown\phi_j(\boldsymbol{s})\, d\boldsymbol{s}$

Stochastic Partial Differential Equations (SPDEs)

- These integrals can be explicitly calculated (Lindgren *et al.*, 2011) and are only non-zero if vertex i is a neighbor of vertex j – hence, these matrices are sparse (and symmetric positive definite), so very convenient for manipulation.

- This leads to

$$\widetilde{\boldsymbol{f}} \stackrel{d}{=} \mathrm{N}(\boldsymbol{0}, \widetilde{\boldsymbol{Q}}^{-1}),$$

where

$$\widetilde{\boldsymbol{Q}} = (\kappa^2 \widetilde{\boldsymbol{C}} + \boldsymbol{G})^{\mathsf{T}} \widetilde{\boldsymbol{C}}^{-1} (\kappa^2 \widetilde{\boldsymbol{C}} + \boldsymbol{G}).$$

Stochastic Partial Differential Equations (SPDEs)

- Unfortunately $\widetilde{\boldsymbol{C}}$ is not a diagonal matrix and so $\widetilde{\boldsymbol{Q}}$ is not sparse.
- Lindgren *et al.* (2011) replace $\widetilde{\boldsymbol{C}}$ by the diagonal matrix $\boldsymbol{C}$ that has on its diagonal the row sums of $\widetilde{\boldsymbol{C}}$.
- This gives the approximating GMRF as

$$\boldsymbol{f} \stackrel{d}{=} \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}),$$

  where

$$\boldsymbol{Q} = (\kappa^2 \boldsymbol{C} + \boldsymbol{G})^{\mathsf{T}} \boldsymbol{C}^{-1} (\kappa^2 \boldsymbol{C} + \boldsymbol{G})$$

- This gives a sparse precision matrix, so that computation can be carried out efficiently.

The SPDE Method

- This method is available within the `INLA` package – see Section 2.2.2 of Krainski *et al.* (2018).
- They walk through a numerical example showing the constituent matrices, and the details of the `INLA` implementation, including a description of the `inla.mesh.fem` function that constructs the finite elements.

Appendix: Penalized Complexity (PC) Priors

Base models

- Over the last 40 years, with the advent of MCMC, and then INLA, the range of complex models that can be fitted is vast, and intoxicating.
- This invites the inclination to try to account for every possible nuance and dependence in the data, by adopting flexible mean functions and complex random effects specifications – but then we fall into the third of Dante's layers of hell (gluttony) – more conventionally referred to as overfitting.
- To control against overfitting we would like to use priors that are "pulling us back" (from the brink of hell) to simpler models – this leads to the vague idea of a base model.

Base models

Definition 1: (Simpson *et al.*, 2017) *For a model component with density $\pi(\boldsymbol{s} \mid \xi)$ controlled by a flexibility parameter ξ, the base model is the "simplest" model in the class. For notational clarity, we will take this to be the model corresponding to $\xi = 0$. It will be common for ξ to be nonnegative. The flexibility parameter is often a scalar, or a number of independent scalars, but it can also be a vector-valued parameter.*

Examples:

- Example: Gaussian random effects model: let $\boldsymbol{s} \mid \xi$ be multivariate Gaussian with zero mean and precision matrix $\tau \boldsymbol{I}$ and we take $\xi = 1/\tau$. The base model therefore puts all mass at $\xi = 0$.
- Example: In the BYM2 model, we can take no spatial variation, i.e., $\phi = 0$, as the base model.
- Example: In an AR1(ρ) model we can take as base model either $\rho = 0$ (no temporal structure) or $\rho = 1$ (a constant function over time).

Principles for PC prior construction

The method is described in the context of four principles:

Principle 1: Occam's razor[2]: We should prefer simpler models until the data tell us that we should go with the more complex model. The simpler model is the base model, and our prior should penalize deviations from it.

[2] "Entities should not be multiplied without necessity", attributed to English Franciscan friar William of Ockham (c. 1287–1347)

Principles for PC prior construction

Principle 2: Measure of complexity: Complexity is based on on the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951):

$$\text{KLD}(\pi(\boldsymbol{s} \mid \xi), \pi(\boldsymbol{s} \mid \xi = 0)) = \int \pi(\boldsymbol{s} \mid \xi) \log\left(\frac{\pi(\boldsymbol{s} \mid \xi)}{\pi(\boldsymbol{s} \mid \xi = 0)}\right) \, d\boldsymbol{x}$$

The KLD measures the average distance between two distributions, where in our context the average is with respect to the full model.

Normal example: The KLD between $\text{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\text{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is

$$\text{KLD}(f_1, f_0) = \frac{1}{2}\left[\text{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\text{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - p - \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right]$$

Three of these are measuring distance between the densities.

PC prior construction

We want the prior to have high mass in areas of the parameter space where replacing the more complex model by the base model will not lead to much information loss.

One way of standardizing the KLD is to take the distance between the two densities as,

$$d = d(\pi(\boldsymbol{s} \mid \xi), \pi(\boldsymbol{s} \mid \xi = 0)) = \sqrt{2\text{KLD}(\pi(\boldsymbol{s} \mid \xi), \pi(\boldsymbol{s} \mid \xi = 0))}.$$

In the context of PC priors, $d = d(\xi)$ is a measure of the complexity of model $\pi(\boldsymbol{s} \mid \xi)$, when compared to model $\pi(\boldsymbol{s} \mid \xi = 0)$.

PC prior construction

The procedure places a prior on $d = d(\xi)$, in such a way that models close to the base model are favored.

Principle 3: Constant rate penalization: The deviation from the base model, as measured by d, is assumed to satisfy a memoryless property, to give an exponential prior on d:

$$\pi(d) = \lambda \exp(-\lambda d).$$

This prior satisfies:

$$\frac{\pi(d+\delta)}{\pi(d)} = r^{\delta},$$

for $\delta \geq 0$, with $r = e^{-\lambda}$ – the multiplicative change in the density does not depend on d.

This induces a prior on ξ via

$$\pi(\xi) = \lambda \exp(-\lambda d(\xi)) \left| \frac{\delta d(\xi)}{\delta \xi} \right|.$$

If d has upper bound then a truncated exponential distribution used.

Linking PC and Jeffreys' prior

Simpson *et al.* (2017) link the PC prior to Jeffreys' prior:

$$\text{KLD}(\pi(\boldsymbol{s} \mid \xi), \pi(\boldsymbol{s} \mid \xi = 0)) = \frac{1}{2} I(0) \xi^2 + \text{higher order terms},$$

where $I(\xi)$ is Fisher's information for ξ.

The PC prior behaves like:

$$\pi(\xi) = I(\xi)^{1/2} \exp(-\lambda m(\xi)) + \text{higher order terms},$$

for $\lambda\xi$ close to zero, and

$$m(\xi) = \int_0^{\xi} \sqrt{I(s)}\, ds.$$

So close to the base model, the PC prior is a tilted Jeffreys' prior for $\pi(\boldsymbol{s} \mid \xi)$.

PC prior construction

Principle 4: User-defined scaling: The final part is to select λ, and this is done by having the user specify numbers U and α to satisfy,

$$\Pr(Q(\xi) > U) = \alpha,$$

where $Q(\xi)$ is an interpretable transformation of the flexibility parameter ξ, U is some quantile with associated probability α.

The PC prior is invariant to parameterization, since the prior is defined on d, and then transformed to ξ.

It's a bit confusing, as we start with a prior on the distance, which implies a prior for ξ, and the way we calibrate the latter is by reparameterizing to a more meaningful parameter $Q(\xi)$.

Example: Gaussian random effects model

- Consider the Gaussian random effects model,

$$\boldsymbol{s} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_1),$$

with $\boldsymbol{\Sigma}_1 = (\tau \boldsymbol{R})^{-1}$ where $\boldsymbol{R}$ is a fixed matrix, and $\boldsymbol{\Sigma}_0 = \mathbf{0}$.

- We include the improper case here, where $\boldsymbol{R}$ does not have full rank.

- In this case $\xi = 1/\tau$ and it can shown that

$$\mathrm{KLD} =\!\!\rightarrow \frac{p}{2}\frac{\tau_0}{\tau},$$

as $\tau_0 \to \infty$ (Simpson *et al.*, 2017, Appendix A.1).

Example: Gaussian random effects model

- Hence, the distance between the complex and the base model is,

$$d(\tau) = \sqrt{p\tau_0/\tau} \tag{11}$$

and with an exponential prior of rate $\widetilde{\lambda}$ for d gives a type-2 Gumbel distribution for τ:

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\frac{\lambda}{\sqrt{\tau}}\right),$$

for $\tau > 0, \lambda > 0$.

- This implies an exponential prior of rate λ for σ (which we knew from (11) anyway).

- We calibrate in terms of this standard deviation: $\Pr(1/\sqrt{\tau} > U) = \alpha$, which (because the prior is exponential) implies

$$\lambda = -\log(\alpha)/U.$$

Choosing U and α

- For simplicity, consider the univariate random effect $x \mid \tau \sim_{iid} \mathrm{N}(0, 1/\tau)$ with type-2 Gumbel prior $\pi(\tau \mid \lambda)$.
- The above prescription requires choosing α and U such that $\Pr(\sigma > U) = \alpha$.
- Specifying the prior on the standard deviation $\sigma = 1/\sqrt{\sigma}$, is intuitive, but one must consider the marginal standard deviation when one integrates out τ:

$$p(x \mid \lambda) = \int p(x \mid \tau) \times p(\tau \mid \lambda)\, d\tau,$$

which is not tractable.

Choosing U and α

- For example, on p.9 of Simpson *et al.* (2017) $U = 0,968, \alpha = 0.01$ are specified which gives $\lambda = -\log(\alpha)/U = 4.8$, and leads to a marginal SD of 0.3:

```
rsd <- rgamma(10000,1,-log(.01)/.968);mean(rsd)
x <- rnorm(10000,0,rsd);sd(x)
[1] 0.21    # Mean of prior for sigma
[1] 0.3     # sd(x)
```

- For fixed α, trial and error would be one way of obtaining the desired value for sd(X).

- As an alternative Simpson *et al.* (2017), simulate from the prior for τ, for different values of U – call the empirical SD, Y.

- They assume a model, $\mathrm{E}[Y] = \beta \times U$, estimate the slope, which allows one to move from U to SD(X). For $U = 0.968, \alpha = 0.01$, $\widehat{\beta} = 0.31$ so that $\widehat{\beta} \times 0.968 = 0.3$, as expected.

Choosing U and α

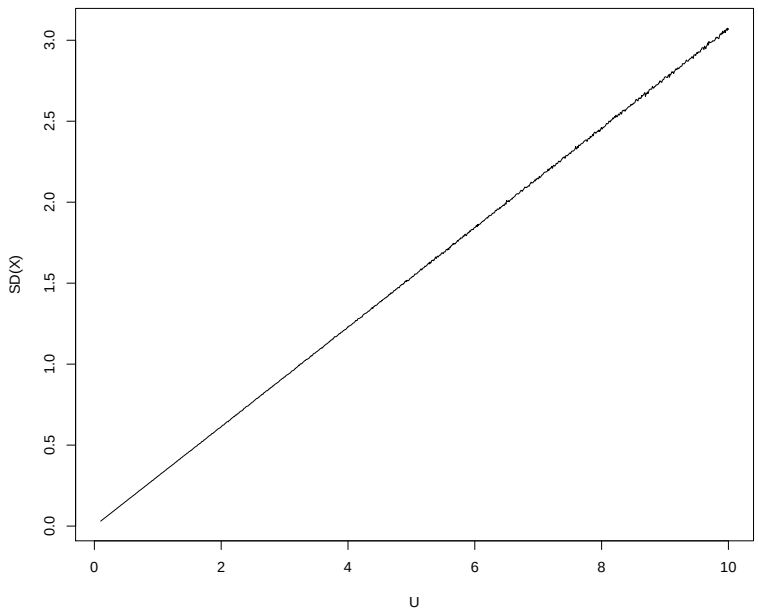


Figure 22: Relationship between SD(X) and U, for $\alpha = 0.01$. Slope is $\widehat{\beta} = 0.31$.

Example: RW1 models

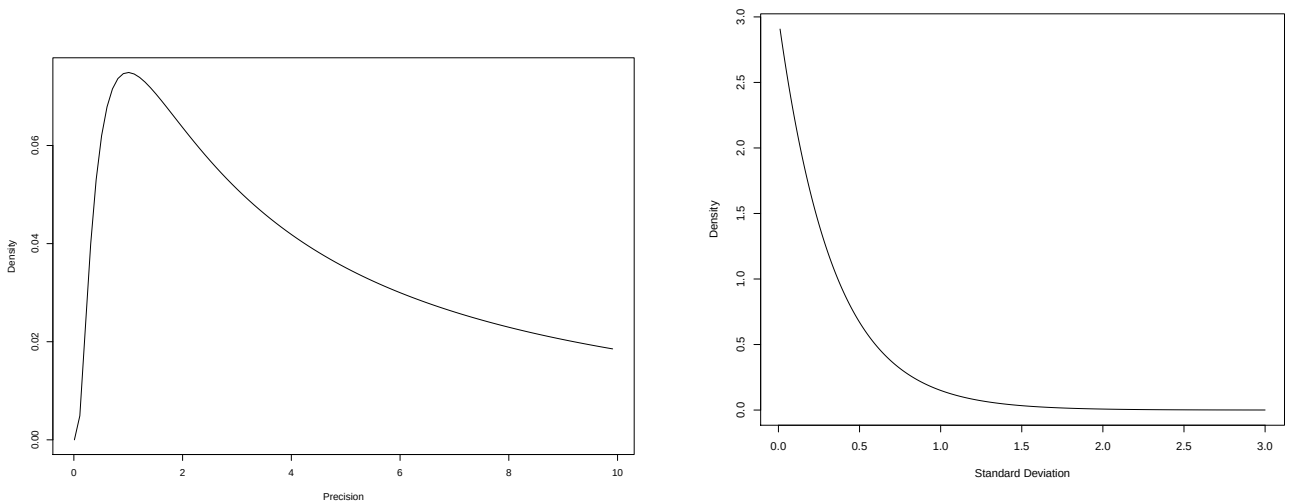


Figure 23: Prior specification on the standard deviation in a normal random effects model is $\Pr(1/\sqrt{\tau} > 1) = 0.05$. Left: Type-2 Gumbel prior for the precision. Right: Exponential prior $\text{Exp}(-\log(\alpha)/U)$ for σ.

Example: RW1 models

- The above method works for improper models, so long as scaling is used.

- Below we show the sensitivity of the RW1 model with prevision τ, on a particular dataset (strontium ratio versus age data).

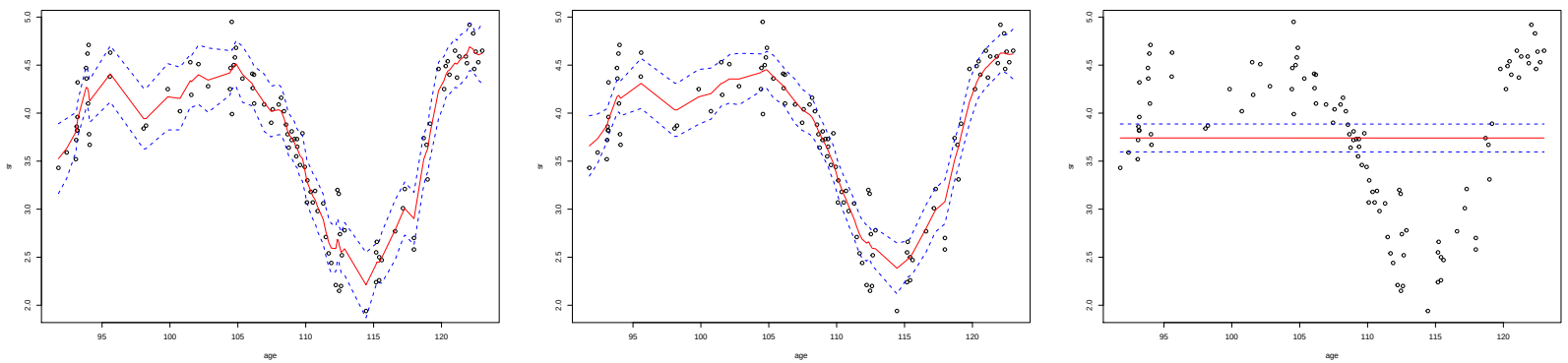


Figure 24: Sensitivity of RW1 smoothing to prior specification: $\alpha = 0.01$ in all cases, with (from left to right) $U = \overline{y} = 0.76$, $U = 0.03$ and $U = 0.01$ (these values are all on the standard deviation $\sqrt{1/\tau}$ scale. In the third case, we shrink to the base model, which is the null space, a horizontal line.

Example: BYM2 model

- Recall the BYM2 model with random effects,

$$\boldsymbol{u} = \sigma_u[\sqrt{1-\phi}\boldsymbol{e} + \sqrt{\phi}\boldsymbol{s}_\star]$$

  where $\boldsymbol{s}_\star$ have been rescaled to have a comparable (generalized) variance to the IID terms.

- Then,

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

- For a prior on σ_u^2, the previous derivation for normal random effects is relevant and we obtain a type-2 Gumbel distribution for the precision $1/\sigma_u^2$, with the λ parameter chosen through specifying a quantile on the standard deviation.

Example: BYM2 model

- In terms of the mixing parameter, $\phi = 0$ corresponds to the base model, i.e., no spatial variation.
- Let $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_1(\phi) = (1-\phi)\boldsymbol{I} + \phi\boldsymbol{Q}_\star^-$, then (Simpson *et al.*, 2017, Appendix A.2)

$$2\text{KLD}(\phi) = n\phi\left(\frac{1}{n}\text{tr}(\boldsymbol{Q}_\star^-) - 1\right) - \log|(1-\phi)\boldsymbol{I} + \phi\boldsymbol{Q}_\star^-|.$$

- Then turn the handle with (U, α) set through $\Pr(\phi < U) = \alpha$.

Example: AR1 model

- There are two possibilities for the base model for the AR1(ρ) temporal structure – either $\rho = 0$ (no temporal dependence) or $\rho = 1$ (complete temporal dependence).

- Franco-Villoria *et al.* (2019) give more details – both versions are available in `INLA`.

Example: Gaussian Process Models

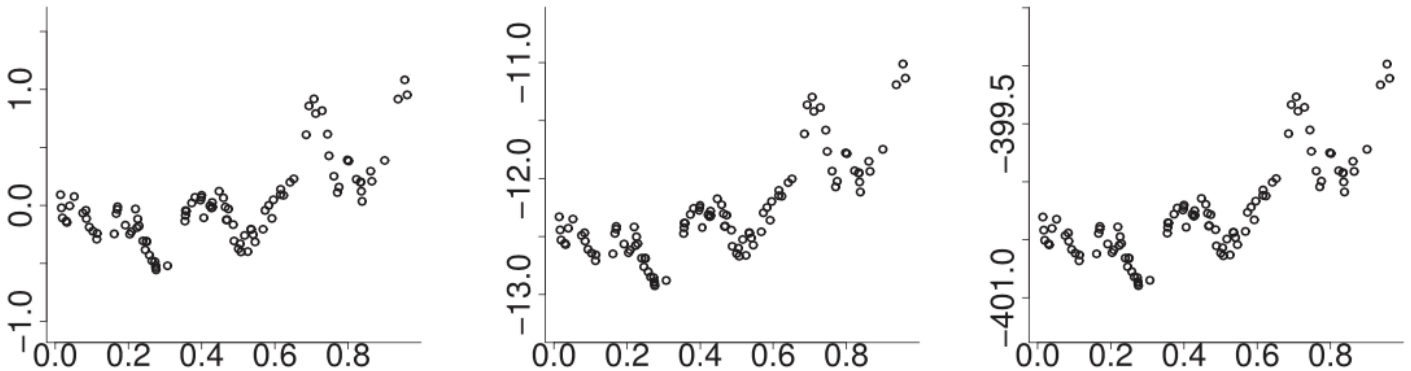
- Fuglstad *et al.* (2019) consider 1D, 2D and 3D GPs with a Matérn covariance model.
- Priors for this family require care, since there is little information in the data.
- The Matén covariance model produces a ridge in the likelihood for the range and spatial variance parameter – there is no consistent estimator under in-fill asymptotics for the parameters when the GP is of dimension 3 or less (Stein, 1999). Zhang and Zimmerman (2005) discuss in-fill and expanding domain asymptotics.

Example: Gaussian Process Models

- For example, for a 1D GP with an exponential covariance function observed on the interval $[0, 1]$, it is only the ratio of the range and the marginal variance that can be estimated consistently, and not the range or the marginal variance separately.

- This ratio also determines the asymptotic properties of predictions under in-fill asymptotics with the exponential covariance function (Stein, 1999).

Non-identifiability



(a) $\rho = \sigma^2 = 1$ (b) $\rho = \sigma^2 = 1000$ (c) $\rho = \sigma^2 = 1000000$

Figure 25: From Fuglstad *et al.* (2019). Simulations with the exponential covariance function $C(d) = \sigma_f^2 \exp(-d/\rho)$ for different values of $\rho = \sigma_f^2$ using the same underlying realization of independent standard Gaussian random variables. The patterns of the values are almost the same, but the levels differ.

PC prior for the Matérn covariance function

- Consider the Matérn covariance function:

$$\text{cov}(f(\boldsymbol{s}), f(\boldsymbol{s}')) = \frac{\sigma_f^2}{\Gamma(\upsilon)2^{\upsilon-1}}(\kappa||\boldsymbol{s}-\boldsymbol{s}'||)^{\upsilon}K_{\upsilon}(\kappa||\boldsymbol{s}-\boldsymbol{s}'||)$$

  where $K_{\upsilon}(\cdot)$ is a modified Bessel function of the second kind, $\sigma_f^2 > 0$ is a variance parameter, where $\upsilon > 0$ is a smoothness parameter, and κ is a scale parameter (with $\rho = \sqrt{8\upsilon}/\kappa$ being the distance at which the correlation falls to 0.1).

- Fuglstad *et al.* (2019) assume fixed υ and parameterize as κ and

$$\delta = \sigma_f \kappa^{\upsilon} \sqrt{\frac{\Gamma(\upsilon + d/2)(4\pi)^{d/2}}{\Gamma(\upsilon)}}$$

  This parameterization has the benefit that it describes what can, δ, and what cannot, κ, be consistently estimated under in fill asymptotics when the dimension of the base space $d \leq 3$.

PC prior for the Matén covariance function

- The joint prior is defined in two steps:
  - First: $\pi(\kappa \mid \delta)$.
  - Second: $\pi(\delta)$.
- We state the joint PC prior for the base model corresponding to $\kappa = 0$ (i.e., the range is infinite), and $\delta = 0$ (i.e., the marginal sd is zero), for details see Fuglstad *et al.* (2019).
- For $d \leq 3$, the joint PC prior is,

$$\pi(\kappa, \delta) = \frac{d}{2}\lambda_1\lambda_2(\kappa)\kappa^{d/2-1}\exp\left(-\lambda_1\kappa^{d/2} - \lambda_2(\kappa)\delta\right),$$

  where the user-specified values $\rho_0, \alpha_1, \sigma_{f0}, \alpha_2$ in $\Pr(\rho < \rho_0) = \alpha_1$ and $\Pr(\sigma_f > \sigma_{f0}) = \alpha_2$ allows identification of λ_1 and $\lambda_2(\kappa)$.
- This prior is easily transformed to a prior for σ_f, ρ.

PC prior for the Matén covariance function

Rationale for range=∞, via multiple emails with Geir-Arne Fuglstad (a member of STAB!):

1. The GP with range = 0 is not a meaningful model component. It is a stochastic process that does not make sense in continuous space.

2. The GP with range = ∞ is a meaningful model component. It is a stochastic process that does make sense in continuous space. It degenerates to an intercept.

3. Choose range = ∞ as a base model as this is the one that actually makes sense as a model component (and it is also the simplest "spatial" model that G-A can imagine).

Advantages of PC priors

A number of desirable features:

- PC priors are defined on individual components of the model, which is useful and means they can be defined in complex situations.
- Support of Occam's razor.
- Invariance to parameterization.
- Empirical robustness to the choice of the flexibility parameter.
- Avaialble in INLA for many models!