Bayesian SAE using Complex Survey Data: Methods and Applications
Lecture 3: Small Area Estimation

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

CSSS Short Course: Thursday 23rd May, 2019

Outline

Motivation

Motivation

- Small area estimation (SAE) is an important endeavor since many agencies require estimates of health, demographic, economic indices, education and environmental measures in order to plan and allocate resources and target interventions.
- SAE is an example of domain (sub-population) estimation.
- “Small” here refers to the fact that we will typically base our inference on a small sample from each area (so it is not a description of geographical size).
- In the limit there may some areas in which there are no data.

Small Area Estimation

- Consider a study region partitioned into n disjoint and exhaustive areas, labeled by i, $i = 1, \ldots, n$.
- As a concrete example, suppose we are interested in a particular condition so that the response is a binary outcome, Y_{ik}, for $k = 1, \ldots, N_i$, individuals in area i.
- Based on samples that are collected in the areas (though some areas may contain no samples), common targets of estimation are of:
  - The population totals:

$$T_i = \sum_{k=1}^{N_i} Y_{ik}.$$

  - The prevalence of the condition in each area:

$$P_i = \frac{1}{N_i} \sum_{k=1}^{N_i} Y_{ik} = \frac{T_i}{N_i}.$$

Background reading on SAE

- The classic text on SAE is Rao and Molina (2015); not the easiest book to read, and little material on spatial smoothing models.
- An excellent review of SAE is Pfeffermann (2013) (though not much on spatial models).
- The SAE literature distinguishes between direct estimation, in which data from the area only is used to provide the estimate in an area, and indirect estimation, in which data from other areas is used to provide the estimate.

Non-Spatial Methods for SAE

Design based inference based on weighted estimators

- Suppose we undertake a complex design and obtain outcomes y_{ik} in area i, $k \in s_i$, where s_i is the set of samples that were in area i.
- Along with the outcome, there is an associated design weight w_{ik}.
- Under the design-based approach to inference, it is common to use the weighted estimator of the prevalence:

$$\widehat{P}_i = \frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}}.$$

  There is an associated variance, that acknowledges the design, $\widehat{V}_i$.
- This variance estimate may be obtained analytically, or through resampling techniques such as the jackknife.
- Asymptotically (that is, in large samples):

$$\widehat{P}_i \sim \mathrm{N}(P_i, V_i).$$

Direct Estimation

- The simplest approach is to simply map the direct estimates $\widehat{P}_i$.
- To assess the uncertainty, one may map the lower and upper ends of (say) a 95% confidence interval:

$$\widehat{P}_i \pm 1.96 \times \sqrt{\widehat{V}_i}.$$

- If the samples in each area are large, so that $\widehat{V}_i$ is small, then this approach works well.

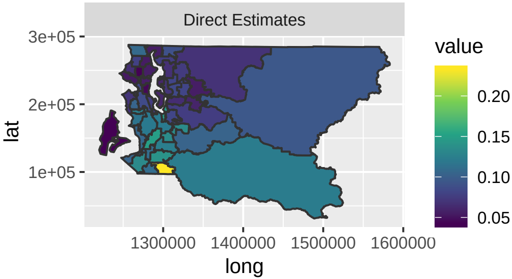


Figure 1: Direct estimates of diabetes prevalence for HRAs in King County.

- We would like to carry out some form of smoothing, but in the case of complex survey sampling, how should we proceed?

Lower and Upper Endpoints of 95% CI

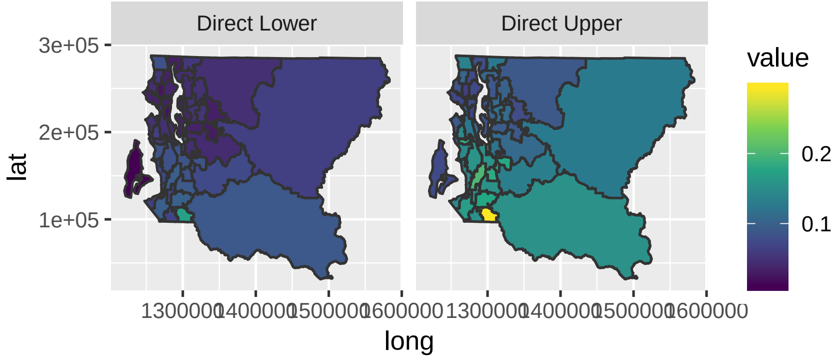


Figure 2: 2.5% and 97.5% points from asymptotic confidence interval.

Synthetic Estimates

Many approaches have been suggested to obtain estimators with greater precision – we discuss three, to give a flavor.

We consider estimation of a generic mean.

- **Synthetic Estimator:**

$$\widehat{Y}_i^{syn} = \frac{1}{N_i} \sum_{k=1}^{N_i} \boldsymbol{x}_{ik}^{\top} \widehat{B},$$

  where

$$\widehat{B} = \left[\sum_{i=1}^{n} \sum_{k \in S_i} w_{ik} \boldsymbol{x}_{ik}^{\top} \boldsymbol{x}_{ik} \right]^{-1} \sum_{i=1}^{n} \sum_{k \in S_i} w_{ik} \boldsymbol{x}_{ik}^{\top} y_{ik}.$$

- Note: Covariates needed for all of population.
- Assumes regression model is appropriate for all areas.
- In general gives high precision estimates – variance is $O(1/n)$, but possibility of large bias.

Survey-Regression Estimates

- **Survey-Regression Estimator:**
  In order to deal with the potential large bias, the bias is estimated to give

$$\begin{aligned}\widehat{Y}_i^{s-r} &= \frac{1}{N_i}\sum_{k=1}^{N_i} \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{B} + \frac{1}{N_i}\sum_{k\in S_i} w_{ik}(y_{ik} - \boldsymbol{x}_{ik}^{\mathsf{T}}\widehat{B} \\ &= \widehat{Y}_i^{ht} + (\overline{\boldsymbol{X}}_i - \widehat{\overline{\boldsymbol{X}}}_i^{ht})^{\mathsf{T}}\widehat{B}\end{aligned}$$

  where $\widehat{Y}_i^{ht}$ and $\widehat{\overline{\boldsymbol{X}}}_i^{ht}$ are the Horvitz-Thompson estimates of $\overline{Y}_{Ui}$ and $\boldsymbol{X}_i$.
- Variance is unfortunately $O(1/n_i)$.
- **Composite estimator** is of the form

$$\widehat{\overline{Y}}_i^{com} = \delta_i \widehat{Y}_i^{s-r} + (1-\delta_i)\widehat{Y}_i^{syn}$$

  with $0 \leq \delta_i \leq 1$ estimated in such a way that for larger n_i we have larger δ_i.

Spatial Methods for SAE

Smoothed Direct Estimation

- Let $\widehat{P}_i$ be the weighted estimator of a prevalence P_i, then consider
$$\widehat{\theta}_i = \text{logit}\ (\widehat{P}_i) = \log\left(\frac{\widehat{P}_i}{1-\widehat{P}_i}\right),$$
which is on the whole of the real line.
- "Data" Model: We take as data the estimator:
$$\widehat{\theta}_i \sim \text{N}(\theta_i, \widehat{V}_i),$$
where $\widehat{V}_i$, its variance, is known.
- Prior Random Effects Model:
$$\theta_i = \beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta} + \epsilon_i,$$
where $\boldsymbol{x}_i$ are area-level covariates and the random effects $\epsilon_i \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$.

Smoothed Direct Estimation

- Fay and Herriot (1979) suggested this hierarchical model, in a landmark paper.
- This model acknowledges the design and also smooths, and it is straightforward to add spatial random effects.
- The spatial model was investigated and applied with simulated and real data (without covariates) in Chen *et al.* (2014); Mercer *et al.* (2014) and (in a space-time setting) in Mercer *et al.* (2014, 2015) and Li *et al.* (2019).

Smoothed Direct Estimation

The spatial version of the model has:

- "Data" Model:

$$\widehat{\theta}_i \sim \mathrm{N}(\theta_i, \widehat{V}_i),$$

  where $\widehat{V}_i$ is known variance.
- Prior Model:

$$\theta_i = \beta_0 + \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta} + \epsilon_i + S_i,$$

  with
  - $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$.
  - $S_i \sim \mathrm{ICAR}(\sigma_s^2)$.
- Known as an area-level SAE model.

Smoothed Direct Estimation

- The area-level SAE model has been used by Gutreuter *et al.* (2019) in the context of estimating HIV prevalence and burden in districts of South Africa, using household survey data.
- Among the covariates considered for the prevalence model were:
  - prevalence estimates from antenatal clinics data,
  - population density,
  - percentages of housing units were were “formal dwellings”),
  - dependency ratio (ratio of the numbers of residents aged 15–64 years to those younger than 15 years and older than 64 years,
  - socio-economic quintile,
  - maternal mortality rate.
- The `sae` package is used.
- For more detail on models, see Marhuenda *et al.* (2013).

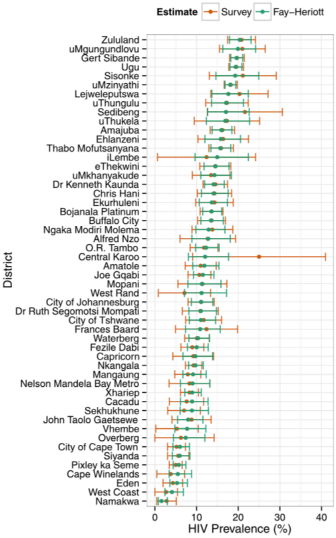


Figure 3: Direct and Fay-Herriot estimates of HIV prevalence in South African districts in 2012, from Gutreuter *et al.* (2019).

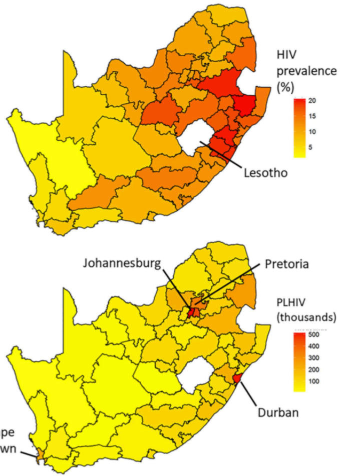


Figure 4: Estimates of HIV prevalence and people living with HIV in South African districts in 2012, from Gutreuter *et al.* (2019).

Smoothed Direct Estimation

The model is implemented in the `R` package `SUMMER` (Martin *et al.*, 2018):

- A design object is created in the `survey` package, and direct estimates formed.
- The `INLA` package to fit the above model.
- It is computationally inexpensive, producing country-specific estimates in seconds.
- Currently undergoing a major upgrade – next version will include many new models, and allow covariates to be included.

Discrete Spatio-Temporal Model

- Let $\widehat{P}_{it}$ be the design-based estimate of a prevalence in area i and period t.
- Take logit of direct estimates $\widehat{P}_{it}$ with appropriate design-based estimator and model as in Mercer *et al.* (2015):

$$
\begin{aligned}
\text{logit}(\widehat{P}_{it}) &\sim \text{N}(\ \theta_{it}, \widehat{V}_{it}\ ) \\
\theta_{it} &= \beta_0 + \underbrace{\omega_t}_{\text{Temporal Smooth}} + \underbrace{\phi_t}_{\text{Temporal Noise}} + \underbrace{S_i}_{\text{Spatial Smooth}} + \underbrace{\epsilon_i}_{\text{Spatial Noise}} + \underbrace{\delta_{it}}_{\text{Interaction}}
\end{aligned}
$$

- Alleviates small sample size problems via temporal, spatial and space-time smoothing.
- Interaction terms are as described by Knorr-Held (2000).

Smoothed Direct Model in Practice (Li *et al.*, 2019)

- The smoothed direct model has been used for 35 African countries to estimate U5MR in Admin-1 regions, by year.
- Data enter at the 5-year level (to give stable variances), but the RWs are defined on the 1-year scale.
- Data:
  - 121 DHS in 35 countries
  - 1.2 million children
  - 192 million child-months
- Takes around 2.5 hours to obtain estimates for all countries – separate models for each country.
- United Nations Inter-agency Group for child Mortality Estimation (UN IGME) involvement:
  - They have supported this research and endorsed these estimates.
  - Methods and software workshops in Ecuador, Jordan and South Africa.
  - Aim is to have health departments produce their own estimates.

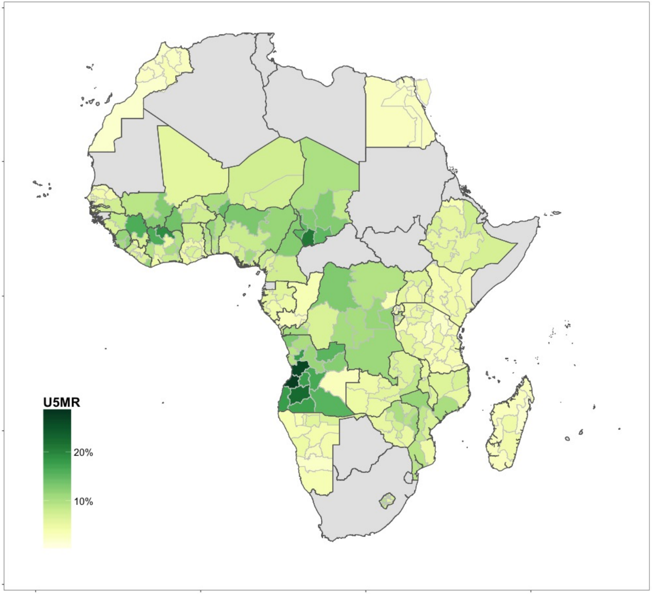


Figure 5: Predictions of U5MR for 2015, in 35 countries of Africa.

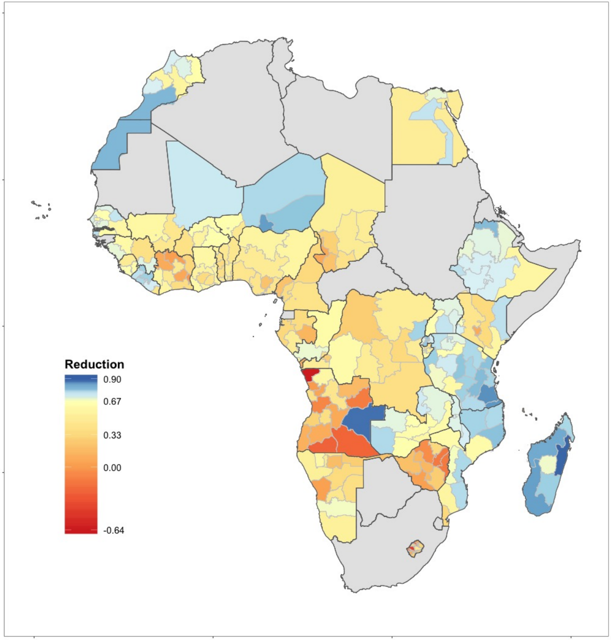


Figure 6: Percent reduction from 1990 to 2015, in 35 countries of Africa.

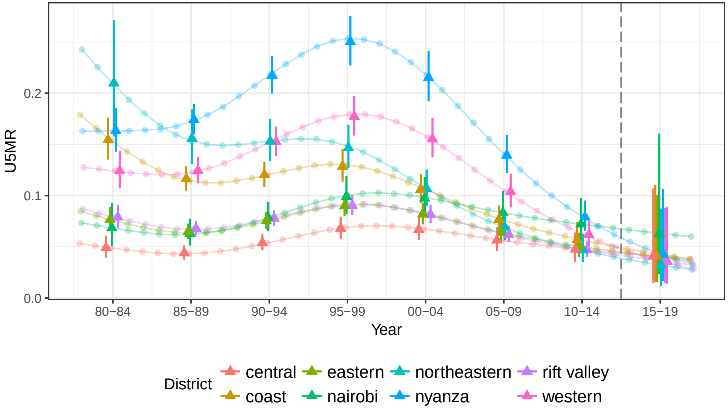


Figure 7: Posterior median estimates for Kenya districts.

Model-Based Continuous Spatial Modeling

- The successful use of smoothed direct estimation hinges on reliable direct (weighted) estimates.
- To produce estimates at a fine spatial/temporal scale, a model-based approach is needed.
- A common approach is we have $Y(\boldsymbol{s}_c, t)$ responses from $N(\boldsymbol{s}_c, t)$ sampled individuals at location $\boldsymbol{s}_c$ in year t, for $c = 1, \ldots, n$, clusters.
- The geostatistical model is,

$$\begin{aligned} Y(\boldsymbol{s}_c, t) \mid q(\boldsymbol{s}_c, t) &\sim \text{Binomial}(\ N(\boldsymbol{s}_c, t), q(\boldsymbol{s}_c, t)\ ) \\ \log\left(\frac{q(\boldsymbol{s}_c, t)}{1 - q(\boldsymbol{s}_c, t)}\right) &= \beta_0 + \gamma I(c \in \text{ urban }) + u(\boldsymbol{s}_c, t) + \epsilon_c \end{aligned}$$

- The **stochastic partial differential equations (SPDE)** approach performs calculations over a triangular mesh.

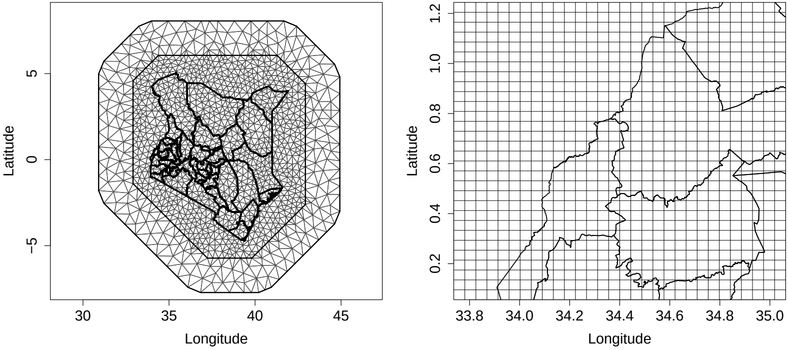


Figure 8: Mesh on which SPDE calculations are carried out (top left), zoomed in grid on which predictions are performed (right).

Surface Reconstructions for U5MR in Kenya

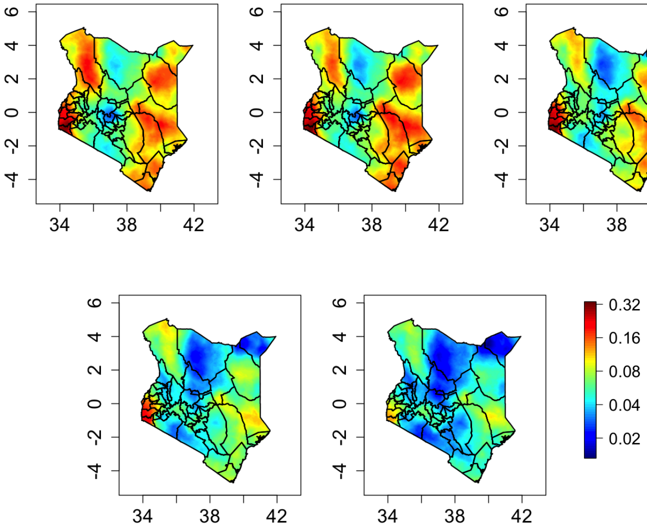


Figure 9: Posterior medians of U5MR for 1990, 1995, 2000, 2005, 2010.

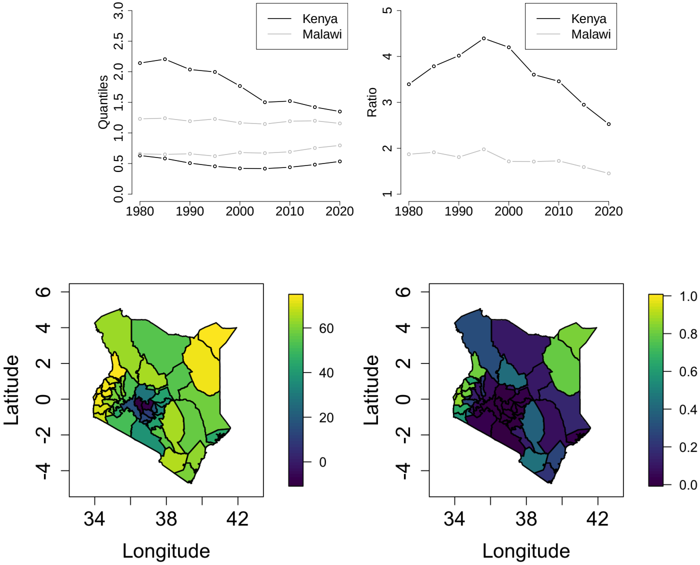


Figure 10: Top row: Kenya and Malawi within-country variability in U5MR (5% and 95% quantiles of pixel distribution). Bottom row: percentage drop from 1990–2015 (left), posterior prob of attaining MDG goal (right), both for Kenya.

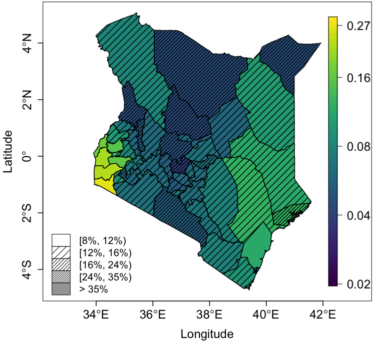
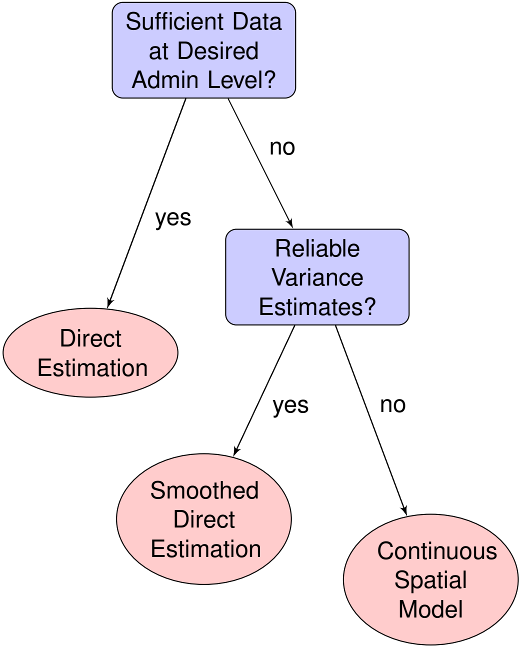


Figure 11: Kenya U5MR estimates in 2000 using discrete spatial model (left), and continuous spatial model (right). Hatching represents uncertainty.

- Point estimates are very similar, but more uncertainty associated with the discrete spatial model estimates.

Recommended Methods for Routine Work



Discussion

Discussion

- Direct smoothed estimates builds on the strengths of weighted estimates and spatial smoothing models.
- In the limit, as we obtain larger data in an area, the weighted estimates will dominate, which is exactly what we want!
- If insufficient samples in areas, then estimated variance is unacceptably large (or undefined), and then we need to resort to continuous spatial modeling which is more difficult.
- Often we will want to combine multiple data sources; the challenge then is modeling differences between study types, e.g., for U5MR estimation, we may have full birth history, summary birth history and vital registration data.

Chen, C., Wakefield, J., and Lumley, T. (2014). The use of sample weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-Temporal Epidemiology*, **11**, 33–43.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Gutreuter, S., Igumbor, E., Wabiri, N., Desai, M., and Durand, L. (2019). Improving estimates of district hiv prevalence and burden in south africa using small area estimation techniques. *PloS one*, **14**, e0212445.

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

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.

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

Martin, B. D., Li, Z. R., Hsiao, Y., Godwin, J., Wakefield, J., and Clark, S. J. (2018). *SUMMER: Spatio-Temporal Under-Five Mortality Methods for Estimation*. R package version 0.2.0.

Mercer, L., Wakefield, J., Chen, C., and Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, **8**, 69–85.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Mosanja, H., and Clark, S. (2015). Small area estimation of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*, **9**, 1889–1905.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

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

Appendix: `R` Packages

- `sae`, by Molina and Marhuenda
  - area-levels (Fay-Herriott (FH), FH with spatial correlation, FH with spatio-temporal correlation) and unit-level models (BHF)
  - estimators: direct Horvitz-Thompson under general sampling designs, post-stratified synthetic estimator and composite estimator
  - fitting and estimation (frequentist) methods: FH, ML, REML, bootstrap
- `rsae`, by Schoch
  - area-levels and unit-level models
  - fitting and estimation (frequentist) methods: ML, Huber-type M-estimation
- `JoSae`, by Breidenbach
  - unit-level models
  - estimators: EBLUP (BHF1988) and GREG (Sarndal 1984)
- `SUMMER` by Martin, Zhang, Wakefield, Clark, Mercer
  - U5MR models using method of Mercer *et al.* (2015).

R Packages

- `hbsae`, by Boonstra
  - area-levels and unit-level models
  - fitting and estimation (frequentist and Bayesian) methods: REML, HB (based on MCMC)
- `mme`, by Lopez-Vizcaino et. al.
  - area-levels multinomial models (area random effects and time random effects)
  - fitting and estimation (frequentist) methods: analytical (PQL and REML) and bootstrap
- `saery`, by Esteban et al.
  - area-level model Rao-Yu 1994
  - fitting and estimation (frequentist) methods: REML
- `sae2`, by Fay and Diallo
  - time series area-level models, Rao-Yu 1994 and extensions
  - fitting and estimation (frequentist) methods: ML and REML

R Packages

- `BayesSAE`, by Shi and Zhang
  - area-levels models: FH and extensions (You-Chapman, spatial models and more)
  - fitting and estimation (Bayesian) methods: HB (based on MCMC)
- `saeSim`, by Warnholz and Schmid
  - useful tools to simulate data for sae studies
- `small area`, by Nandy
  - area-level model (FH)
  - fitting and estimation (frequentist) methods: FH, Prasad and Rao, REML

Note that only `hbsae` and `BayesSAE` use Bayesian methods for the estimation, both use MCMC.