Bayesian Subnational Estimation using Complex Survey Data: Bayesian Inference and Smoothing Models

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation for Smoothing Models

Bayesian Inference

Temporal Smoothing

Spatial Smoothing

Spatio-Temporal Smoothing

Discussion

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.
- To this end, we formulate statistical approaches to model the totality of data, which allows us to disentangle signal from sampling variability.
- In data sparse situations, when one expects similarity, smoothing data locally (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.
- We start with temporal modeling, since time is easier to think about! One dimensional and an obvious direction.
- In this lecture, we will often take modeling a prevalence as our generic objective.
- We suppose data are collected via simple random sampling, so that we will not consider methods for data from complex surveys – that's coming in the next lecture.

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 counts Y_t over $t = 1, \ldots, 60$ months as

$$Y_t | p \sim \text{Binomial}(n, p_t),$$

  with $n = 10, 20, 200$ and $p_t = 0.2$ (shown in blue).
- In the top plot in particular, we might conclude large temporal variation, but it's a mirage, all we are seeing is sampling variation.

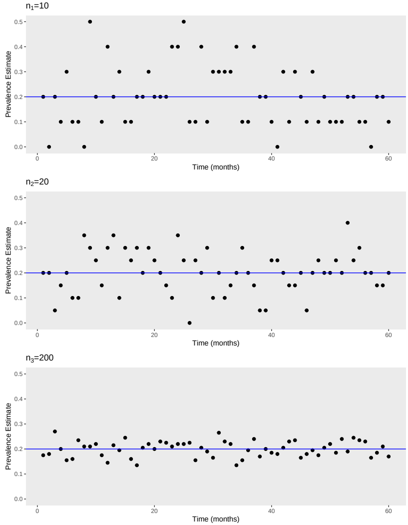


Figure 1: Simulated prevalence estimates over time.

Motivation for Smoothing: Temporal Case

- Figure 2 summarizes estimates from a second simulation.
- The sample sizes are again $n = 10, 20, 200$ but now there is a non-constant temporal pattern (shown in blue).
- For these data, we would not want to oversmooth and remove the trend.
- Later we will apply temporal smoothing models to these two sets of data.

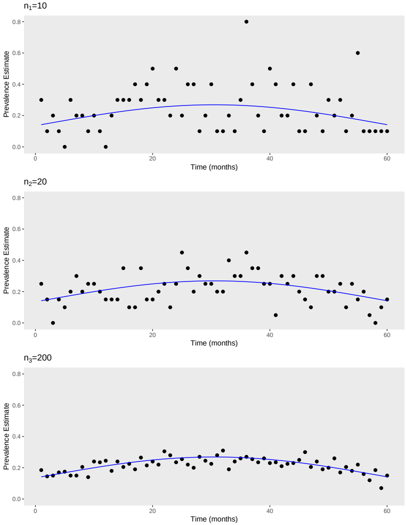


Figure 2: Simulated prevalence estimates over time.

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 the same sample size n in each area:

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

- We look at varying sample sizes $n = 50, 100, 500$, 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.

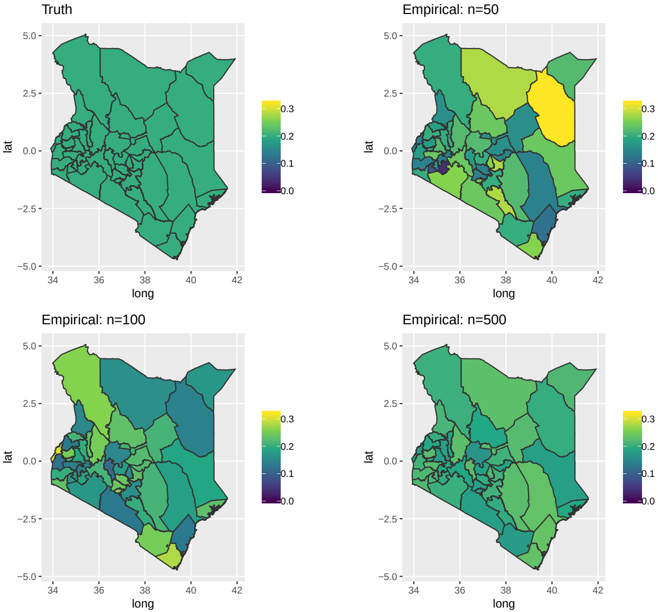


Figure 3: Prevalence estimates over space for simulated data with sample sizes of $n = 50, 100, 500$. True prevalence is 0.2 in all areas.

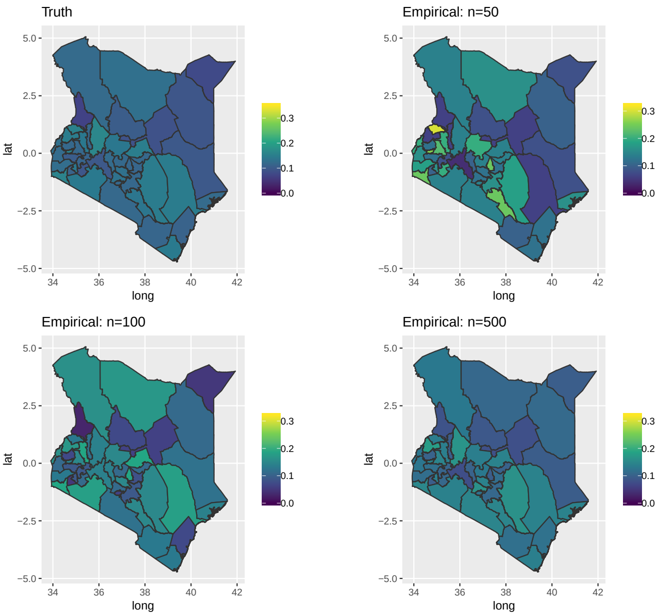


Figure 4: Prevalence estimates over space for simulated data with sample sizes of $n = 50, 100, 500$. True prevalence is spatially varying (top left).

Smoothing

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

- The true underlying prevalences are all the same.
- The true underlying prevalences are distinct but not linked.
- The true underlying prevalences 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. This case corresponds to penalizing deviations from an overall level.
- The prevalences are correlated over time – this case corresponds to penalizing deviations from a local level.

- These are both examples of hierarchical or random effects models — a key element is estimating the smoothing parameter.

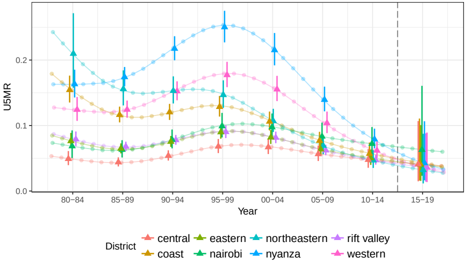


Figure 5: Temporally smoothed U5MR estimates in Kenya.

Bayesian Inference

Bayesian Inference

Bayesian modeling is convenient for implementing notions of smoothing.

There are two key elements that must be specified:

- The sampling model (likelihood) describes the distribution of the data – this model depends on unknown parameters, that we will denote θ.
- The prior distribution expresses beliefs about the parameters θ and provides a mechanism by which penalization/smoothing can be imposed.

These elements are probabilistically combined 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:

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

Bayesian Inference

- In a Bayesian analysis the complete set of unknowns (parameters) is summarized via the multivariate posterior distribution – one or two dimensional marginal posterior distributions can be vizualised.
- 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.

Bayesian Computation

- The computations required for Bayesian inference (integrals) are often not trivial and may be carried out using a variety of analytical, numerical and simulation based techniques.
- We use the integrated nested Laplace approximation (INLA), introduced by Rue *et al.* (2009).
- `R-INLA` is a package that implements the INLA approach.
- Book-length treatments on INLA:
  - Blangiardo and Cameletti (2015) – space-time modeling.
  - Wang *et al.* (2018) – general modeling.
  - Krainski *et al.* (2018) – advanced space-time modeling.

Bayes Example

- Imagine the data model is normal with an unknown mean μ:
$$y_i \mid \mu \sim \mathrm{N}(\mu, \sigma^2),$$
$i = 1, \ldots, n$, with σ^2 assumed known.
- This is equivalent to:
$$\overline{y} \mid \mu \sim \mathrm{N}(\mu, \sigma^2/n),$$
where $\sigma/\sqrt{n}$ is the standard error.
- Suppose a normal prior is appropriate:
$$\mu \sim \mathrm{N}(m, v),$$
so that values of the mean μ that are (relatively) far from m are penalized – v is the smoothing parameter, more/less smoothing if small/big.
- The log posterior is:
$$\underbrace{\log p(\mu \mid y}_{\textbf{Updated Beliefs}}) = -\underbrace{\frac{n}{2\sigma^2}(\overline{y} - \mu)^2}_{\textbf{Data Model}} - \underbrace{\frac{1}{2v}(\mu - m)^2}_{\textbf{Penalization}}.$$

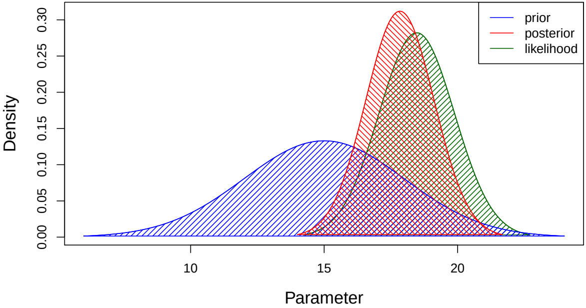


Figure 6: Normal data model with $n = 10$, $\overline{y} = 19.3$ and standard error 1.41. The prior for μ has mean $m =$15 and $v = 3^2$. The posterior for the parameter μ is a compromise between the two sources of information: the posterior mean is 18.5 and the posterior standard deviation is 1.28.

Temporal Smoothing

Smoothing over Time

In the simple normal example, we had a single parameter that was pulled towards the prior mean – now we consider the situation in which we have multiple parameters, indexed by time, and we wish to specify a joint prior that encourages similarity between "close-by" means.

Rationale and overview of models for temporal smoothing:

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

Random Walk Models

We use random walk models which encourage the mean of the response at time t, ϕ_t, to be similar to its neighbors.

We will start by describing the model for a continuous, unbounded variable, i.e., not constrained to lie between 0 and 1 like the prevalence is – to be concrete we will describe the model in the context of a logit prevalence.

Random walk models can be described in different (equivalent) ways:

1. Via the joint distribution: $p(\phi_1, \ldots, \phi_T)$.
2. One-dimensional distributions conditional on the past: $p(\phi_t | \phi_{t-1}, \phi_{t-2}, \ldots)$.
3. One-dimensional distributions conditional on neighbors: $p(\phi_t | \boldsymbol{\phi}_{-t})$ where $\boldsymbol{\phi}_{-t} = [\phi_1, \ldots, \phi_{t-1}, \phi_{t+1}, \ldots, \phi_T]$ is the full collection with the t-th entry removed.

The first and third can be easily generalized to spatial processes.

Random Walk Models

In the first-order random walk model, denoted RW1, the mean of the logit prevalence at time t, ϕ_t , is modeled as a function of its immediate neighbors via:

$$\phi_t \mid \phi_{t-1}, \phi_{t+1}, \sigma^2 \sim \mathrm{N}\left(\frac{1}{2}(\phi_{t-1} + \phi_{t+1}), \frac{\sigma^2}{2}\right),$$

where σ^2 is a smoothing parameter:

- the smoothing parameter is estimated from the data and determines the extent to which deviations from the neighbors are penalized,
- small values enforce strong smoothing,
- large values enforce weak smoothing.

Random Walk Models

- The explicit penalty term for the RW1 model is:

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

- Hence:
  - Values of ϕ_t that are close to $\frac{1}{2}(\phi_{t-1} + \phi_{t+1})$ are favored (higher density).
  - The relative favorability is governed by σ^2 – if this variance is small, then ϕ_t can't stray too far from its neighbors.
- S time step ahead predictions from the RW1 model follow distribution

$$\phi_{T+S} | \phi_1, \ldots, \phi_T, \sigma^2 \sim \mathrm{N}(\phi_T, \sigma^2 \times S),$$

so that the level is determined by the last value ϕ_T and the variance increases linearly in S

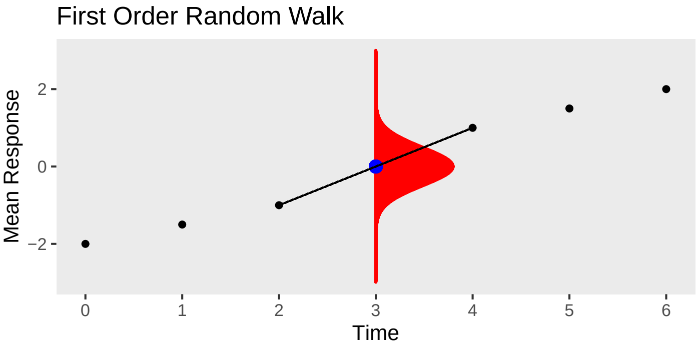


Figure 7: Illustration of the RW1 model for smoothing at time 3. The mean of the smoother is the average of the two adjacent points (and is highlighted as •), and deviations from this mean are penalized via the normal distribution shown in **red**.

RW1 Model

- Letting $\boldsymbol{\phi} = [\phi_1, \dots, \phi_T]^\mathsf{T}$, the form of the prior RW1 density is:

$$
\begin{aligned}
p(\boldsymbol{\phi}|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(\phi_{t+1}-\phi_t)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{t\sim t'}(\phi_t-\phi_{t'})^2\right) = \exp\left(-\frac{1}{2}\boldsymbol{\phi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\phi}\right)
\end{aligned}
$$

  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 random walk (RW2) model produces smoother trajectories than the RW1, and has more reasonable short term predictions, which is desirable for many health and demographic indicators.
- The model is defined in terms of second differences:

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

  showing that deviations from linearity are discouraged.
- S time step ahead predictions from the RW2 model follow distribution

$$\phi_{T+S} \mid \phi_1, \dots, \phi_T, \sigma^2 \sim \mathrm{N}\left(\phi_T + S(\phi_T - \phi_{T-1}), \frac{\sigma^2}{6} \times S(S+1)(2S+1)\right),$$

  so that the mean is a linear function of the values at the last two time points, and the variance is cubic in the number of periods S, so blows up very quickly.

RW2 Model

- Form of the prior RW2 density is:

$$
\begin{aligned}
p(\phi|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T-2}(\phi_{t+2}-2\phi_{t+1}+\phi_t)^2\right) \\
&= \exp\left(-\frac{1}{2}\phi^{\mathsf{T}}\boldsymbol{Q}\phi\right)
\end{aligned}
$$

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, so again sparse.

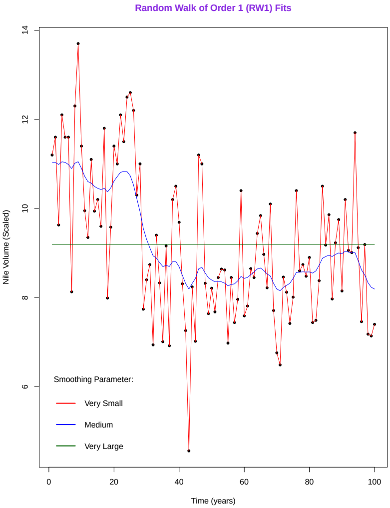


Figure 8: Nile data with RW1 fits under different priors for smoothing parameter $1/\sigma^2$.

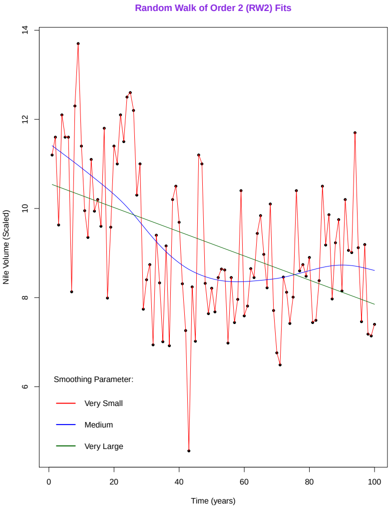


Figure 9: Nile data with RW2 fits under different priors for smoothing parameter $1/\sigma^2$.

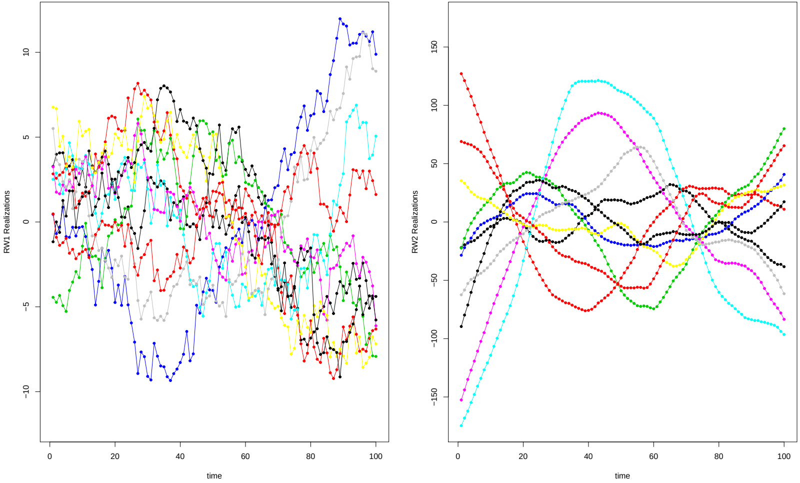


Figure 10: Ten simulated realizations from RW1 (left) and RW2 (right) models. In both cases $\sigma^2 = 1$.

Temporal Smoothing Model Summary

We have three models:

IID MODEL:

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

smooth towards zero.

RW1 MODEL:

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

smooth towards the previous value.

RW2 MODEL:

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

smooth towards the previous slope.

The RWs are examples of Gaussian Markov Random Field (GMRF) models, which have many appealing properties, and are computationally convenient (Rue and Held, 2005).

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 11 and 12 the fitted values are shown in red – in both examples the reconstruction is reasonable.

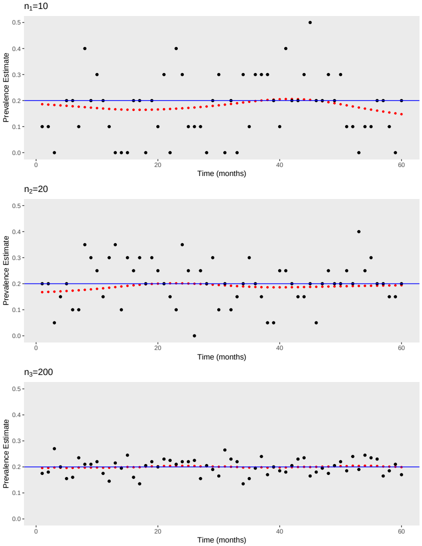


Figure 11: Prevalence estimates over time from simulated data, true prevalence $p = 0.2$ (blue solid lines). Smoothed random walk estimates in **red**.

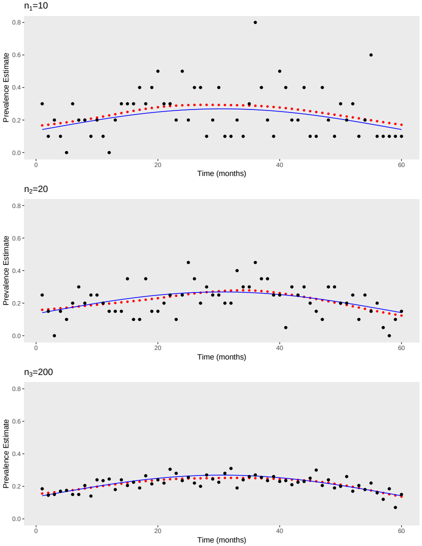


Figure 12: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line. Smoothed random walk estimates in **red**.

RW2 Model Applied to Ecuador Survey Data

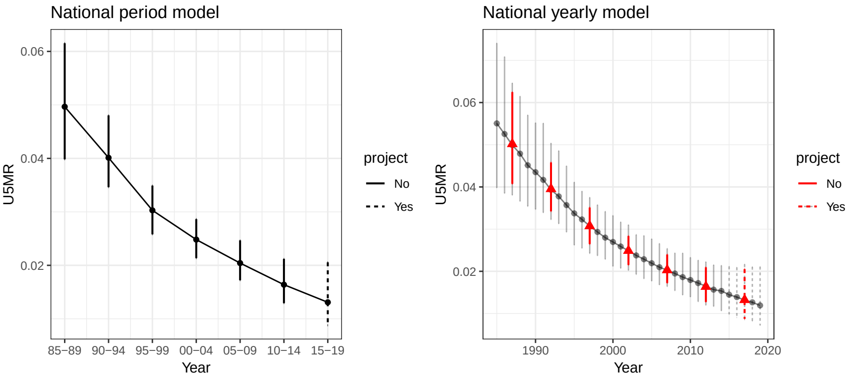
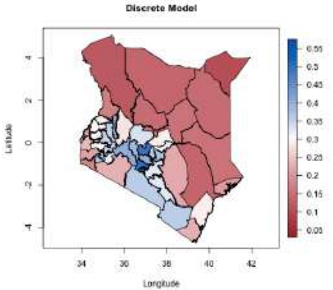


Figure 13: Yearly RW2 smoothing of weighted estimates of under-5 mortality in Ecuador, with 95% uncertainty intervals. On the left we apply to data aggregated over 5 years and on the right by 1 year. The dashed lines on the right of each plot are projections.

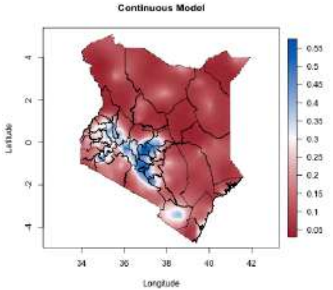
Spatial Smoothing

Two Approaches to Spatial Modeling

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



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



Spatial Models for Binomial Data

Point Data: Suppose we carry out sampling at a cluster with location $\boldsymbol{s}_c$:

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

- Discrete spatial random effects:

$$p(\boldsymbol{s}_c) = \text{expit}(\alpha + \underbrace{S(\boldsymbol{s}_{i[s_c]})}_{\text{Discrete Spatial}} + \underbrace{e_{i[s_c]}}_{\text{Discrete Random Noise}}),$$

  where $i[s_c]$ is the spatial area within which the cluster at $\boldsymbol{s}_c$ lies.

- Continuous spatial random effects:

$$p(\boldsymbol{s}_c) = \text{expit}(\alpha + \underbrace{S(\boldsymbol{s}_c)}_{\text{Continuous Spatial}} + \underbrace{\epsilon_c}_{\text{“Nugget” Random Noise}}).$$

Spatial Models for Binomial Data

Aggregate Data: For a count in area i,

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

- Discrete spatial random effects:

$$p_i = \text{expit}(\alpha + \underbrace{S_i}_{\text{Discrete Spatial}} + \underbrace{e_i}_{\text{Discrete Random Noise}}).$$

- Continuous spatial random effects:

$$\begin{aligned} p_i &= \int_{s \in A_i} p(\boldsymbol{s}) d(\boldsymbol{s})\, d\boldsymbol{s} \\ &= \int_{s \in A_i} \text{expit}(\quad \alpha + \underbrace{S(\boldsymbol{s})}_{\text{Continuous Spatial}} \quad)\, d(\boldsymbol{s})\, d\boldsymbol{s}, \end{aligned}$$

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

The BYM Model

In the popular BYM model Besag, York and Mollié (1991):

$$p_i = \text{expit}(\alpha + S_i + e_i),$$

where

- $e_i \sim_{iid} \text{N}(0, \sigma_e^2)$.
- The spatial effects S_i are modeled conditional on the neighbors:

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

  where $\bar{s}_i = \frac{1}{m_i} \sum_{j \sim i} s_j$ is the mean of the neighbors 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.
- This form is known as an intrinsic conditional autoregression (ICAR) and is the spatial extension of the RW1 model.

Example of a Neighborhood Scheme

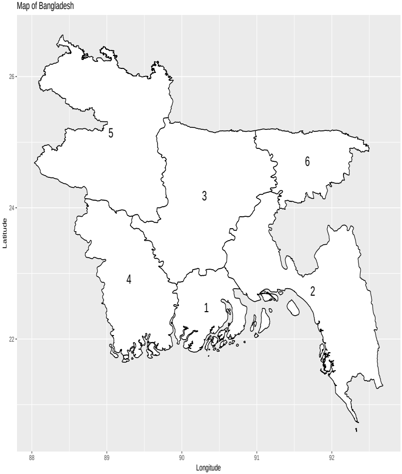


Figure 14: Common boundary neighbor scheme for Bangladesh divisions.

ICAR prior for spatial random 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, $\boldsymbol{Q} = \boldsymbol{R}/\sigma_s^2$, $\boldsymbol{R}$:

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

Continuous Spatial Models

- Continuous spatial models are routinely used by both WorldPop (Utazi *et al.*, 2018) and the Institution for Health Metrics and Evaluation (IHME) (Golding *et al.*, 2017), two large producers of maps of health and demographic indicators.
- Here, we focus on discrete spatial models.
- Continuous modeling is theoretically appealing:
  - It avoids the arbitrariness of the discrete spatial model, in which one needs to specify a neighborhood scheme.
  - It allows data with different geographical information to be combined.
- However, continuous modeling requires greater statistical and computational expertise, and so is more hazardous in practice.
- Also, for area-level summaries, one needs to aggregate the continuously varying prevalence surface with respect to a population density surface $d(\boldsymbol{s})$, $\int_{A_i} p(\boldsymbol{s})d(\boldsymbol{s})\, d\boldsymbol{s}$, which can be sensitive to the choice of $d(\boldsymbol{s})$.

For more discussion and a comparison of discrete and spatial models, see Wakefield *et al.* (2019).

Comparison of Spatial Models

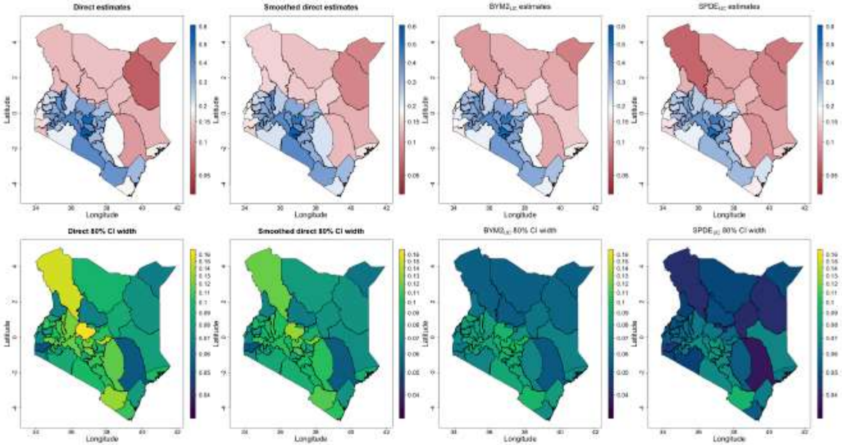


Figure 15: Kenya county level 2014 secondary education predictive estimates (top) and 80% uncertainty interval width (bottom) for women aged 20–29.

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:

$$\frac{p_i}{1 - p_i} = \exp(\alpha + S_i + e_i),$$

where:

- $e_i \sim_{iid} \text{N}(0, \sigma_e^2)$ represent independent "shocks" with no spatial structure in area i.
- $[S_1, \ldots, S_n]$ are ICAR(σ_S^2).
- The model is completed with priors on σ_e^2 and σ_S^2 – we won't discuss here, but see Simpson *et al.* (2017).

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

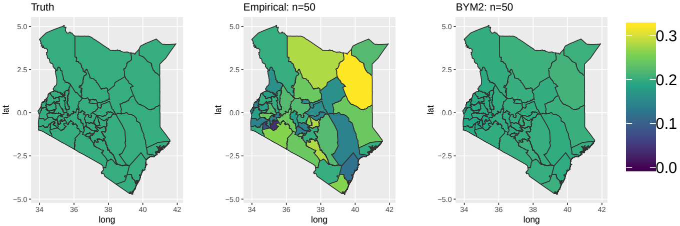


Figure 16: Results with $n = 50$ when true prevalence is 0.2. Left: Truth. Middle: Raw proportions. Right: Smoothing with BYM.

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

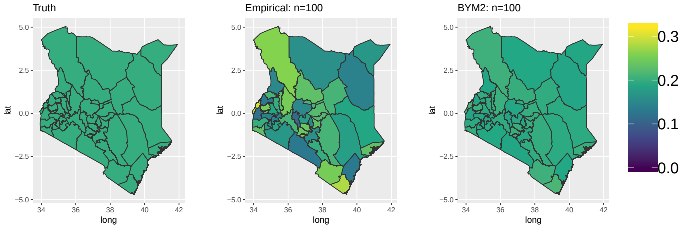


Figure 17: Results with $n = 100$ when true prevalence is 0.2. Left: Truth. Middle: Raw proportions. Right: Smoothing with BYM.

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

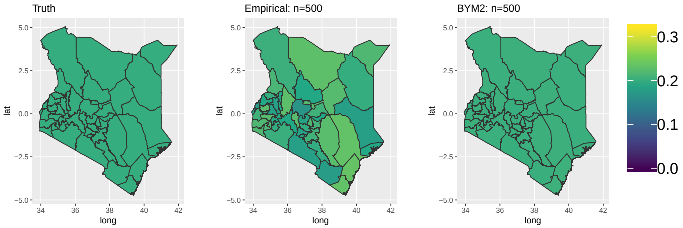


Figure 18: Results with $n = 500$ when true prevalence is 0.2. Left: Truth. Middle: Raw proportions. Right: Smoothing with BYM.

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

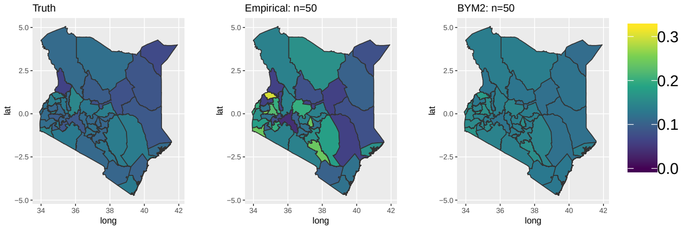


Figure 19: Results with $n = 50$ when true prevalence is varying. Left: Truth. Middle: Raw proportions. Right: Smoothing with BYM.

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

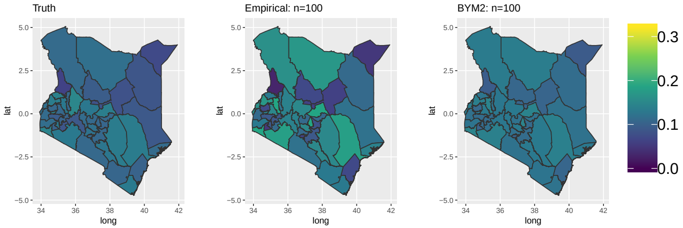


Figure 20: Results with $n = 100$ when true prevalence is varying. Left: Truth. Middle: Raw proportions. Right: Smoothing with BYM.

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

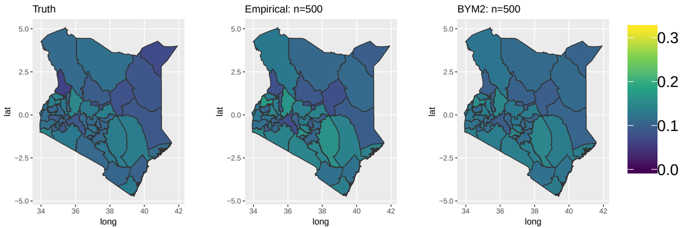


Figure 21: Results with $n = 500$ when true prevalence is varying. Left: Truth. Middle: Raw proportions. Right: smoothing with BYM.

2013 Nigeria DHS

- In Nigeria, almost a third of the Admin2 areas have no data (colored white in left plot below).
- We fit a discrete spatial model in which the rates in neighboring areas (as defined by sharing a boundary) are “encouraged” to be similar (right plot below).

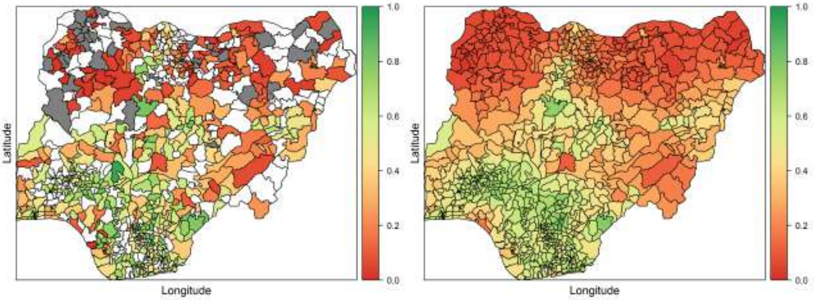


Figure 22: Vaccination prevalences in Nigeria in 2013. Left: Weighted estimates. Right: Estimates from BYM smoothing model.

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

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

- **Interpretation:** η_i is the effect of being at level i for factor A, regardless of the level assumed by B, and ϕ_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

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

- **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 an “ordering” and we might expect similarity.

Main Effects Model

- First, consider the space-time model,

$$
\begin{aligned}
Y_{it}|p_{it} &\sim \text{Binomial}(n_{it}, p_{it}) \\
p_{it} &= \text{expit}(\alpha + e_i + S_i + \omega_t + \phi_t)
\end{aligned}
$$

- Components:
  - Unstructured spatial term $e_i \sim_{iid} \text{N}(0, \sigma_e^2), i = 1, \ldots, n$.
  - Smooth spatial term $[S_1, \ldots, S_n]$ smooth in space, e.g., from an ICAR model.
  - Unstructured temporal term $\omega_t \sim_{iid} \text{N}(0, \sigma_\omega^2), t = 1, \ldots, T$.
  - Smooth temporal term $[\phi_1, \ldots, \phi_T]$ 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:

$$p_{it} = \text{expit}(\alpha + e_i + S_i + \omega_t + \phi_t + \delta_{it}),$$

  with e_i, S_i, ω_t, ϕ_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:

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

- `R-INLA` implements this model.

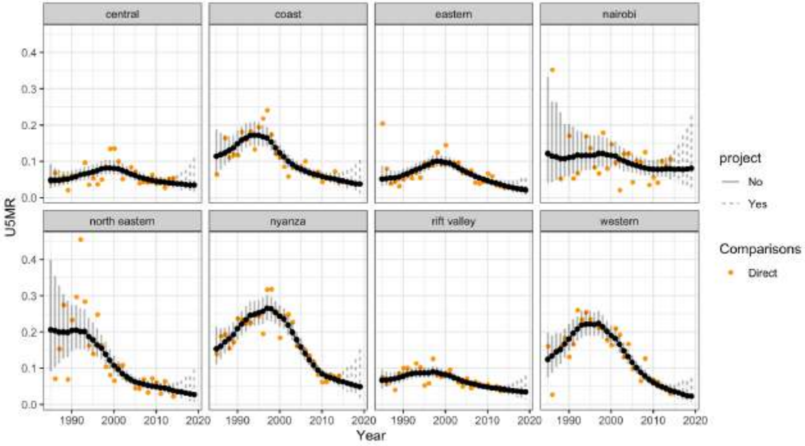


Figure 23: Weighted estimates and smoothed fits over time for 8 provinces of Kenya.

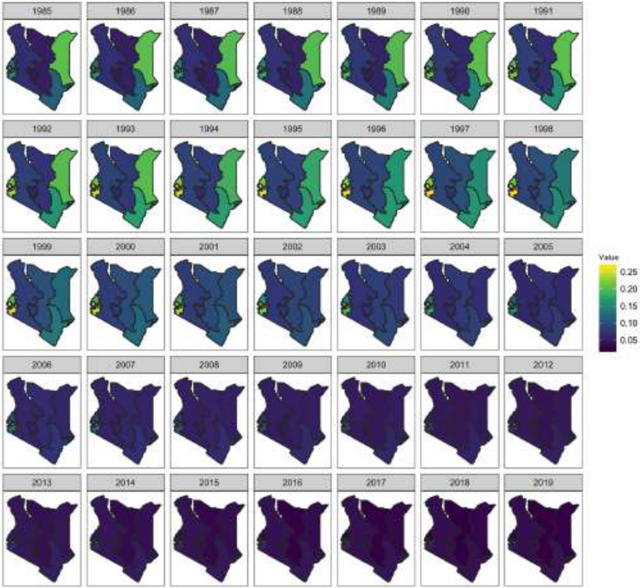


Figure 24: Maps of smoothed estimates over time for 8 provinces of Kenya.

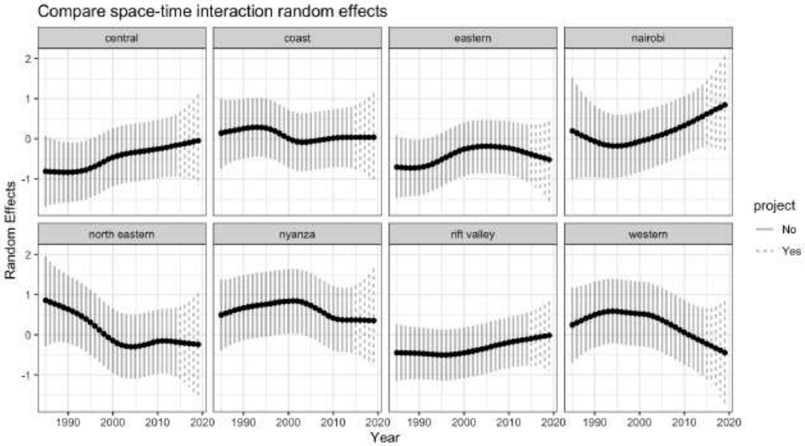


Figure 25: Space-time interactions δ_{it} for 8 provinces of Kenya.

Discussion

Discussion

- Lots of applications have used discrete spatial models – they are well understood and relatively easy to use.
- Computation for discrete spatial models is fast.
- Continuous spatial models pose greater modeling and computational challenges.
- These difficulties mean that continuous spatial models are harder to "package up".
- In the next lecture we see how the spatial and space-time models we have considered here can be used with data collected from complex surveys.

Books

Bayes:

- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B., *Bayesian Data Analysis*, Chapman and Hall/CRC Press.
- Hoff, P.D. (2009), *A First Course in Bayesian Statistical Methods*, Springer.
- Wakefield, J. (2013). *Bayes and Frequentist Regression Methods*, Springer.

Spatial:

- Banerjee, S., Gelfand, A.E. and Carlin, B.P. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*, Chapman and Hall/CRC Press.
- Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*, Second Edition, Springer.
- Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*, Wiley.

Acknowledgments

This lecture series was supported by the Hewlett Foundation and the International Union for the Scientific Study of Population (IUSSP).

The research reported in this series has grown out of a longstanding collaboration between Jon Wakefield, Zehang Richard Li and Sam Clark.

Many other people have contributed, however, for full details and for links to other aspects of this work, check out:

http://faculty.washington.edu/jonno/space-station.html

References

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.

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 neontal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet*, **390**, 2171–2182.

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.

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*. Published January 22, 2019.

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., 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.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, **36**, 1583–1591.

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.

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