Bayesian SAE using Complex Survey Data: Methods and Applications
Lecture 2: Spatial Smoothing Models

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

CSSS Short Course: Thursday 23rd May, 2019

Outline

Motivation for Smoothing Models

Smoothing/Penalization

- When looking at estimates over space or time, we want to know if the differences we see are “real”, or simply reflecting sampling variability.
- In data sparse situations, when one expects similarity smoothing local patterns (in time, space, or both) can be highly beneficial.
- This can equivalently be thought of penalization, in which large deviations from “neighbors”, suitably defined, are discouraged.
- In this section we will generically think of modeling prevalence.
- We start with temporal modeling, since time is easier to think about! One dimensional and an obvious direction...

Motivation for Smoothing: Temporal Case

- Temporal setting: Even if the underlying prevalence is the same over time, we will see differences in the empirical estimates.
- Figure 1 demonstrates: We sampled binomial data with $n = 10, 20, 200$ and $p = 0.2$ (shown in blue) in all cases.
- In the top plot in particular, we might conclude large temporal variation, but all we are seeing is sampling variation.
- Figure 2 summarizes estimates from a second simulation in which there is a real temporal pattern – here we would not want to oversmooth and remove the trend.
- Later we will apply temporal smoothing models to these two sets of data.

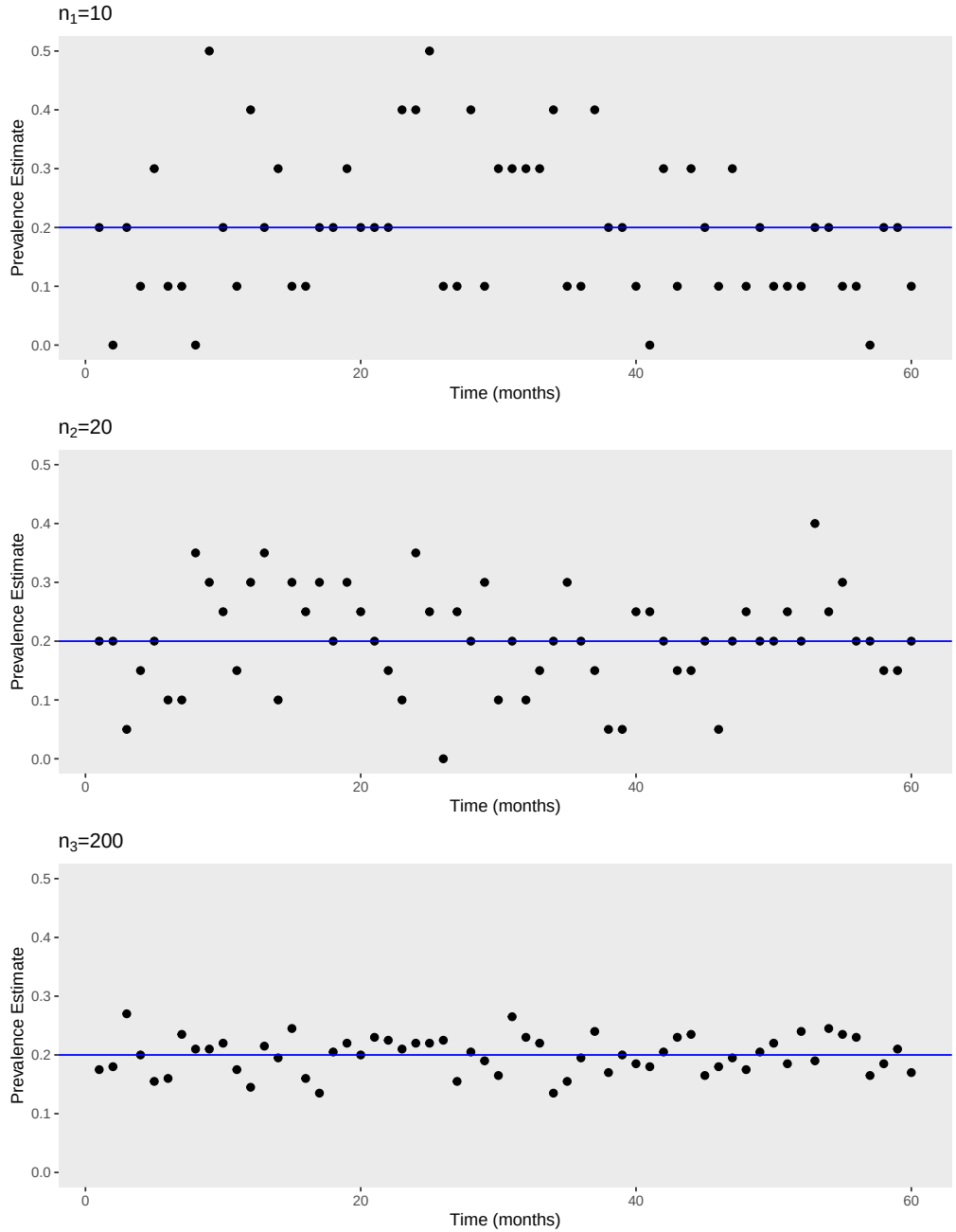


Figure 1: Prevalence estimates over time from simulated data with true prevalence of $p = 0.2$ (blue solid lines).

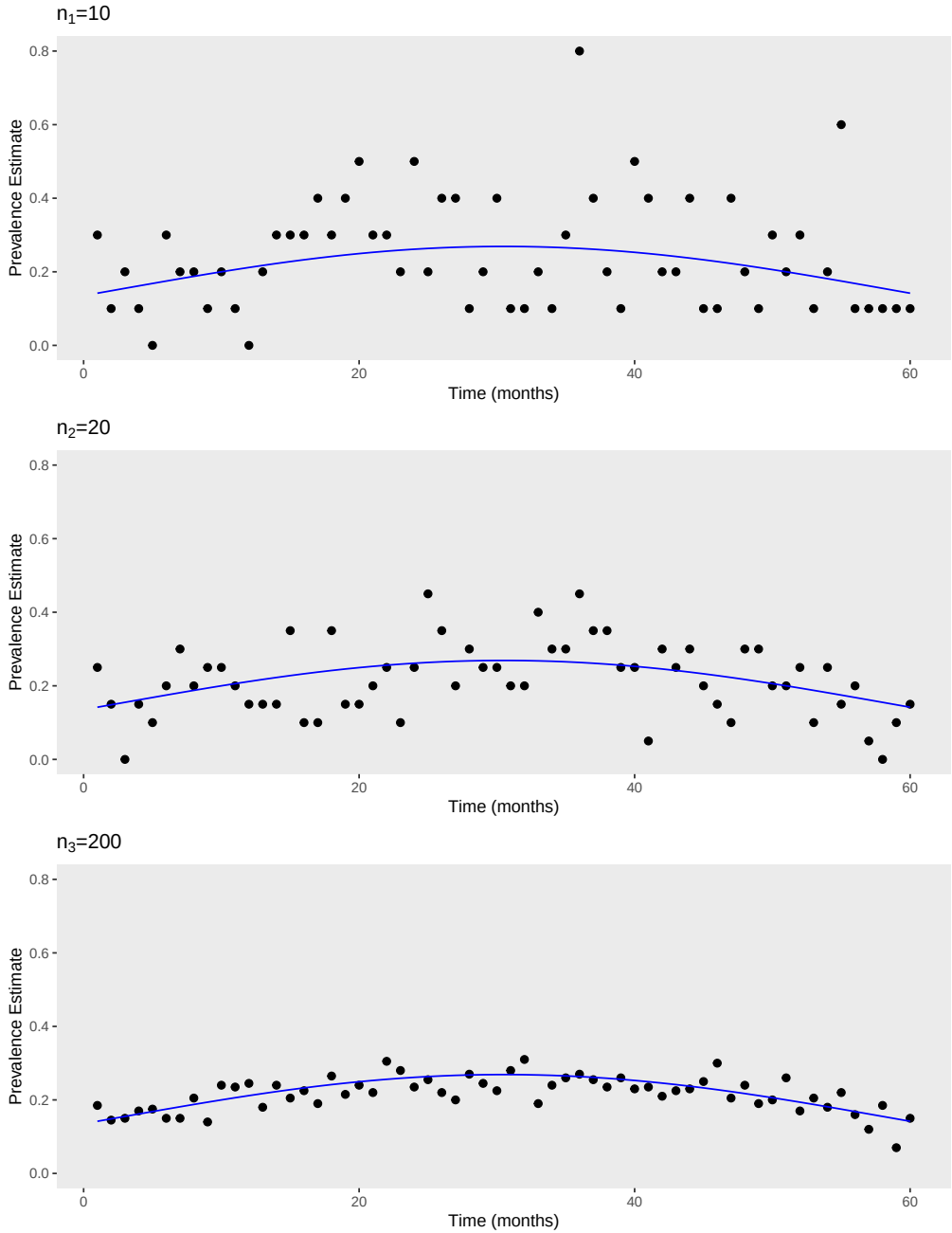


Figure 2: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line.

Motivation for Smoothing: Spatial Case

- We repeat the previous simulation example, but now for spatial data.
- Counts Y_i are simulated for each area i from a binomial distribution with prevalence p_i and sample size n_i:

$$Y_i \mid p_i \sim \text{Binomial}(n_i, p_i).$$

- We look at varying sample sizes $n_i = 50, 100, 500$, so that the influence of sampling variability can be examined.
- We examine two sets of simulated data:
  - Figure 3: Constant prevalence.
  - Figure 4: Spatially varying prevalence.

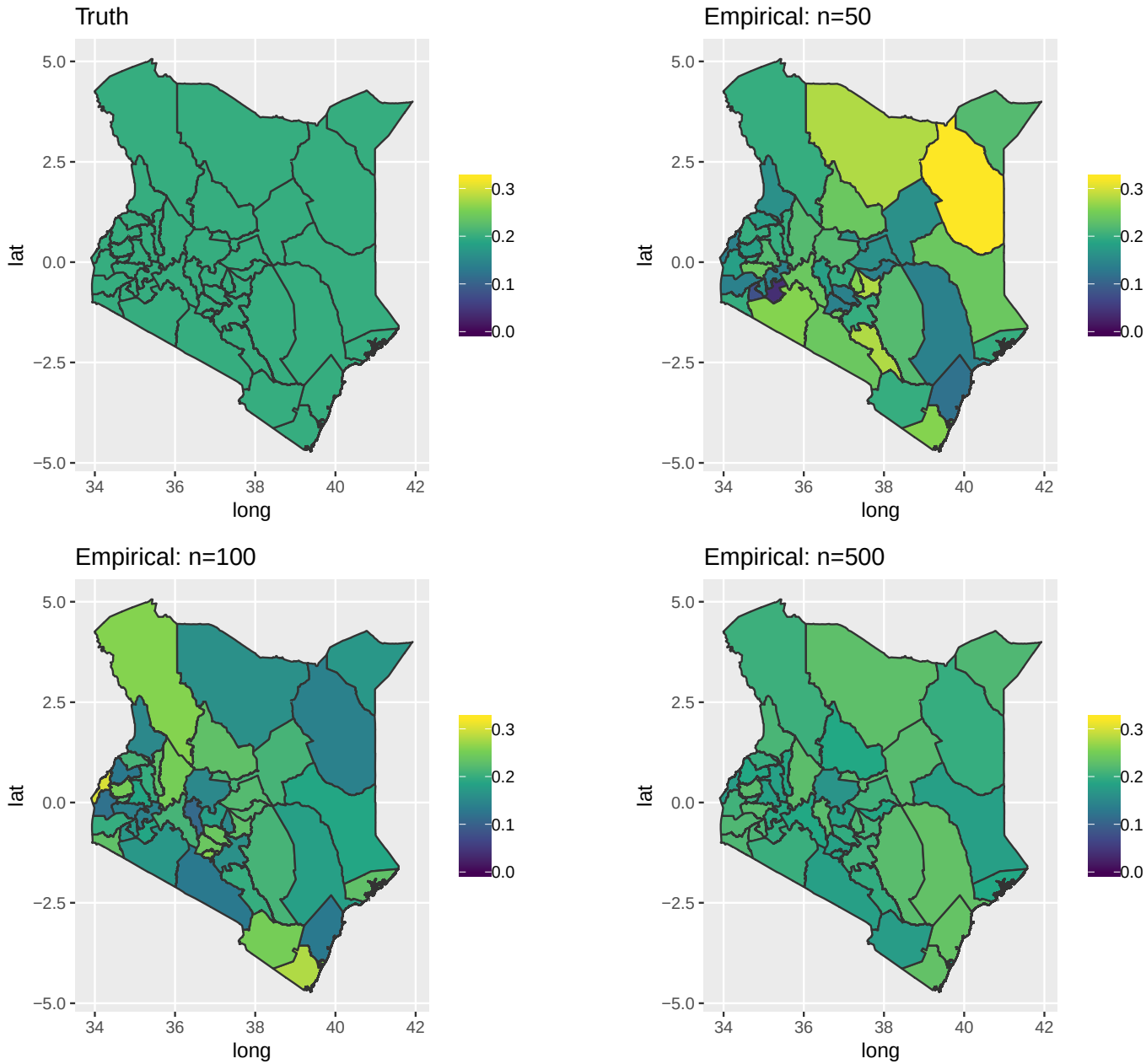


Figure 3: Prevalence estimates over space for simulated data with sample sizes of $n = 50, 100, 500$. True prevalence is 0.2 in all areas.

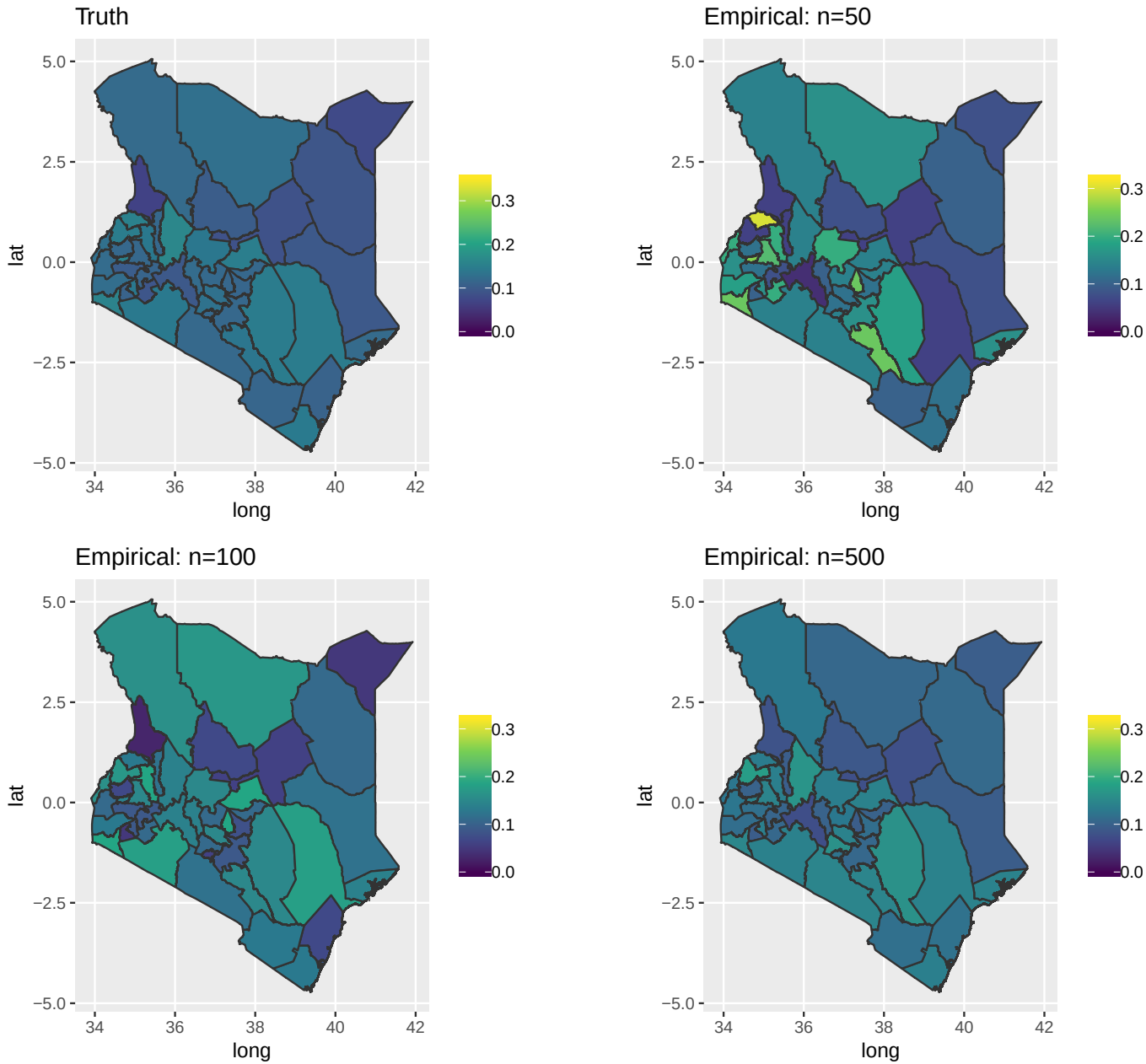


Figure 4: Prevalence estimates over space for simulated data with sample sizes of $n = 50, 100, 500$. True prevalence is spatially varying.

Smoothing

When faced with estimation n different quantities of the **prevalence** under different conditions, there are three model choices:

- The true underlying prevalence risks are **ALL THE SAME**.
- The true underlying prevalence risks are **DISTINCT** but not linked.
- The true underlying prevalence risks are **SIMILAR IN SOME SENSE**.

The third option seems plausible when the conditions are related, but how do we model “similarity”?

Smoothing

There are a number of possibilities for **SMOOTHING** models:

- The prevalences are drawn from some **COMMON** probability distribution, but are not ordered in any way. We refer this as the independent and identically distributed, or **IID** model. We could think of this as saying we think the prevalences are likely to be of the same order of magnitude.
- The prevalences are **CORRELATED** over time.

These are both examples of **HIERARCHICAL** or **RANDOM EFFECTS MODELS** — a key element is estimating the **SMOOTHING PARAMETER**.

Temporal Smoothing

Smoothing over Time

Rationale and overview of models for temporal smoothing:

- We often expect that the true underlying prevalence in an area will exhibit some degree of smoothness over time.
- A linear trend in time is unlikely to be suitable for more than a small number of years, and higher degree polynomials can produce erratic fits.
- Hence, local smoothing is preferred.
- Splines and random walk models have proved successful as local smoothers.
- And to emphasize again, in either approach, the choice of smoothing parameter is crucial.

Random Walk Models

We use random walk models which encourage the mean responses (e.g., prevalences) across time to not deviate too greatly from their neighbors.

The true underlying mean of the prevalence at time t is modeled as a function of its neighbors:

$$\mu_t \mid \mu_{\text{NE}(t)} \sim \text{N}(m_t, v_t),$$

where

- μ_t is the mean prevalence (or some function of it such as the logit) at time t.
- $\mu_{\text{NE}(t)}$ is the set of neighboring means – with the number of neighbors chosen depending on the model used – typically 2 or 4.
- m_t is the mean of some set of neighbors – for a **first order random walk** or **RW1** it is simply $\frac{1}{2}(\mu_{t-1} + \mu_{t+1})$.
- v_t is the variance, and depends on the number of neighbors – for the RW1 model it is $\sigma^2/2$, where σ^2 is a smoothing parameter – small values give large smoothing.

Random Walk Models

- The smoothing parameter σ^2 is estimated from the data, and determines the extent deviations from the mean are penalized.
- The penalty term for the RW1 model is:

$$p(\mu_t \mid \mu_{t-1}, \mu_{t+1}, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\mu_t - \tfrac{1}{2}\left(\mu_{t-1} + \mu_{t+1}\right)\right]^2\right\}.$$

- Hence:
  - Values of μ_t that are close to $\frac{1}{2}(\mu_{t-1} + \mu_{t+1})$ are favored (higher density).
  - The relative favorability is governed by σ^2 – if this variance is small, then μ_t can't stray too far from its neighbors.
- Predictions from the RW1 are

$$\mu_{T+s} | \mu_1, \ldots, \mu_T, \sigma^2 \sim \mathrm{N}(\mu_T, \sigma^2 \times s).$$

Figure 5: Illustration of the RW1 model for smoothing at time 3. The mean of the smoother is the average of the two adjacent points (and is highlighted as •), and deviations from this mean are penalized via the normal distribution shown in **red**.

RW1 Model

- Form of the prior density is:

$$\begin{aligned}\pi(\boldsymbol{\mu}|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(\mu_{t+1}-\mu_t)^2\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}\sum_{t\sim t'}(\mu_t-\mu_{t'})^2\right) = \exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\mu}\right)\end{aligned}$$

where $t \sim t'$ indicates t is a neighbor of t' and the precision is $\boldsymbol{Q} = \boldsymbol{R}/\sigma^2$ with

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}$$

and zeroes everywhere else.

- This sparsity leads to big gains in computational efficiency.

RW2 Model

- The second order RW (RW2) model produces smoother trajectories than the RW1, and has more reasonable short term predictions, which is desirable for modeling child prevalence.
- In terms of second differences:

$$(\mu_t - \mu_{t-1}) - (\mu_{t-1} - \mu_{t-2}) \sim \mathrm{N}(\,0, \sigma^2\,),$$

  showing that deviations from linearity are discouraged.
- Forecasts S steps ahead have a normal distribution with mean:

$$\mathrm{E}[\mu_{T+S} \mid \mu_1, \ldots, \mu_T] = \mu_T + S(\mu_T - \mu_{T-1})$$

  which is a linear function of the values at the last two time points.
- The variance is

$$\mathrm{var}(\mu_{T+S} \mid \mu_1, \ldots, \mu_T) = \frac{\sigma^2}{6} \times S(S+1)(2S+1)$$

  which is cubic in the number of periods S, so blows up very quickly.

RW2 Model

- Form of the prior density is:

$$\begin{aligned}\pi(\boldsymbol{\mu}|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T-2}(\mu_{t+2}-2\mu_{t+1}+\mu_t)^2\right)\\ &= \exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\mu}\right)\end{aligned}$$

where the precision is $\boldsymbol{Q} = \boldsymbol{R}/\sigma^2$ with

$$\boldsymbol{R} = \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & 1 & -4 & 6 & -4 & 1 & \\ & & . & . & . & . & . \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}$$

and zeroes everywhere else.

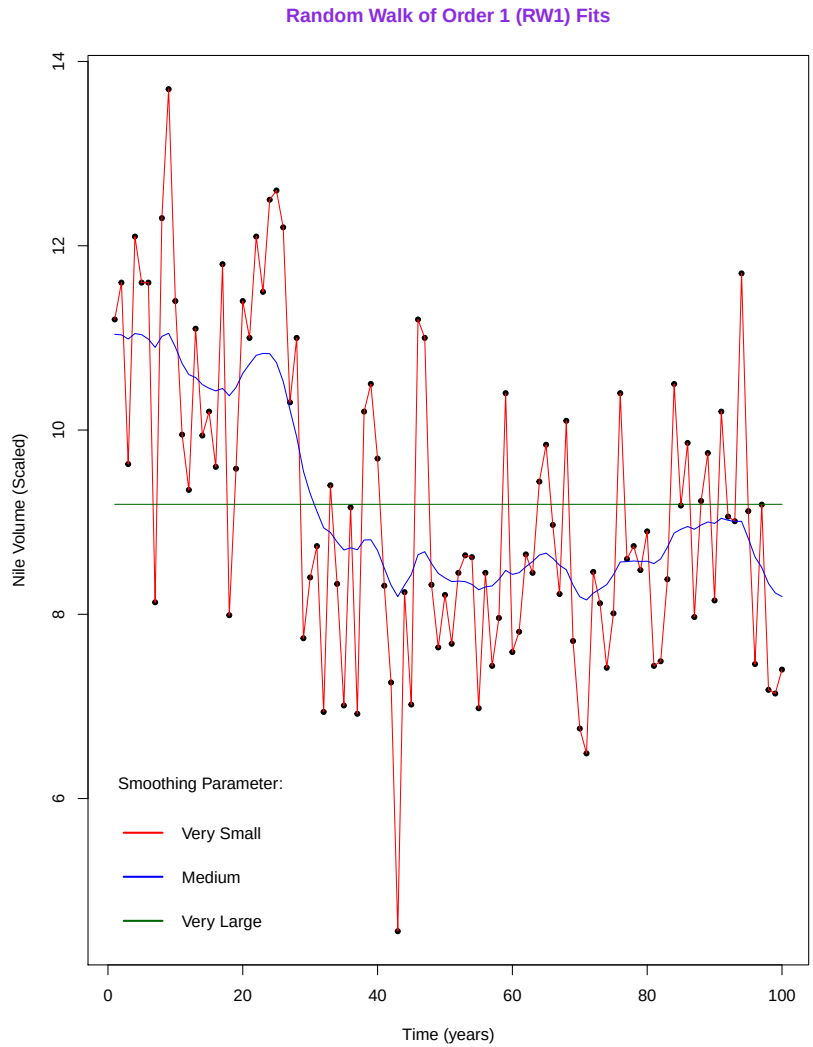


Figure 6: Nile data with RW1 fits under different priors for smoothing parameter σ^{-2}.

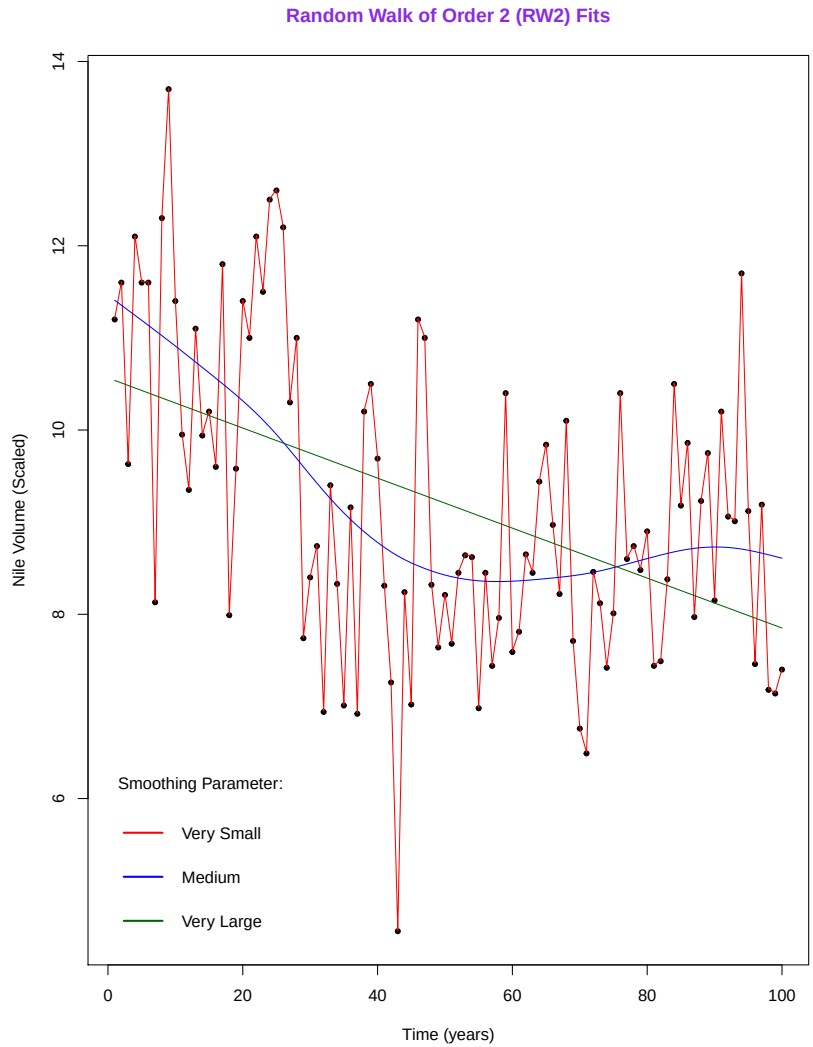


Figure 7: Nile data with RW2 fits under different priors for smoothing parameter σ^{-2}.

Temporal Smoothing Model Summary

We have three models:

IID MODEL:

$$\mu_t \sim \mathrm{N}(0, \sigma^2),$$

smooth towards zero.

RW1 MODEL:

$$\mu_t - \mu_{t-1} \sim \mathrm{N}(0, \sigma^2),$$

smooth towards the previous value.

RW2 MODEL:

$$(\mu_t - \mu_{t-1}) - (\mu_{t-1} - \mu_{t-2}) \sim \mathrm{N}(0, \sigma^2),$$

smooth towards the previous slope.

Bayesian Inference

Bayesian Inference

Bayesian inference is a convenient framework within which to implement smoothing models.

- A Data Model (Likelihood) is probabilistically combined with
- A Penalization (Prior) that expresses beliefs about the parameters θ encoding the model.
- Combination occurs via **Bayes Theorem:**

$$\underbrace{p(\theta|y)}_{\textbf{Posterior}} \propto \underbrace{L(\theta)}_{\textbf{Likelihood}} \times \underbrace{\pi(\theta)}_{\textbf{Prior}}.$$

- On the log scale:

$$\underbrace{\log p(\theta|y)}_{\textbf{Updated Beliefs}} = \underbrace{\log L(\theta)}_{\textbf{Data Model}} + \underbrace{\log \pi(\theta)}_{\textbf{Penalization}}.$$

Bayesian Inference

- In a Bayesian analysis the complete set of unknowns (parameters) is summarized via the **multivariate posterior distribution**.
- The marginal distribution for each parameter may be summarized via its mean, standard deviation, or quantiles.
- It is common to report the **posterior median** and a **90% or 95% posterior range** for parameters of interest.
- The range that is reported is known as a credible interval.
- The computations required for Bayesian inference (integrals) is often not trivial and many be carried out using a variety of analytic, numeric and simulation based techniques.
- We use the integrated nested Laplace approximation (INLA), introduced by Rue *et al.* (2009).
- Book-length treatments:
  - Blangiardo and Cameletti (2015) – space-time models.
  - Wang *et al.* (2018) – general models.
  - Krainski *et al.* (2018) – advanced space-time models.

Bayes Example

- Imagine the data model is normal with an unknown mean μ:

$$\overline{y} \mid \mu \sim \mathsf{N}(\mu, \sigma^2/n),$$

  where σ^2/n is assumed known ($\sigma/\sqrt{n}$ is the standard error).
- We also imagine the prior is normal:

$$\mu \sim \mathsf{N}(m, v),$$

  so that values of the mean μ that are (relatively) far from m are penalized.
- The log posterior is:

$$\underbrace{\log p(\mu \mid y\ )}_{\text{Updated Beliefs}} = \underbrace{-\frac{n}{2\sigma^2}(\overline{y} - \mu)^2}_{\text{Data Model}} \underbrace{-\frac{1}{2v}(\mu - m)^2}_{\text{Penalization}}.$$

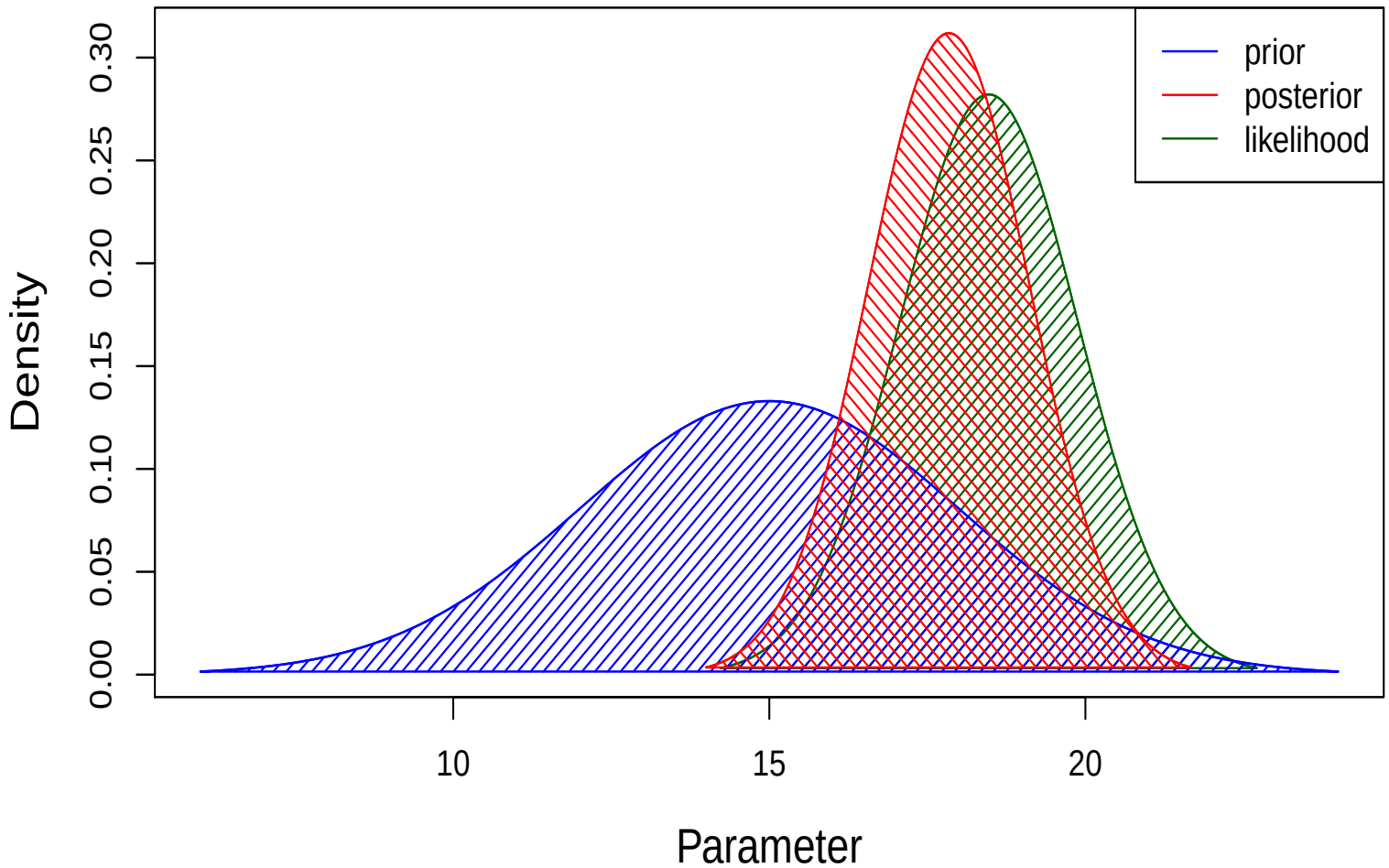


Figure 8: Normal data model with $n = 10$, $\overline{y} = 19.3$ and standard error 1.41. The prior for μ has mean m =15 and $v = 3^2$. The posterior for the parameter μ is a compromise between the two sources of information: the posterior mean is 18.5 and the posterior standard deviation is 1.28.

RW Fitting to Simulated Data

- We illustrate fitting with the RW2 model, using the simulated data seen earlier.
- The model is:

$$
\begin{aligned}
Y_t | p_t &\sim \text{Binomial}(n_t, p_t) \\
\frac{p_t}{1 - p_t} &= \exp(\alpha + \phi_t) \\
(\phi_1, \ldots, \phi_T) &\sim \text{RW2}(\sigma^2) \\
\sigma^2 &\sim \text{Prior on Smoothing Parameter} \\
\alpha &\sim \text{Prior on Intercept}
\end{aligned}
$$

- Fit using `R-INLA`.
- On Figures 9 and 10 the fitted values are shown in red – in both the constant prevalence and curved prevalence cases, the reconstruction is reasonable.

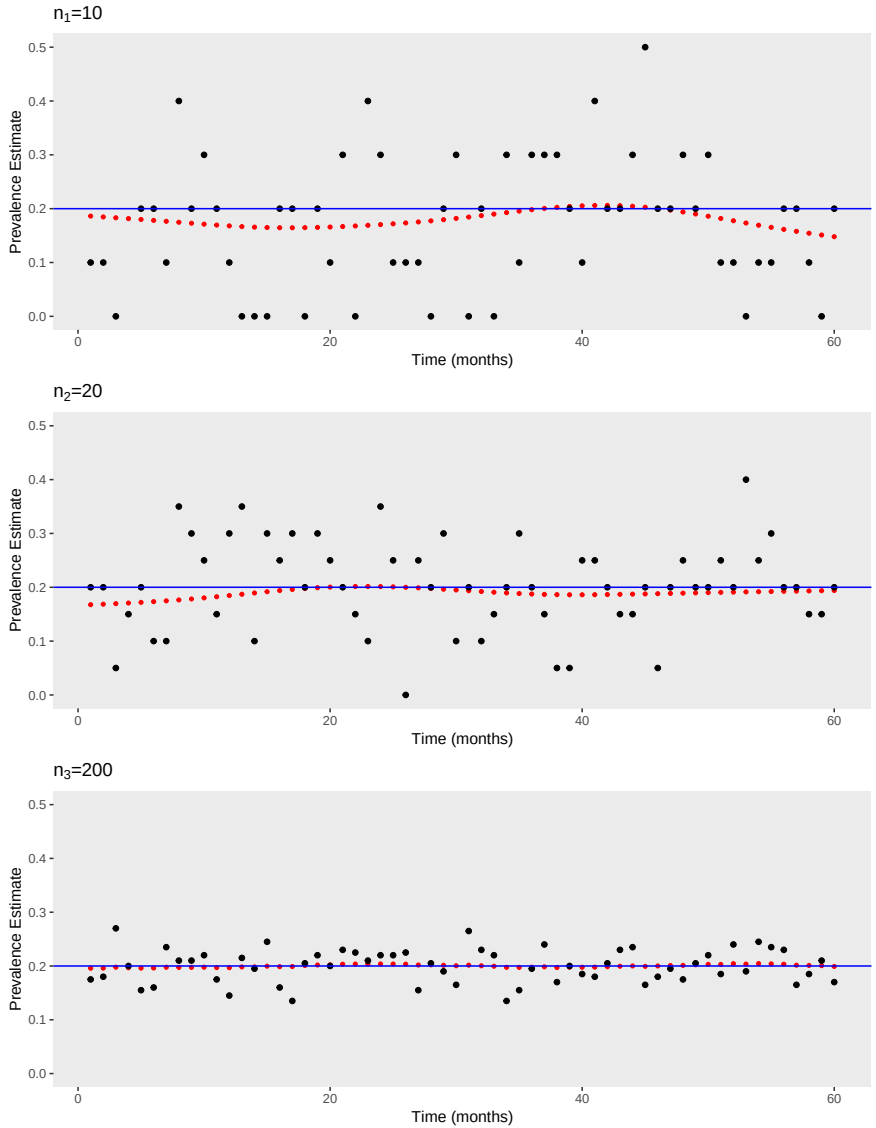


Figure 9: Prevalence estimates over time from simulated data, true prevalence $p = 0.2$ (blue solid lines). Smoothed random walk estimates in **red**.

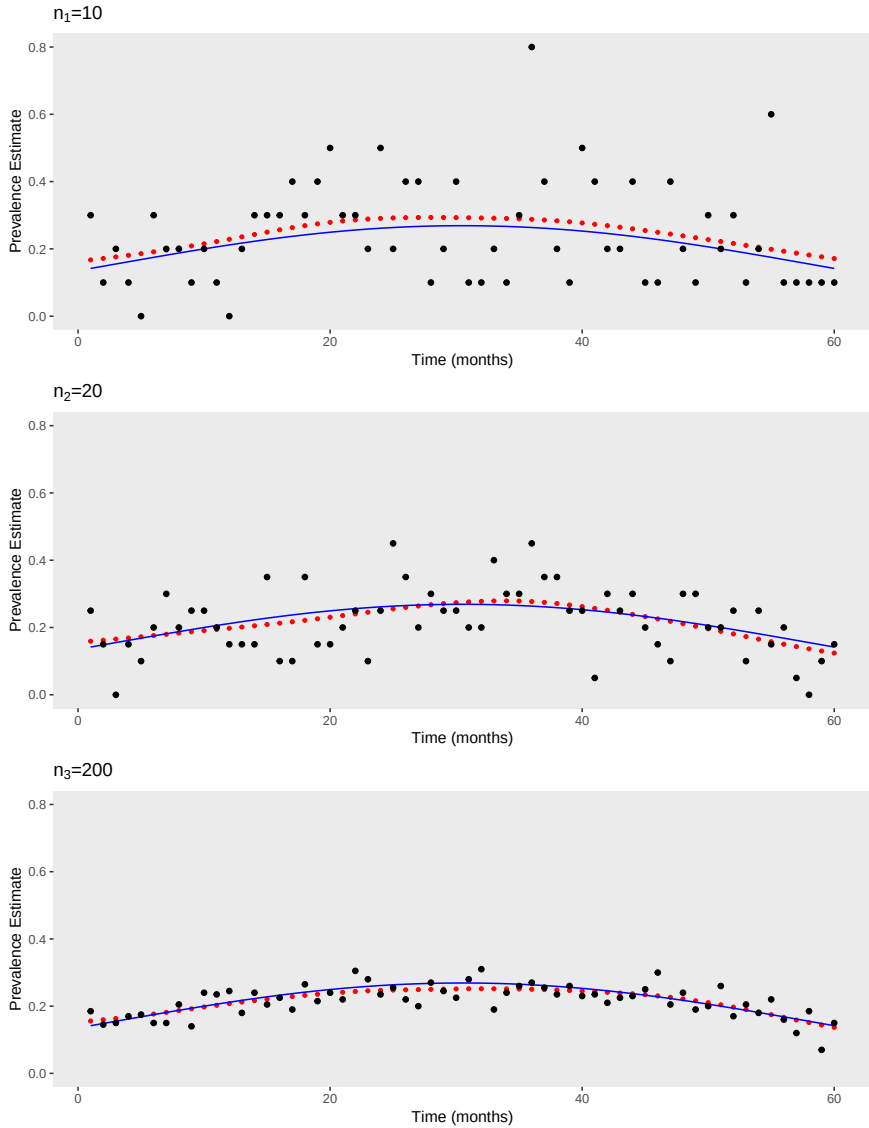


Figure 10: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line. Smoothed random walk estimates in **red**.

Spatial Smoothing

Spatial Modeling

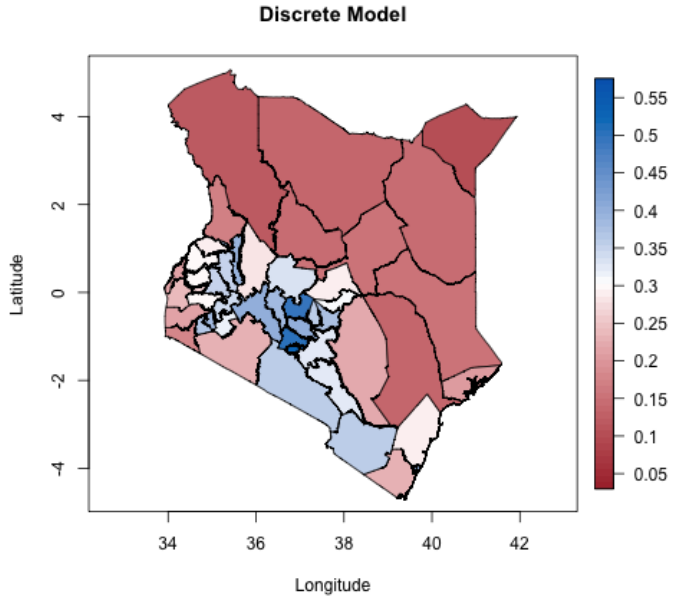
Two approaches to modeling spatial dependence:

1. **Covariance matrix approach (Stein, 1999):**
   - Kriging is the label attached to prediction using this approach,
   - Intuitive isotropic correlation models based on distance, leads to dense matrices.
   - Unfortunately, if n is large, computation is a nightmare, because we need to manipulate $n \times n$ matrices, which involves $O(n^3)$ operation (Rue and Held, 2005).
   - A large number of approximations have been proposed: fixed rank Kriging, lattice Kriging, predictive processes, SPDE,...
   - Known as Gaussian Process (GP), or geostatistical, models.
2. **Precision matrix approach (Besag, 1974):**
   - Model local structure, leads to sparse matrices but less intuition on the implied covariances.
   - Computation is very efficient with either MCMC, INLA (Rue *et al.*, 2009) or TMB.

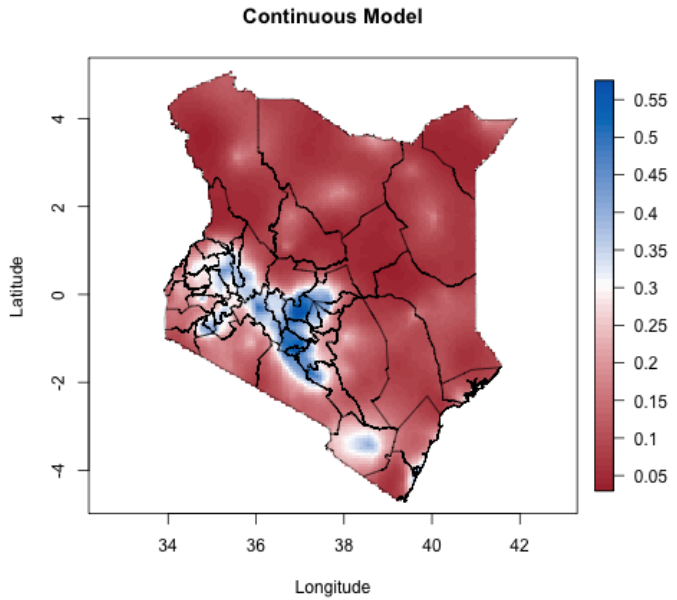
Each of these approaches can be used for point or area data.

The Two Approaches to Spatial Smoothing

- Model at the area level using a discrete spatial model.
  These are the models that are implemented in the SUMMER package.



- Model at the point level using a continuous spatial model.
  Gaussian Process (GP) models abound and have many different implementations.



Spatial Models for Binomial Data

Point Data: For a cluster at $\boldsymbol{s}_c$:

$$Y(\boldsymbol{s}_c)|p(\boldsymbol{s}_c) \sim \text{Binomial}(N(\boldsymbol{s}_c), p(\boldsymbol{s}_c)).$$

- Continuous spatial random effects:

$$p(\boldsymbol{s}_c) = \text{expit}(\beta_0 + \boldsymbol{x}(\boldsymbol{s}_c)^\top \beta_1 + \underbrace{S(\boldsymbol{s}_c)}_{\substack{\text{Continuous} \\ \text{Spatial}}} + \underbrace{\epsilon_c}_{\substack{\text{“Nugget”} \\ \text{Random} \\ \text{Noise}}}).$$

- Discrete spatial random effects:

$$p(\boldsymbol{s}_c) = \text{expit}(\beta_0 + \boldsymbol{x}(\boldsymbol{s}_c)^\top \beta_1 + \underbrace{S(\boldsymbol{s}_{i[c]})}_{\substack{\text{Discrete} \\ \text{Spatial}}} + \underbrace{\epsilon_{i[c]}}_{\substack{\text{Discrete} \\ \text{Random} \\ \text{Noise}}}),$$

where $i[c]$ is the spatial area within which cluster c lies.

Spatial Models for Binomial Data

Aggregate Data: For area i,

$$Y_i | p_i \sim \text{Binomial}(N_i, p_i).$$

- Discrete spatial random effects:

$$p_i = \text{expit}(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \underbrace{S_i}_{\substack{\text{Discrete} \\ \text{Spatial}}} + \underbrace{\epsilon_i}_{\substack{\text{Discrete} \\ \text{Random} \\ \text{Noise}}}).$$

- Continuous spatial random effects:

$$\begin{aligned} p_i &= \int_{s \in A_i} p(\boldsymbol{s}) d(\boldsymbol{s}) \, d\boldsymbol{s} \\ &= \int_{s \in A_i} \text{expit}(\quad \beta_0 + \boldsymbol{x}(\boldsymbol{s})^\top \boldsymbol{\beta}_1 + \underbrace{S(\boldsymbol{s})}_{\substack{\text{Continuous} \\ \text{Spatial}}} \quad) \, d(\boldsymbol{s}) \, d\boldsymbol{s}, \end{aligned}$$

where $d(\boldsymbol{s})$ is population density at location $\boldsymbol{s}$.

Spatial Models for Binomial Data

- Key Point: The area-level prevalence is not,

$$p_i = \text{expit}(\beta_0 + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_1 + S_i),$$

with

$$\begin{aligned} S_i &= \int_{s \in A_i} \underbrace{S(\boldsymbol{s})}_{\substack{\text{Continuous} \\ \text{Spatial}}} d(\boldsymbol{s})\, d\boldsymbol{s} \\ \boldsymbol{x}_i &= \int_{s \in A_i} \boldsymbol{x}(\boldsymbol{s})\, d(\boldsymbol{s})\, d\boldsymbol{s} \end{aligned}$$

which is what some have done.

- Classic ecological fallacy (Wakefield, 2008).

Discrete Spatial Models

In spatial epidemiology in regions with disease registries, the data are available as aggregates and so area-based models are popular:

- Besag, York and Mollié (1991), the famous BYM model – see also reparameterized version in Riebler *et al.* (2016), known as BYM2.
- The Leroux model (Leroux *et al.*, 1999) model is increasing in popularity – another reparameterized version in Riebler *et al.* (2016).
- Proper conditional autoregression (CAR) (Cressie and Wikle, 2011).
- Multivariate normal model with points taken as centroids of each area: computationally expensive and not appealing for area-level data. Little used.
- Simultaneous Autoregression (SAR) models (LeSage and Pace, 2009). Popular in econometrics, less so in health and demographic modeling.

Difference between BYM2, Leroux and CAR is often small, but expertise required for priors specification – penalized complexity priors have good theoretical basis and work well in practice (Simpson *et al.*, 2017).

The BYM Model

In the BYM model:

$$p_i = \text{expit}(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + S_i + \epsilon_i),$$

where

- $\epsilon_i \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$.
- The spatial effects S_i are modeled conditional on the neighbors. Specifically,

$$S_i | \{S_j = s_j, j \sim i\}, \sigma_s^2 \sim \text{N}\left(\bar{s}_i, \frac{\sigma_s^2}{m_i}\right),$$

  where $\bar{s}_i = \frac{1}{m_i} \sum_{j \sim i} s_j$ is the mean of the neighbors (defined in some way) of area i and m_i is the number of such neighbors.
- σ_s^2 is a smoothing parameter: large values indicate large spatial variability .
- The distribution of the complete set $\boldsymbol{s} = [s_1, \ldots, s_n]^\top$:

$$p(\boldsymbol{s} | \sigma_s^2) \propto \exp\left(-\frac{1}{2\sigma_s^2} \boldsymbol{s}^\top \boldsymbol{Q} \boldsymbol{s}\right)$$

  is improper, since $\boldsymbol{Q}$ is singular.

Example of a Neighborhood Scheme

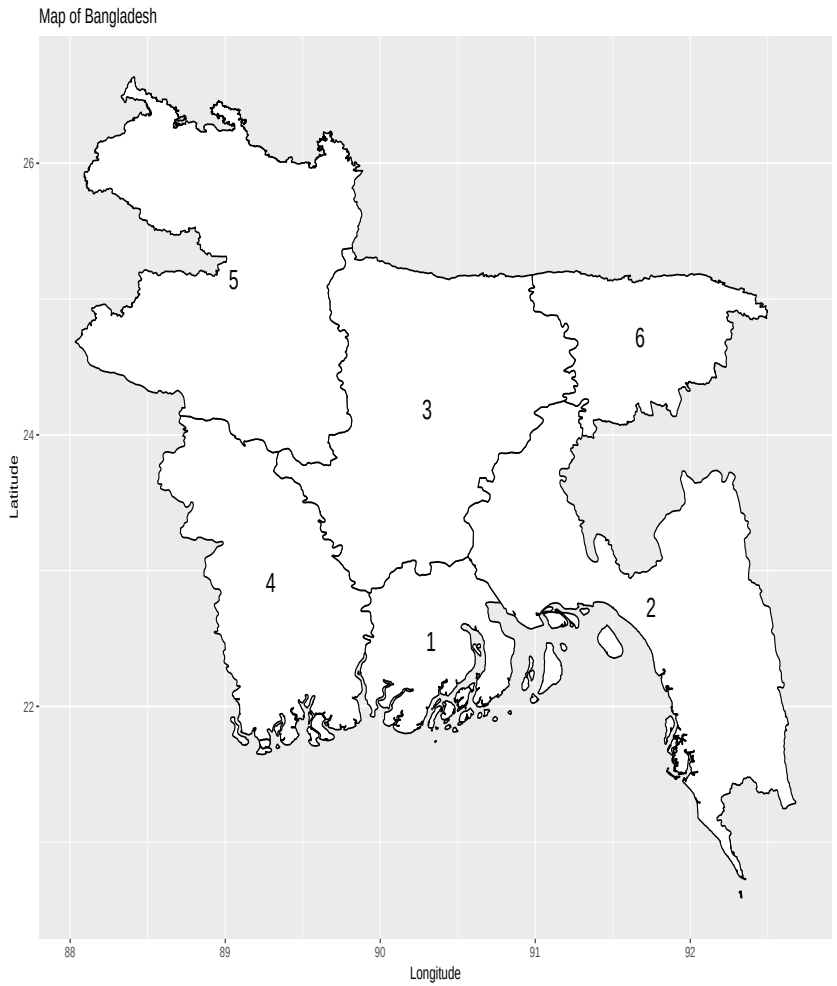


Figure 11: Common boundary neighbor scheme for Bangladesh divisions.

Prior for spatial effects $\boldsymbol{s}$ is,

$$p(\boldsymbol{s}|\sigma_s^2) \propto \exp\left(-\frac{1}{2}\boldsymbol{s}^\top \boldsymbol{Q}\boldsymbol{s}\right).$$

Precision matrix, $\boldsymbol{Q} = \boldsymbol{R}/\sigma_s^2$, $\boldsymbol{R}$:

$$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 & -1 \\ -1 & -1 & 5 & -1 & -1 & -1 \\ -1 & 0 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & -1 & 2 & 0 \\ 0 & -1 & -1 & 0 & 0 & 2 \end{bmatrix}$$

Each row and each column sum to 0, illustrating non-singularity.

Spatial Smoothing of Simulated Data

Data Model: For area i:

$$\underbrace{Y_i}_{\text{Count}} \mid \underbrace{p_i}_{\text{Prevalance}} \sim \underbrace{\text{Binomial}(n_i, p_i)}_{\text{Data Model}}.$$

Smoothing Model: For the odds in area i:

$$\frac{p_i}{1 - p_i} = \exp(\alpha + \phi_i).$$

We consider two choices for the smoothing model:

- IID model: Smooth to the overall mean with no spatial structure $\phi_i \sim \text{N}(0, \sigma^2)$ where σ^2 controls the amount of smoothing — small/large corresponds to strong/weak smoothing.
- BYM[1] model: Add a spatial component that encourages local similarity analogously to the random walk model with a suitable choice of neighbors, sharing a common boundary being the commonest choice.

[1] named after the paper that introdced the model, Besag, York and Mollié (1991)

Spatial Modeling of Simulated Data for $n = 50$ Constant Risk Case

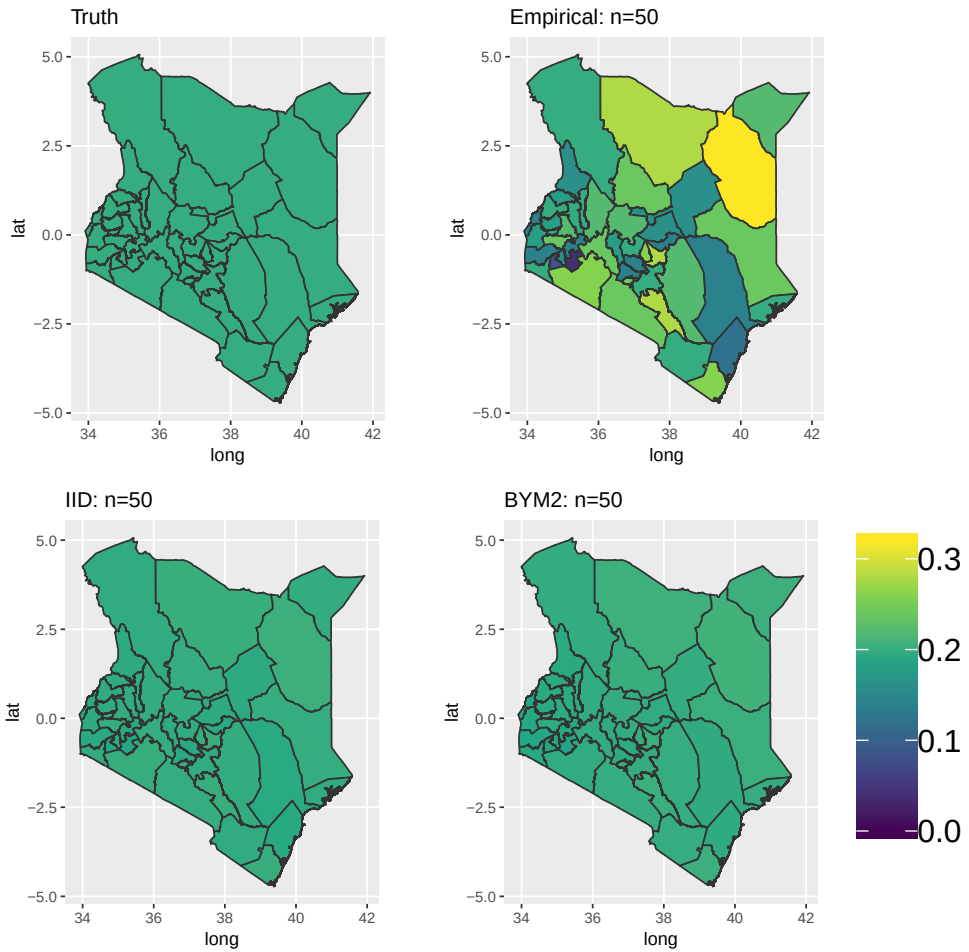


Figure 12: Results with $n = 50$ when true prevalence is 0.2. Top Left: Truth. Top Right: raw proportions. Bottom Left: Estimates with IID model. Bottom Right: smoothing with BYM2.

Spatial Modeling of Simulated Data for $n = 100$ Constant Risk Case

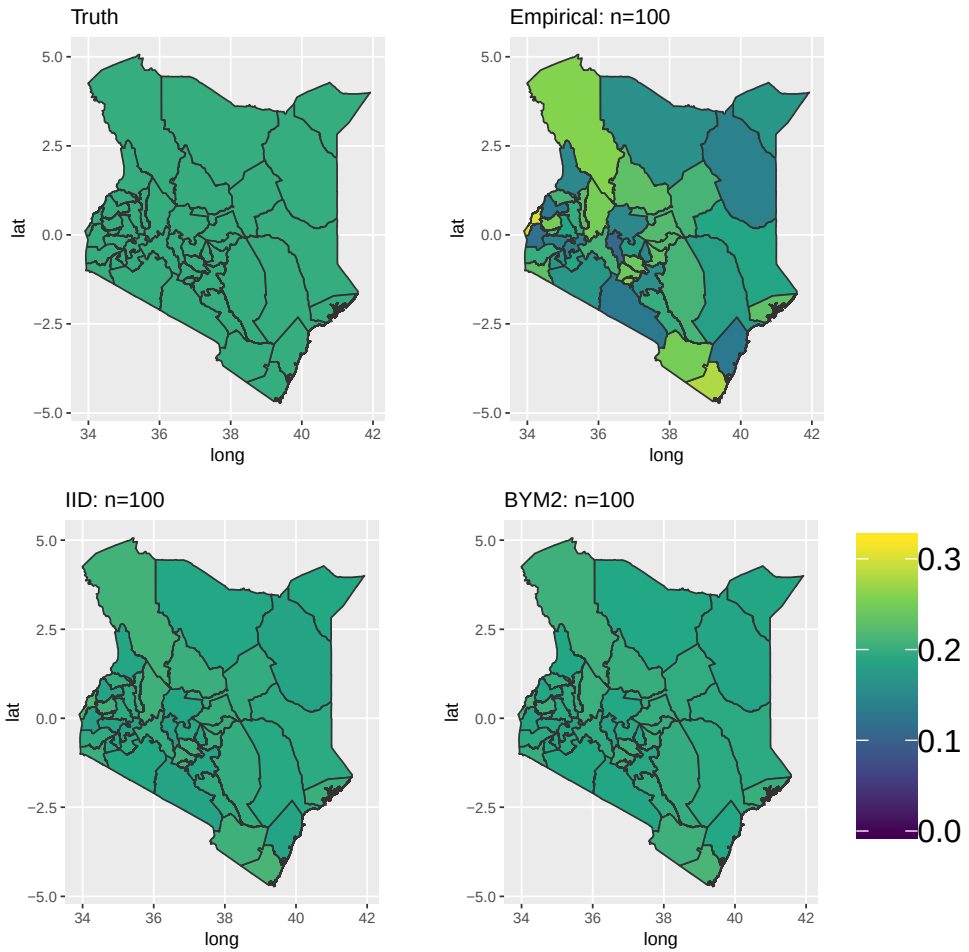


Figure 13: Results with $n = 100$ when true prevalence is 0.2. Top Left: Truth. Top Right: raw proportions. Bottom Left: Estimates with IID model. Bottom Right: smoothing with BYM2.

Spatial Modeling of Simulated Data for $n = 500$ Constant Risk Case

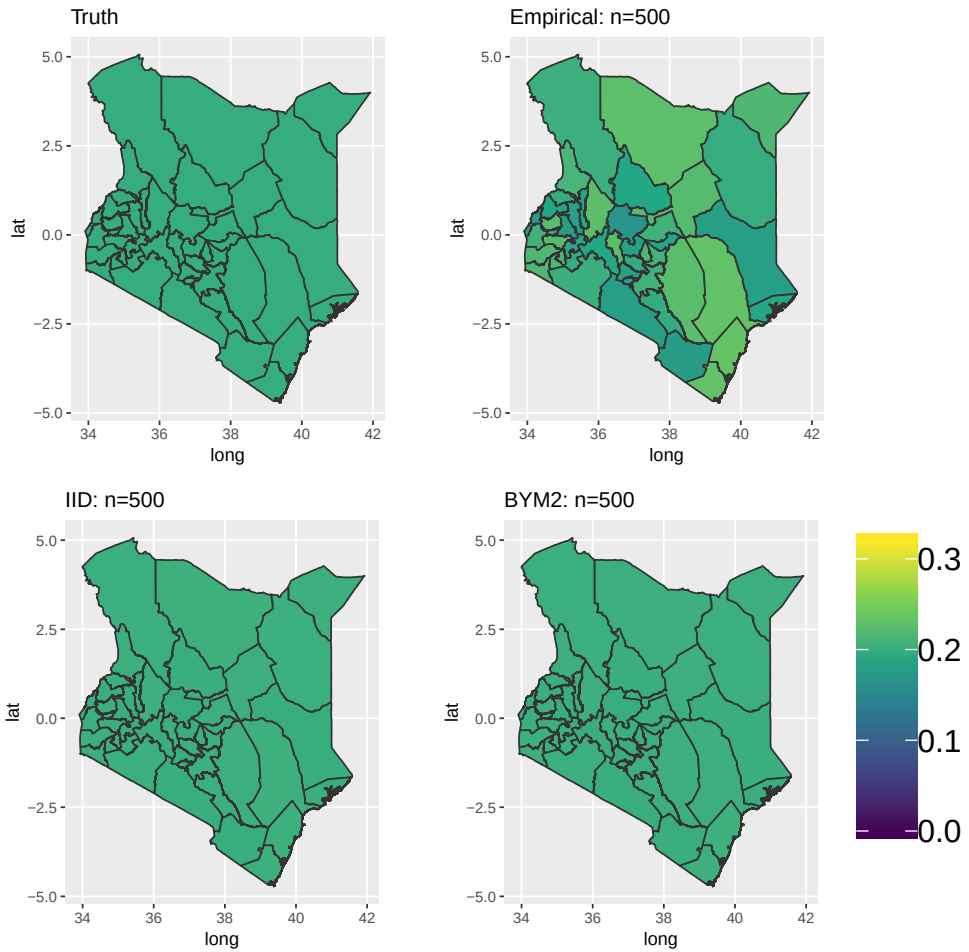


Figure 14: Results with $n = 500$ when true prevalence is 0.2. Top Left: Truth. Top Right: raw proportions. Bottom Left: Estimates with IID model. Bottom Right: smoothing with BYM2.

Spatial Modeling of Simulated Data for $n = 50$ Varying Risk Case

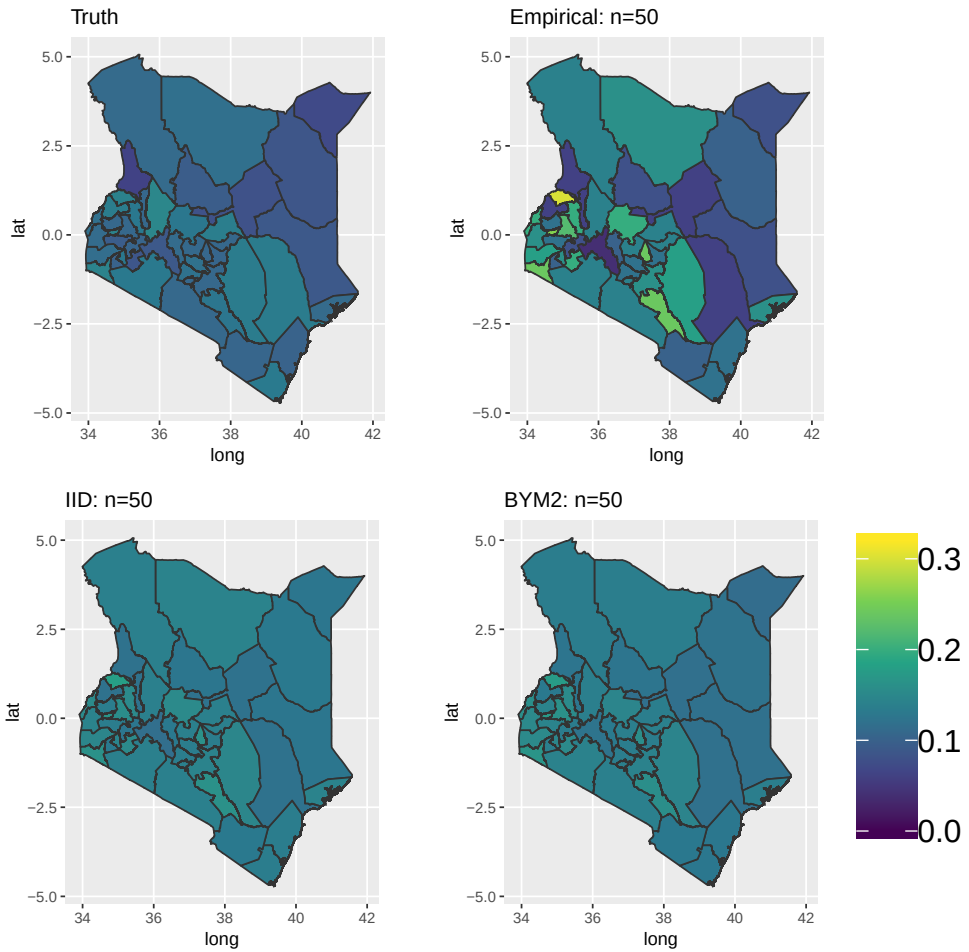


Figure 15: Results with $n = 100$ when true prevalence is varying. Top left: Truth. Top right: Raw proportions, Bottom left: smoothing with IID model. Bottom right: smoothing with BYM2.

Spatial Modeling of Simulated Data for $n = 100$ Varying Risk Case

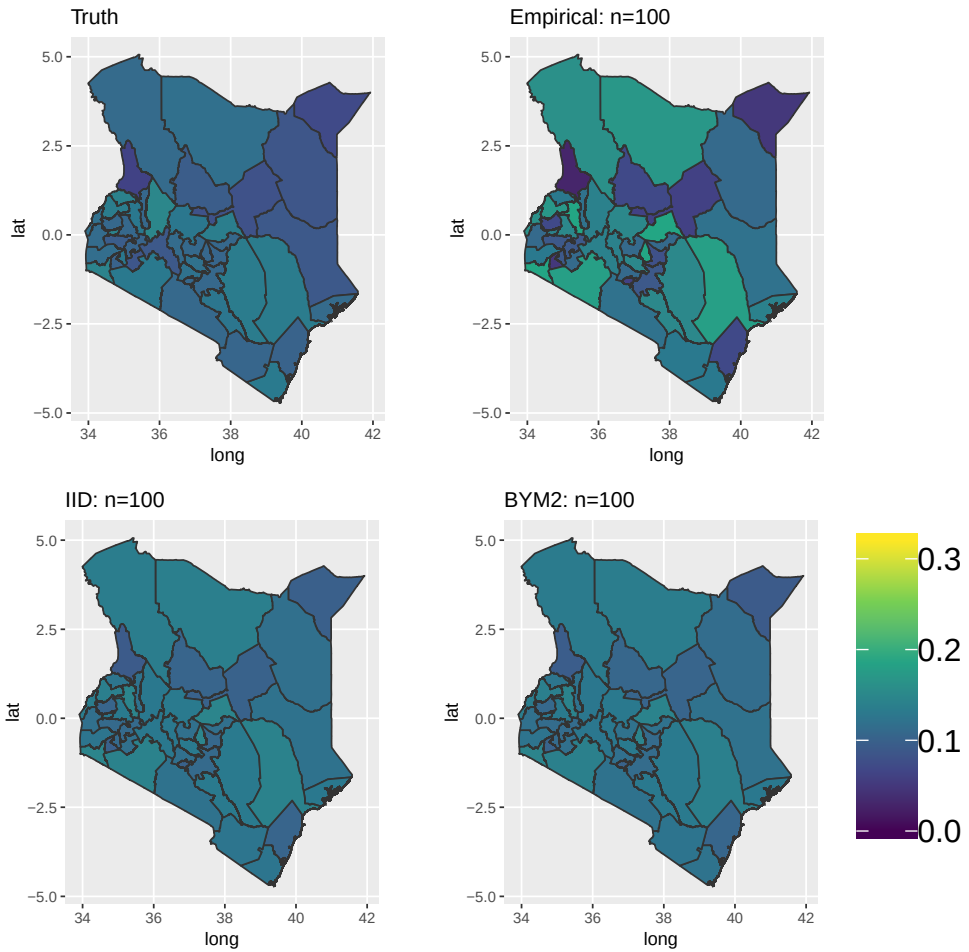


Figure 16: Results with $n = 100$ when true prevalence is varying. Top left: Truth. Top right: Raw proportions, Bottom left: smoothing with IID model. Bottom right: smoothing with BYM2.

Spatial Smoothing of Simulated Data for $n = 500$ Case

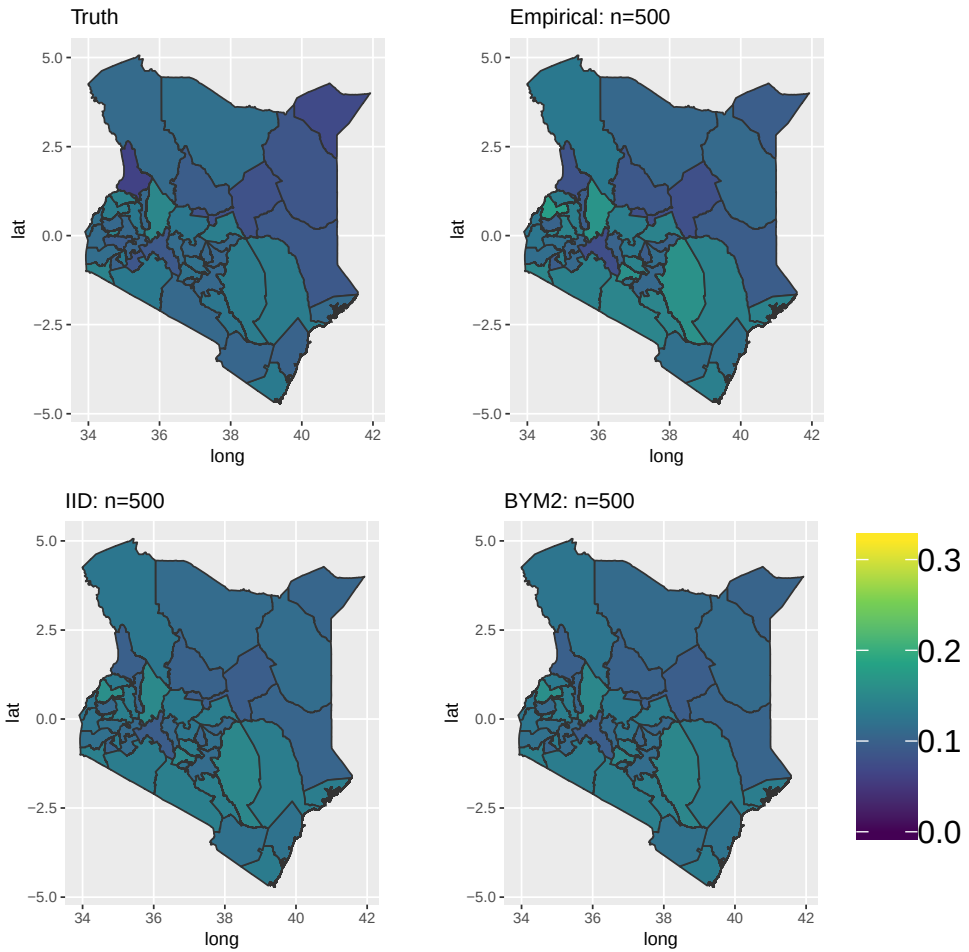


Figure 17: Results with $n = 500$ when true prevalence is varying. Top left: Truth. Top right: Raw proportions. Bottom left: smoothing with IID model. Bottom right: smoothing with BYM2.

Continuous Modeling

Continuous spatial models are popular in health and demography:

- Routinely used by both WorldPop (Wardrop *et al.*, 2018) and IHME (Golding *et al.*, 2017), but continuous modeling is a more hazardous approach to estimation.
- However, it is the way forward to allow **multiple data sources** at different spatial resolutions to be combined.
- And reporting can be on a relevant discrete scale.

For a comparison of discrete and spatial models, see Wakefield *et al.* (2018).

Continuous Spatial Models

For point data $\boldsymbol{s} = [s_1, \ldots, s_n]^\top$, computation and inference is theoretically straightforward, just need to calculate

$$p(\boldsymbol{s}|\theta) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{s}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{s}\right)$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta)$, but determinant and inverse are computationally prohibitive.

Alternatives:

- Lattice Kriging (Nychka *et al.*, 2015).
- Fixed rank Kriging (Cressie and Johannesson, 2008).
- Predictive processes (Banerjee *et al.*, 2008).
- Stochastic Partial Differential Equations (SPDE, Lindgren *et al.* (2011)).

Heaton *et al.* (2017) is a good review paper.

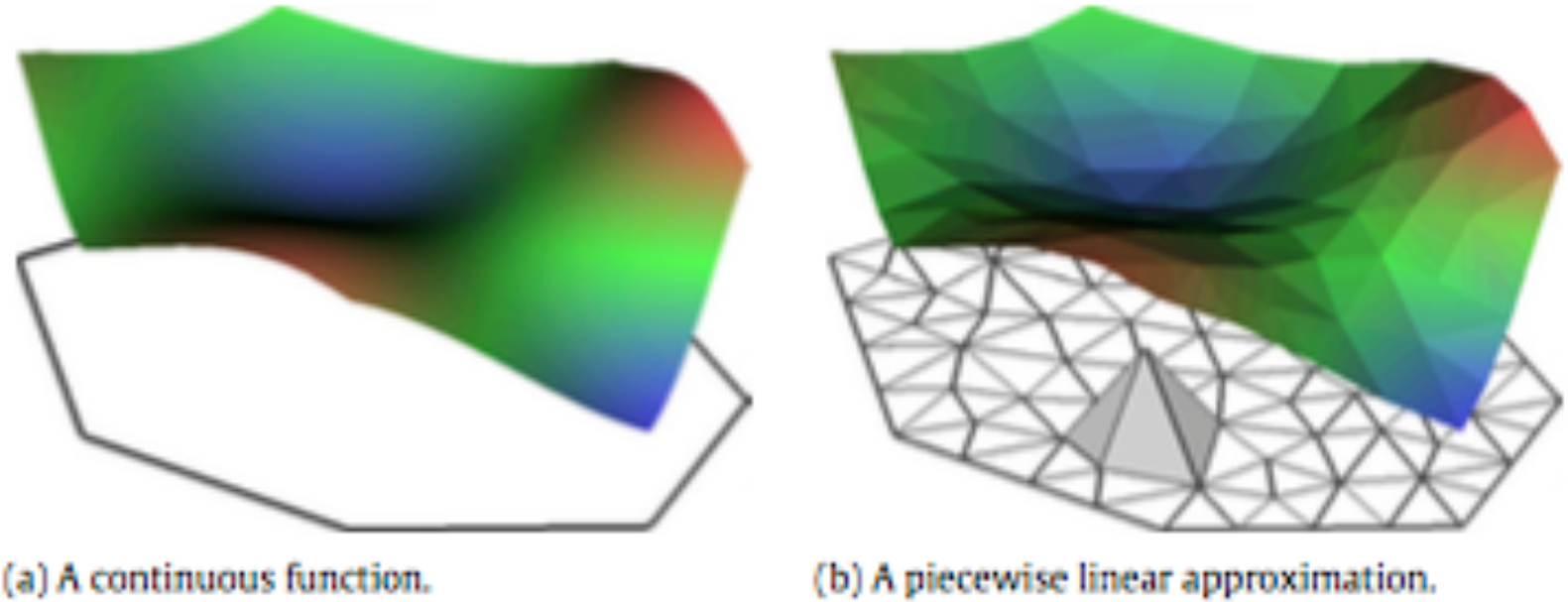
Stochastic Partial Differential Equations (SPDEs)

- The GRF is approximated by

$$S(\boldsymbol{s}) \approx \sum_{i=1}^{m} w_i \phi_i(\boldsymbol{s})$$

  where the w_i are random and $\phi_i(s)$ are a set of basis functions, $i = 1, \ldots, m$.
- We need to identify the distribution of w_i and this can be done using a particular stochastic partial differential equation (SPDE).
- This results in the distribution of $w_1, \ldots, w_m$ being a Gaussian MRF, so that the computation is efficient via INLA.

SPDE Approximation



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

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

Figure 18: GMRF representation of a Markovian GRF, via triangulation, from Simpson *et al.* (2012)

Computation

Continuous spatial models are computationally difficult.

Possibilities:

- MCMC: stochastic simulation – can converge very slowly (Filippone *et al.*, 2013).
- INLA: combination of analytic and numerical approximations (Rue *et al.*, 2009).
- TMB: `R` package built around automatic differentiation routines, with capabilities for Laplace transforms and MCMC (Kristensen *et al.*, 2016; Osgood-Zimmerman and Wakefield, 2019). Less user-friendly but more flexible than `R-INLA`.

Discrete spatial models have multiple efficient implementations (MCMC, INLA, TMB).

Spatio-Temporal Smoothing

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.
- Suppose we have two factors with levels, A and B, with $i = 1, \ldots, I$ and $j = 1, \ldots, J$ indexing the levels.
- A main effects only model takes the form

$$\mathrm{E}[Y_{ij}|\beta_0, \eta_i, \phi_j] = \beta_0 + \eta_i + \phi_j.$$

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

Main Effects and Interactions

- An interaction model adds a set of interaction parameters

$$\mathrm{E}[Y|\beta_0, \eta_i, \phi_j, \delta_{ij}] = \beta_0 + \eta_i + \phi_j + \delta_{ij}.$$

- **Interpretation:** δ_{ij} is the additional effect, beyond $\eta_i + \phi_j$ 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 structure and we would expect similarity.

Separable Main Effects Model

- First, consider a separable space-time model

$$\begin{aligned} Y_{it}|P_{it} &\sim \text{Binomial}(n_{it}, P_{it}) \\ \theta_{it} = \text{logit}(P_{it}) &= \beta_0 + \epsilon_i + S_i + \omega_t + \tau_t \end{aligned}$$

- Components:
  - Unstructured spatial term $\epsilon_i \sim_{iid} \text{N}(0, \sigma_v^2), i = 1, \ldots, n$.
  - Smooth spatial term $[S_1, \ldots, S_n]$ smooth in space (e.g. ICAR model).
  - Smooth temporal term $[\tau_1, \ldots, \tau_T]$ smooth in time (e.g. follows a RW1 or RW2 model).
  - Unstructured temporal term $\omega_t \sim_{iid} \text{N}(0, \sigma_\omega^2), t = 1, \ldots, T$.
- Notice there is no interaction between space and time.
- The spatial effects are constant across time and temporal trends are constant across space.

Inseparable Space-Time Interaction Models

- Knorr-Held (2000) considered the model:

$$\theta_{it} = \beta_0 + \epsilon_i + S_i + \omega_t + \tau_t + \delta_{it}$$

  with ϵ_i, S_i, ω_t, η_t are as in the separable 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.

Inseparable Space-Time Interaction Models

- Type I: $\delta_{it} \sim_{iid} \mathrm{N}(0, \sigma_\delta^2)$.
- Type II: Temporal trends differ between areas but don't have spatial structure.
- For example, an RW(2) model in each area gives the joint distribution:

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

- Realistic to assume that time trends have no spatial structure?

Inseparable Space-Time Interaction Models

- Type III: Spatial patterns differ between time points but without temporal structure:

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

- So this model says we have independent ICAR models at each time point (though with the same variance, σ_δ^2).
- Realistic to assume that spatial structure changes at every time point without smooth patterns in space?

Inseparable Space-Time Interaction Models

- Type IV: Temporal trends differ between areas but more likely to be similar for adjacent 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:

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

- `R-INLA` implements each of these models.

Space-Time Interactions for Continuous Space

- The most common space-time interaction models are separable, which means the variance-covariance matrices have the Kronecker form:

$$\mathbf{\Sigma}_{ST} = \mathbf{\Sigma}_S \otimes \mathbf{\Sigma}_T$$

- So if $\mathbf{\Sigma}_S$ is $n \times n$ and $\mathbf{\Sigma}_T$ is $T \times T$, $\mathbf{\Sigma}_{ST}$ will be $nT \times nT$.
- Suppose we have covariance models in time and space:

$$\begin{aligned} \text{cov}(S(\boldsymbol{s}), S(\boldsymbol{s}')) &= \sigma_S^2 \rho_S^{|\boldsymbol{s}-\boldsymbol{s}'|} \\ \text{cov}(\tau(t), \tau(t')) &= \sigma_T^2 \rho_T^{|t-t'|} \end{aligned}$$

- We can combine to give the separable interaction process:

$$\text{cov}(U(\boldsymbol{s}, t), U(\boldsymbol{s}', t')) = \sigma_S^2 \rho_S^{|\boldsymbol{s}-\boldsymbol{s}'|} \times \sigma_T^2 \rho_T^{|t-t'|}$$

Surface Reconstructions for U5MR in Kenya

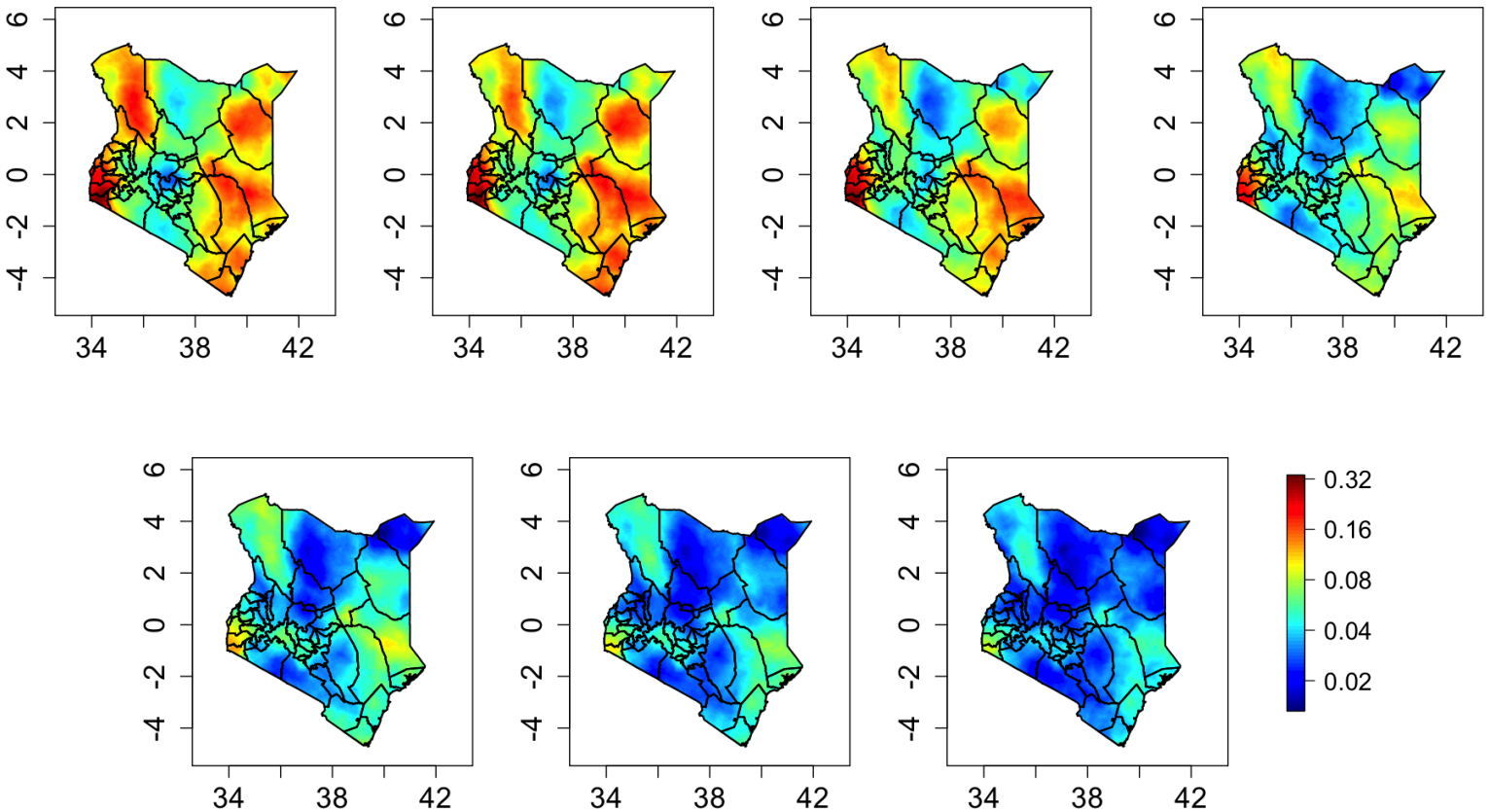


Figure 19: Posterior medians of U5MR for 1990, 1995, 2000, 2005, 2010, 2015, 2020. **Important Point:** These are point estimates and the uncertainty at each pixel is in general very large.

Estimates for U5MR in Malawi

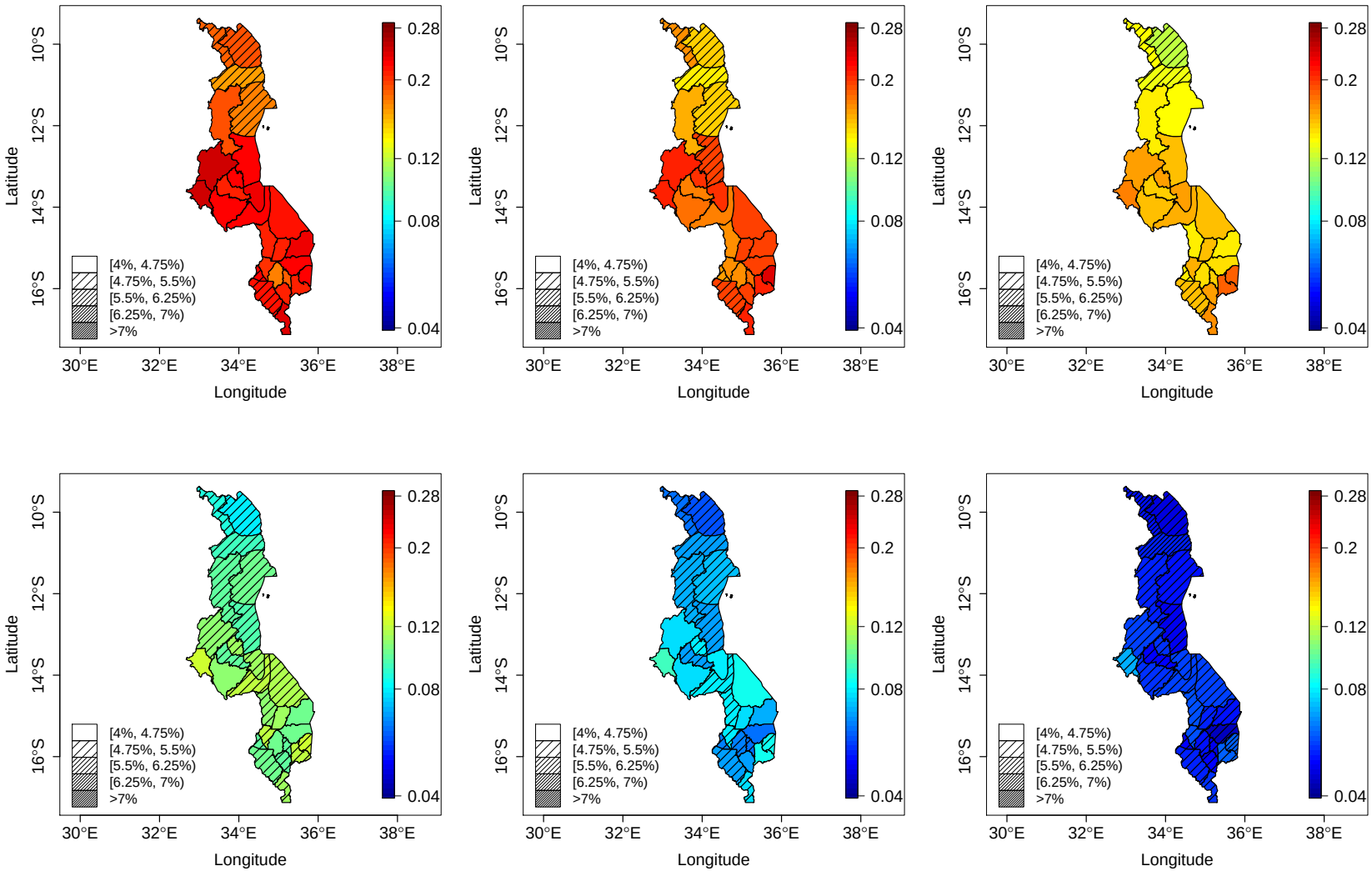


Figure 20: Estimates of U5MR for Malawi for 1990, 1995, 2000, 2005, 2010, 2015.

Discussion

Discussion

- Discrete spatial models are well understood and relatively easy to use.
- Computation for discrete spatial models is fast.
- Continuous spatial models pose a greater computational challenge.
- Computation for continuous spatial models may be challenging, depending on the model – harder to “package up”.

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **70**, 825–848.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

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

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley and Sons.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209–226.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. John Wiley and Sons.

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, **93**, 93–114.

Golding, N., Burstein, R., Longbottom, J., Browne, A., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D., Dwyer-Lindgren, L., Farag, T., Flaxman, A., Fraser, M., Gething, P., Gibson, H., Graetz, N., Krause, L., Kulikoff, X., Lim, S., Mappin, B., Morozoff, C., Reiner, R., Sligar, A., Smith, D., Wang, H., Weiss, D., Murray, C., Moyes, C., and Hay, S. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet*, **390**, 2171–2182.

Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., *et al.* (2017). Methods for analyzing large spatial data: A review and comparison. *arXiv preprint arXiv:1710.05013*.

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.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21.

Leroux, B., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, the Environment and Clinical Trials*, pages 179–192. Springer, New York.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics (Statistics, textbooks and monographs)*. CRC Press.

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.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, **24**, 579–599.

Osgood-Zimmerman, A. and Wakefield, J. (2019). Template Model Builder (TMB), a flexible alternative to the Integrated Nested Laplace Approximation. *In Preparation*.

Riebler, A., Sørbye, S., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, **25**(4), 1145–1165.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (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., Lindgren, F., and Rue, H. (2012). 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.

Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health*, **29**, 75–90.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. (2018). Estimating under five mortality in space and time

in a developing world context. *Statistical Methods in Medical Research*. Published online April 19.

Wang, X., Yue, Y., and Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

Wardrop, N., Jochem, W., Bird, T., Chamberlain, H., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., and Tatem, A. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, **115**, 3529–3537.