2016 SISMID Module 16
Lecture 10: Small Area Estimation

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Background reading on SAE

The classic text on SAE is Rao (2003), with a new edition just appeared (Rao and Molina, 2015).

Lohr (2010, Chapter 14).

Särndal *et al.* (1992) also have a section, but do not describe model-based approach (their general mantra is model-assisted inference).

Valliant et al. (2000, Section 11.5).

An excellent recent review of SAE is Pfeffermann (2013).

The `sae` package in `R` fits a number of models including Fay-Herriot.

Motivation

Small area estimation (SAE) is an important endeavor since many agencies require estimates of health, 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.

So far we have considered situations in which we have either a complete enumeration of cases and populations, a cross-sectional (random sampling) survey, or case-control sampling.

Design-Based Inference

If survey data are collected from a simple random sample (SRS) then there is no problem in using the methods we have already seen.

Often, however, surveys are carried out in which the design is not SRS.

In particular complex sample schemes are often carried out in which certain populations are disproprtionately sampled, clustered samples are collected,...

In this case, each individual response is accompanied with a weighting factor to account for the unequal probability of selection and also for non-response.

Post-stratification or raking may also be carried out in which estimated population totals of demographic groups are matched to the known totals.

We motivate this section with some examples.

Motivating Example: Diabetes in King County

Arises out of a joint project between Laina Mercer/Jon Wakefield and Seattle and King County Public Health.

Aim we will concentrate on here is to estimate the number of 18 years or older individuals with diabetes, by health reporting areas (HRAs) in King County in 2011.

HRAs are city-based sub-county areas with a total of 48 HRAs in King County. Some of these are as are a single city, some are a group of smaller cities, and some are unincorporated areas. Larger cities such as Seattle and Bellevue include more than one HRA.

Data are based on the question, "Has a doctor, nurse, or other health professional ever told you that you had diabetes?", in 2011.

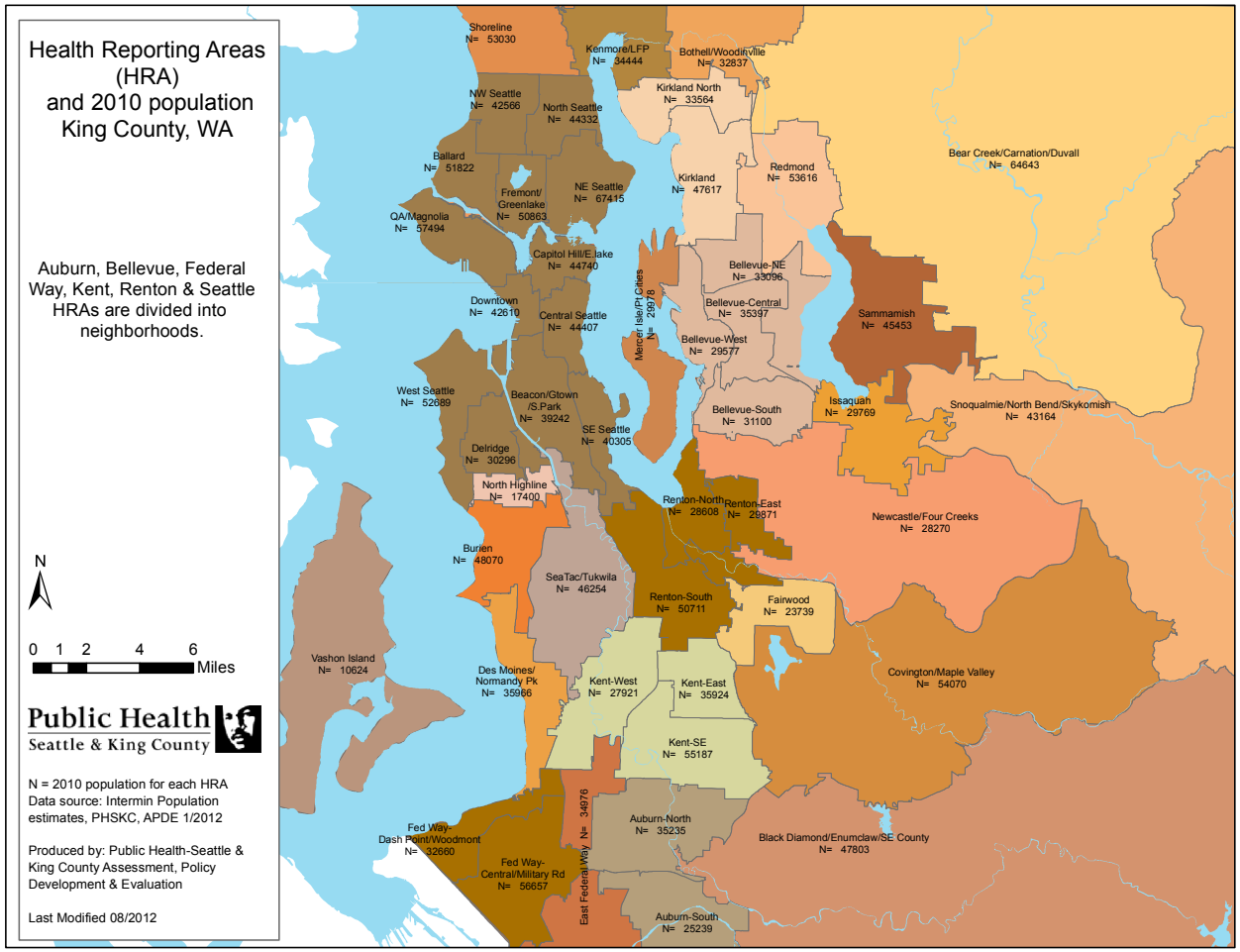

Figure 1 : Health reporting areas (HRAs) in King County.

Motivating BRFSS Example

Estimates are used for a variety of purposes including summarization for the local communities and assessment of health needs.

Analysis and dissemination of place-based disparities is of great importance to allow efficient targeting of place-based interventions.

Because of its demographics, King County looks good compared to other areas in the U.S., but some of its disparities are among the largest of major metro areas.

Estimation is based on Behavioral Risk Factor Surveillance System (BRFSS) data.

The BRFSS is an annual telephone health survey conducted by the Centers for Disease Control and Prevention (CDC) that tracks health conditions and risk behaviors in the United States and its territories since 1984.

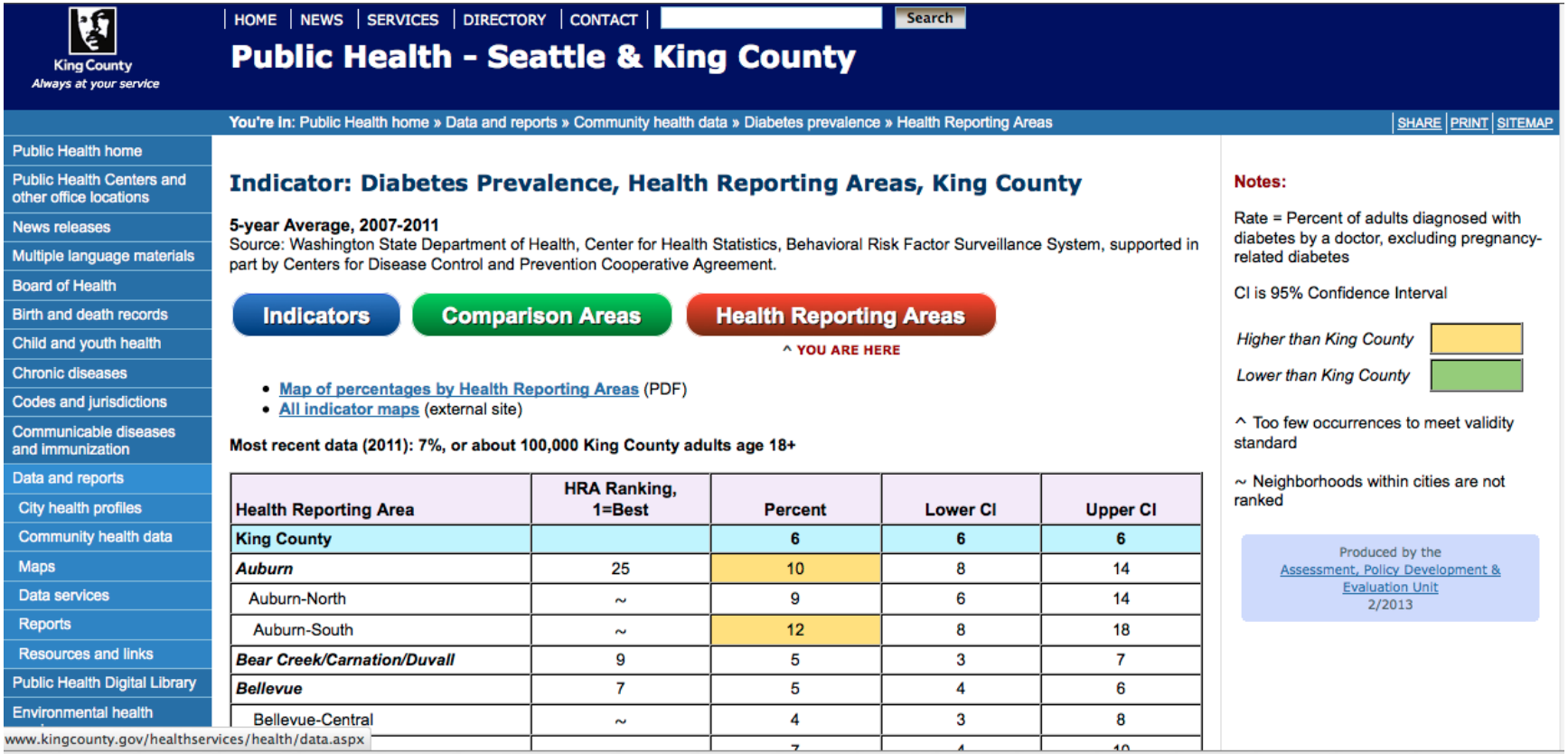

Figure 2 : Public Health: Seattle and King County website.

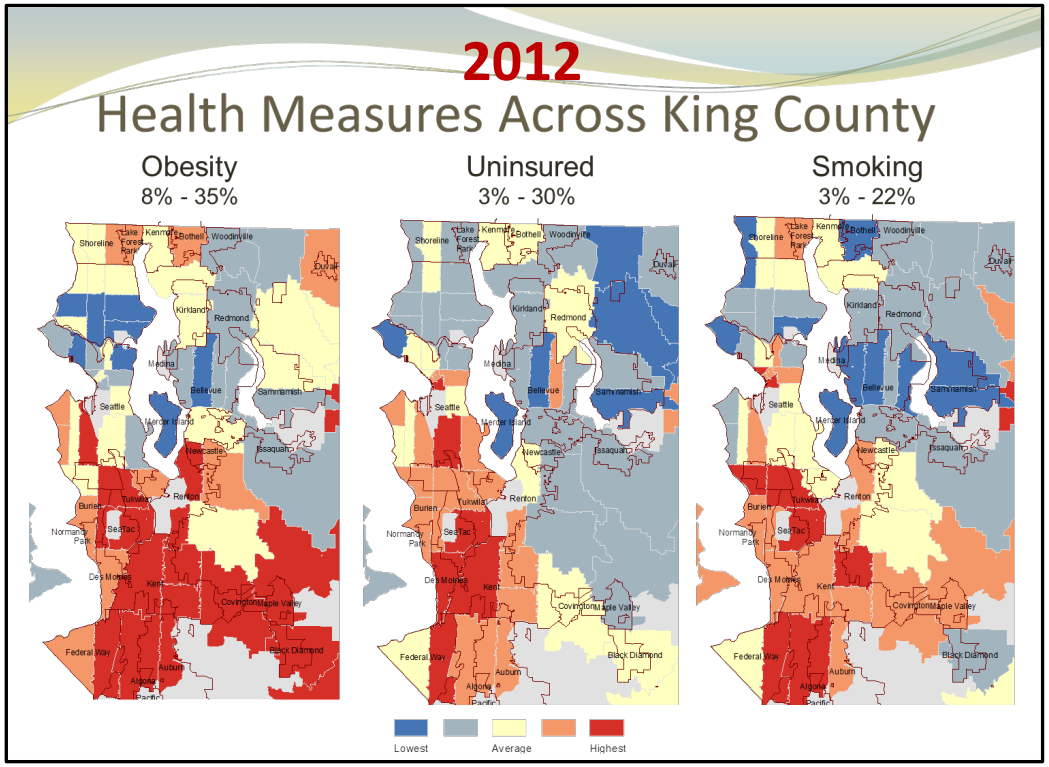

Figure 3 : Summaries from Public Health: Seattle King County.

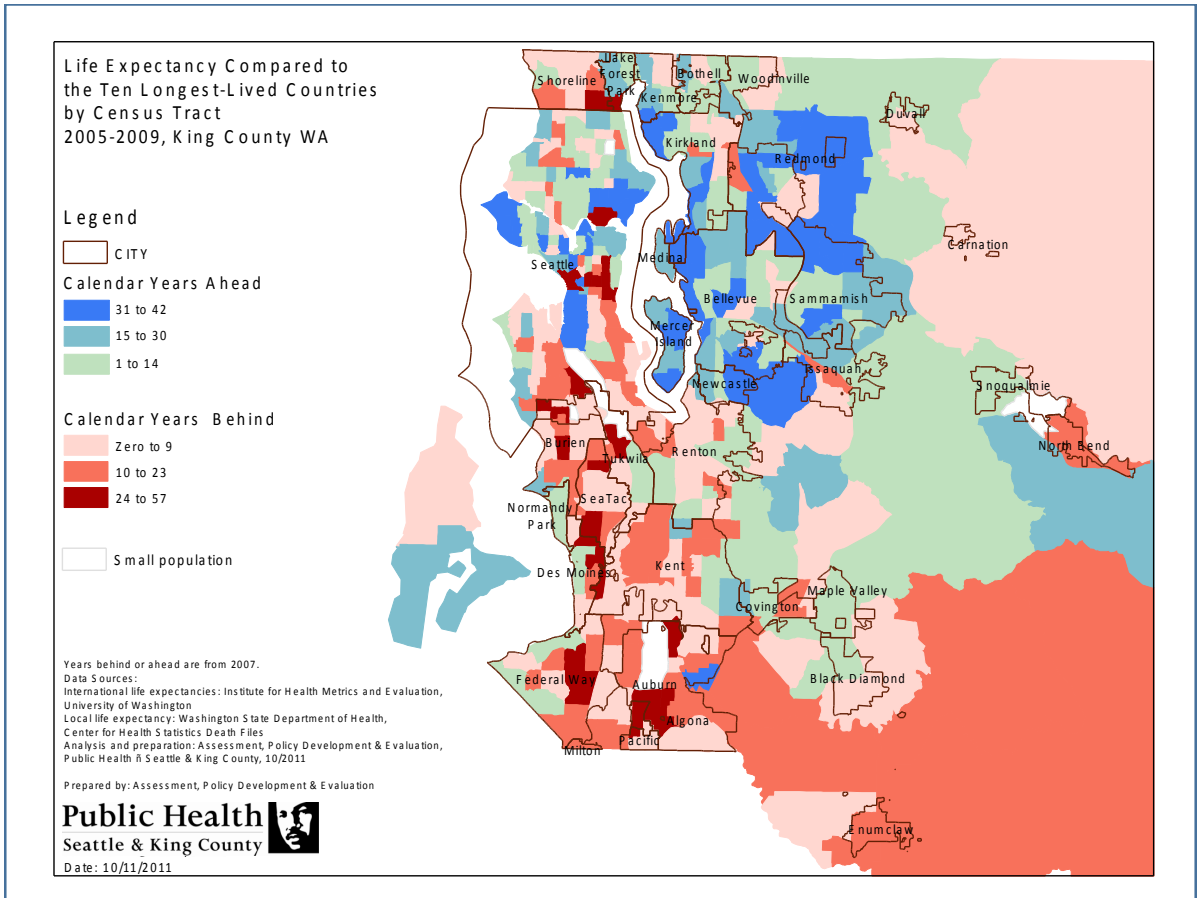

Figure 4 : Summaries from Public Health: Seattle King County.

Motivating BRFSS Example

The BRFSS sampling scheme is complex...

The `Sample Wt`, is calculated as the product of four terms

$$\texttt{Sample Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{No Telephones}} \times \texttt{No Adults} \times \texttt{Post Strat Wt}$$

where `Strat Wt` is the inverse probability of a "likely" or "unlikely" stratum being selected (stratification based on county and "phone likelihood").

Table 1 : Summary statistics for population data, and 2011 King County BRFSS diabetes data, across health reporting areas.

	Mean	*Std. Dev.*	*Median*	*Min*	*Max*	*Total*
Population (>18)	31,619	10,107	30,579	8,556	56,755	1,517,712
Sample Sizes	62.9	24.3	56.5	20	124	3,020
Diabetes Cases	6.3	3.1	6.3	1	15	302
Sample Weights	494.3	626.7	280.4	48.0	5,461	1,491,880

Motivating BRFSS Example

A total of $3,020$ individuals answered the diabetes question.

About 35% of the areas have sample sizes less than 50 (CDC recommended cut-off), so that the diabetes prevalence estimates are unstable in these areas.

We would like to use the totality of the data to aid in estimation in the data sparse areas.

The variability in the weights is high, from 48 to 5,461, with mean 494.

The coefficient of variation (CV) of the weights is 1.27.

Therefore, the inefficiency of using the sample weights under the assumption that unweighted mean is unbiased is about 62%, calculated as $CV^2/(CV^2+1)$ (Korn and Graubard, 1999).

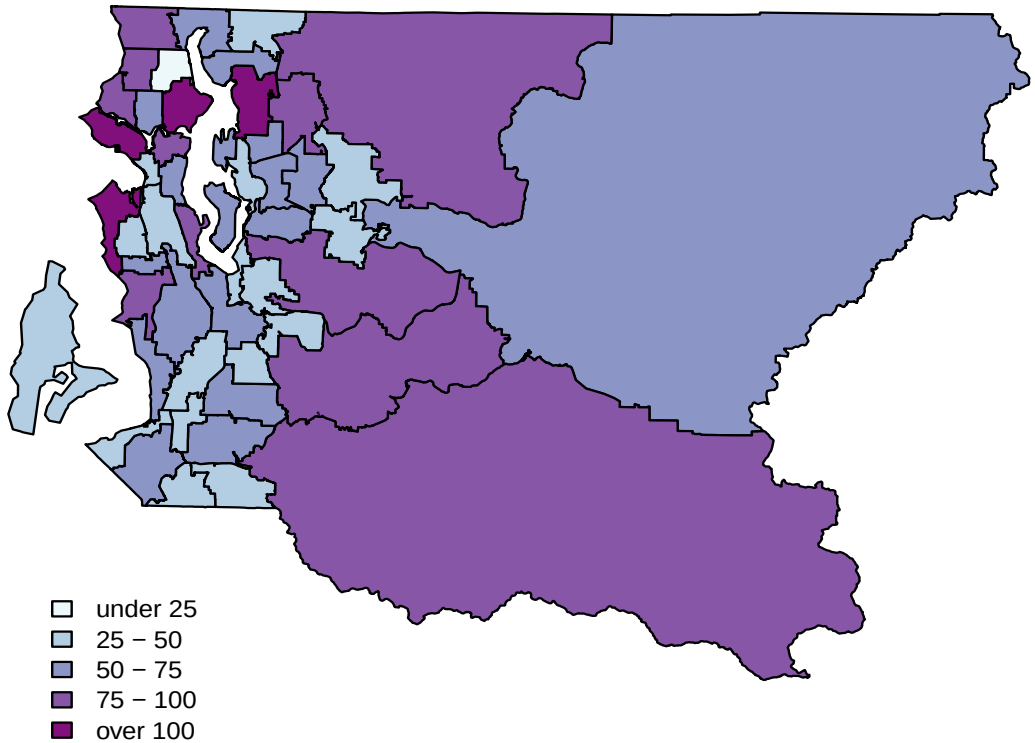

Figure 5 : Sample sizes across 48 HRAs in 2011.

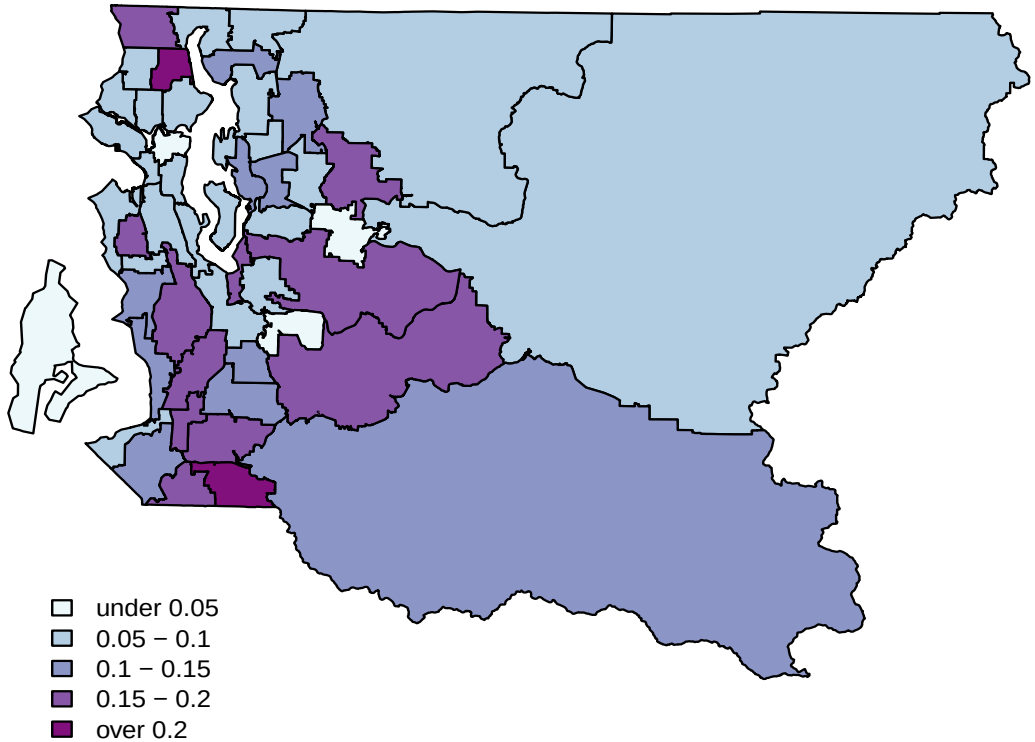

Figure 6 : Diabetes prevalence by HRAs in 2011: crude proportions.

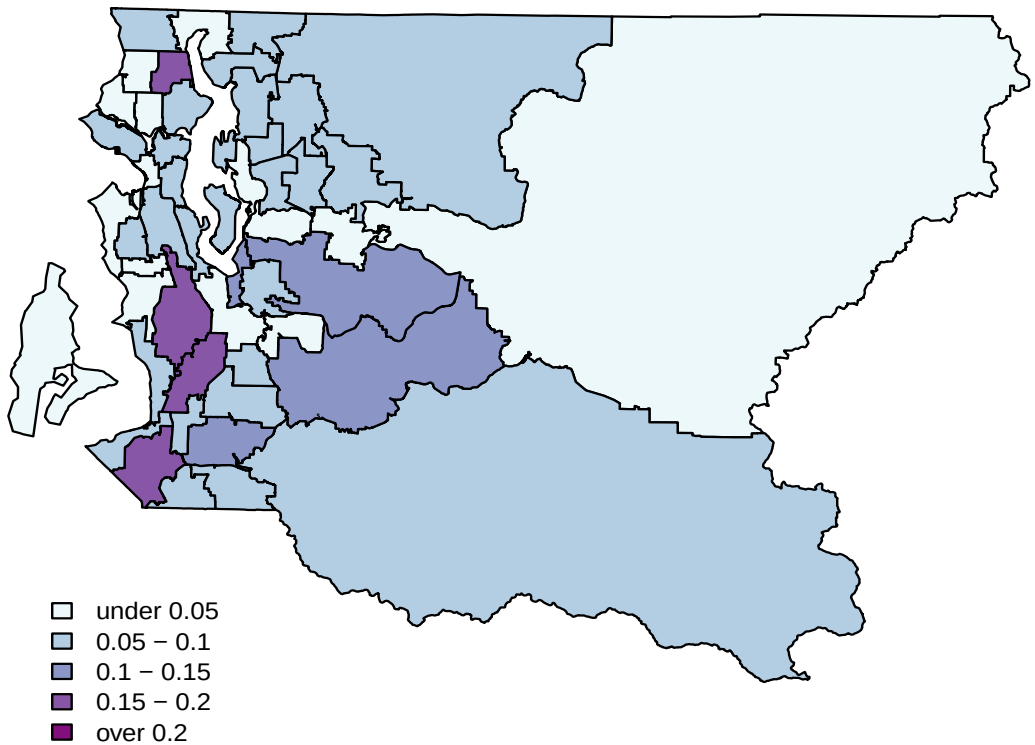

Figure 7 : Diabetes prevalence by HRAs in 2011: Horvitz-Thompson weighted estimator.

Another Example of a Survey: NHANES

The National Health and Nutrition Examination Survey (NHANES) is designed to assess the health and nutritional status of adults and children in the United States.

The sampling uses multistage, probability sampling.

We take details from `http://www.cdc.gov/nchs/tutorials/nhanes/surveydesign/SampleDesign/intro.htm`.

Another Example of a Survey: NHANES

Stage 1: Primary sampling units (PSUs) are selected. These are mostly single counties or, in a few cases, groups of contiguous counties. Sampling carried out with probability proportional to a measure of size (PPS).

Stage 2: The PSUs are divided up into segments (generally city blocks or their equivalent). As with each PSU, sample segments are selected with PPS.

Stage 3: Households within each segment are listed, and a sample is randomly drawn. In geographic areas where the proportion of age, ethnic, or income groups selected for oversampling is high, the probability of selection for those groups is greater than in other areas.

Stage 4: Individuals are chosen to participate in NHANES from a list of all persons residing in selected households. Individuals are drawn at random within designated age-sex-race/ethnicity screening subdomains. On average, 1.6 persons are selected per household.

Approaches to Inference: Overview

If the data are not a random sample, this must be accounted for in the analysis.

We have already seen an example of non-random sampling: case-control studies!

A key distinction is between design- and model-based approaches to inference.

To add to the confusion there is also model-assisted inference.

Standard hierarchical model-based approaches to analysis ignore the sampling mechanism and do not adjust for non-response and are subject to biases, which may be large.

A notable exception is the model due to Fay and Herriot (1979).

Design-Based Inference

Examples of non-SRS designs include:

- Stratified Random Sampling: The population is categorized into strata and simple random sampling is carried out within each strata, with sample sizes set by the sampler.
- Single Stage Cluster Sampling: Partition the total population into "clusters". Randomly sample clusters and then obtain information from all individuals within clusters.
- Two-Stage Cluster Sampling: Partition the total population into "clusters". Randomly sample clusters and then sample individuals within each of the sampled clusters.
- Multistage Sampling: More stages, for example sample clusters within strata.

Design-Based Inference

So in this section we will start by considering a single area only.

Suppose the population is of size N with individual responses $Y_1, \ldots, Y_N$.

A survey of size m is carried out.

We collect measurements y_k along with sampling weights w_k, $k \in S$, where S denotes the random set of indices that are selected, with $m = |S|$, i.e. the number of individuals sampled.

The sampling weight w_k can be thought of as the number of people that response k represents.

The sampling weights are defined as the inverse of the inclusion probabilities $\pi_k = \Pr(\text{ sampling individual } k\ )$, i.e. $w_k = \pi_k^{-1}$.

Design-Based Inference

The mean of the response over the total population is $P = \frac{1}{N}\sum_{k=1}^{N} Y_k$.

Suppose further we wish to estimate the mean θ based on y_k, $k \in S$.

The Horovitz-Thompson (HT) estimator of P is

$$\widehat{P}^{\text{HT}} = \frac{\sum_{k\in S} w_k y_k}{N}.$$

For a SRS, $\pi_k = m/N$ and so $w_k = N/m$. Therefore,

$$\widehat{P}^{\text{HT}} = \frac{\sum_{k\in S} \frac{N}{m} y_k}{N} = \frac{\sum_{k\in S} y_k}{m},$$

the sample mean.

Horvtiz-Thompson Estimator

In design-based inference the data Y_k are viewed as fixed (non-random), it is the binary indicators of inclusion in the sample that are viewed as random.

Let I_k be a 0-1 indicator of whether person k was sampled, with

$$I_k|\pi_k \sim_{ind} \text{Bernoulli}(\pi_k),$$

$k = 1, \ldots, N$.

The HT estimator is an unbiased estimator as we now show.

Horvtiz-Thompson Estimator

Recall, we have fixed constants $Y_1, \ldots, Y_N$ and random variables $I_1, \ldots, I_N$. Hence,

$$\begin{aligned}
\mathrm{E}\left[\frac{\sum_{k \in S} w_k y_k}{N}\right] &= \mathrm{E}\left[\frac{\sum_{k=1}^{N} w_k I_k Y_k}{N}\right] \\
&= \frac{\sum_{k=1}^{N} w_k \mathrm{E}[I_k] Y_k}{N} \\
&= \frac{\sum_{k=1}^{N} \pi_k^{-1} \pi_k Y_k}{N} \\
&= \frac{\sum_{k=1}^{N} Y_k}{N} = P
\end{aligned}$$

Horvtiz-Thompson Estimator

It can also be shown that the variance of $\widehat{P}^{\text{HT}}$ is

$$\text{var}\left(\widehat{P}^{\text{HT}}\right) = \frac{1}{N}\left[\sum_{j=1}^{N} Y_j^2 \frac{(1-\pi_j)}{\pi_j} + 2\sum_{j=1}^{N}\sum_{k>j} Y_j Y_k \frac{(\pi_{jk}-\pi_j\pi_k)}{\pi_j\pi_k}\right] \quad (1)$$

where π_{jk} is the joint inclusion probability of samples j and k.

The estimator of this variance is

$$\widehat{\text{var}}\left(\widehat{P}^{\text{HT}}\right) = \frac{1}{N}\left[\sum_{j\in S} y_j^2 \frac{(1-\pi_j)}{\pi_j} + 2\sum_{j\in S}\sum_{k>j} y_j y_k \frac{(\pi_{jk}-\pi_j\pi_k)}{\pi_j\pi_k}\right] \quad (2)$$

These estimators are known as design-based variance estimators.

An equivalent form for the variance is

$$\widehat{\text{var}}\left(\widehat{P}^{\text{HT}}\right) = \frac{1}{N}\sum_{j\in S}\sum_{k\in S} \frac{y_j y_k}{\pi_{jk}} - \frac{y_j}{\pi_j}\frac{y_k}{\pi_k} \quad (3)$$

The estimator is also asymptotically normal, which allows confidence intervals to be derived.

Horvtiz-Thompson Estimator

For the case of SRS, the variance formula (1) simplies to

$$\widehat{\text{var}}\left(\widehat{P}^{\text{HT}}\right) = \left(1 - \frac{m}{N}\right)\frac{S^2}{m}$$

where

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j - \overline{Y})^2$$

and (2) replaces S^2 by

$$s^2 = \frac{1}{m-1}\sum_{j\in S}(y_j - \overline{y})^2.$$

Sanity checks: If the population is large then $m/N \approx 0$ and we obtain the usual formula. And if $m = N$ we have no uncertainty, since we know the answer.

Weighting

We return to our notation with $i = 1, \ldots, n$ indexing areas.

We have N_i individuals in area i and the indices of those selected in a sample of size m_i is denoted S_i.

The weights are often formed via

$$w_{ik} = w_{ik}^d \times w_{ik}^p \tag{4}$$

where w_{ik}^d is the design weight and w_{ik}^p is the post-stratification weight.

For the design weights

$$w_{ik}^d = \frac{1}{\pi_{ik}}$$

where π_{ik} is the probability of selection.

Weighting

If N_i is not known it may be estimated by

$$\widehat{N}_i = \sum_{k \in S_i} w_{ik}^d$$

is an estimate of the total population in area i, in line with interpreting w_{ik}^d as the number of individuals that this individual represents.

Note that,

$$E[\widehat{N}_i] = \sum_{k=1}^{N_i} E[I_{ik}] \pi_{ik}^{-1} = N_i,$$

so that this estimator is unbiased.

Post-stratification, as the name suggests, adjusts the weights after sampling, so that population totals in a set of stratum (e.g., age/gender) are recovered.

Post-stratification and Raking

If the post-stratification groups are indexed by j the weights are

$$w_{ik}^{p} = \frac{N_{j(k)}}{\widehat{N}_{j(k)}}$$

where $j(k)$ indicates the group to which individual k belongs, N_j are the known totals and $N_{j(k)} = \sum_{k \in S_j} w_{ik}^{d}$. This procedure adjusts the weights so that the known totals are recovered.

Previously in BRFSS in King County, post-stratification was used based only on age and gender.

Raking now used for BRFSS, adjusting for more factors: age, gender, race/ethnicity, marital status, education, owner/renter status, and cell phone/landline status).

Cannot exactly match all cross-classified tables of counts, so instead lower dimensional margins are controlled using a procedure known as iterative proportional fitting.

SAE under SRS

As discussed, the data upon which SAE is based are often gathered via complex designs but we begin with simple random sampling (SRS).

Let Y_i represent the total number in area i with the characteristic of event, and N_i the population size; Y_i is unobserved but assume N_i is known.

Interest focusses on the proportion $P_i = Y_i/N_i$ or the total Y_i.

Note the notation here: P_i is the unobserved proportion and not the proportion in an infinite population from which N_i are drawn.

A SRS is taken in area i of size m_i, of which y_i have the characteristic.

The obvious estimators are

$$\widehat{P}_i = \frac{y_i}{m_i} \tag{5}$$

$$\widehat{Y}_i = N_i \times \frac{y_i}{m_i} \tag{6}$$

SAE under SRS

A serious problem with SAE is that areas with small m_i will have large sampling variability.

To overcome this problem hierarchical modeling is used.

This approach performs global and/or local smoothing, which introduces bias in the estimator, but the variance is reduced, so that usually the mean squared error (MSE)[1] of the estimator is reduced compared to the original estimator.

The hierarchical models we have previously examined can be used, with some post-processing to obtain inference on the quantities of interest.

[1] MSE=Bias^2 + Variance

SAE under SRS

A starting model is,

$$\begin{aligned} y_i|P_i &\sim \text{Binomial}(m_i, P_i) \\ \text{logit } P_i &= \beta_0 + \epsilon_i + S_i \end{aligned}$$

with unstructured residual log odds $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and spatial residual log odds $S_i \sim \text{ICAR}(\sigma_s^2)$.

Finally we specify (hyper-)priors for β_0, σ_ϵ^2 and σ_s^2.

Inference for Y_i is straightforward since $Y_i = y_i + (N_i - m_i) \times P_i$ under the binomial approximation.

So if we have a posterior for P_i we can easily convert into a posterior for Y_i.

Model-Based Inference for Stratified Random Sampling

Suppose we sample on the basis of a discrete (design) variable $X \in (X_1, \ldots, X_J)$, for example, age, gender, race, area,...

Stratified random sampling provides one example: we sample m_{ij} people within area i and stratum j.

Simple example:

- Suppose that gender is the design variable and
- we sample twice as many women as men, and that the outcome of interest is diabetes,
- women have lower risk of diabetes than men.
- if we take the simple proportion (ignoring gender) then our area-level estimate will be downwardly biased.

Model-Based Inference for Stratified Random Sampling

Suppose the stratum membership of all sampled individuals is known, and z_{ij} are the number of responders from a sample m_{ij} in stratum j.

Could set up as a product of hypergeometric distributions.

More simply, we can model as:

$$\begin{aligned} z_{ij}|p_i(X_j) &\sim \text{Binomial}(m_{ij}, p_i(X_j)) \\ \log\left[\frac{p_i(X_j)}{1-p_i(X_j)}\right] &= \beta_j + \epsilon_i + S_i \end{aligned}$$

with (say) $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $S_i \sim \text{ICAR}(\sigma_s^2)$.

Hyperpriors for β_j, σ_ϵ^2 and σ_s^2.

Model-Based Inference

Then prediction is

$$\widehat{Y}_i = \sum_{j=1}^{J} N_{ij} \widehat{p}_i(X_j)$$

where N_{ij} is the size of the population in group j and area i; this is the post-processing step.

Problems:

- X's are not routinely known for sampled individuals in public-use databases.
- N_{ij} are also not routinely known.

For example, in BRFSS, we would need to know which telephone list each individual came from, and the population numbers in each area from each telephone list.

Alternative: Design-based inference (the appendix contains details on direct and indirect approaches to SAE).

Model-based approaches

Indirect methods provide a link, often with an implicit model, between different areas; we now turn to explicit models that provide such a link.

The models we describe are hierarchical (or mixed-effects) models which aim to accurately describe between area differences.

Such models offer several advantages:

- Models can be tuned to the application, building on the existing theory and practical experience of mixed models, including non-linear models, such as logistic mixed models.
- Area-specific measures of uncertainty are produced.
- Estimates for areas with no data can be formed.
- One can attempt to check assumptions using diagnostics.
- A variety of area-specific random effects, including spatial, are available.

Model-based approaches

The use of explicit models has not been carried out greatly in survey sampling, where design-based inference is historically the norm.

The design consistency of model-based estimators is therefore of interest.

Disadvantages of mixed models:

- How to incorporate the design weights/acknowledge the design?
- It is often difficult to check modeling assumptions.
- Computation can be demanding, though this is improving.

Models can be specified at the level of the area or the unit

Hierarchical modeling: notes

Inference may be carried out via likelihood or Bayes, with the latter placing priors on $\boldsymbol{\beta}, \sigma^2_\varepsilon, \sigma^2_\epsilon$.

If a likelihood approach is taken, the random effect estimates $\widehat{\varepsilon_i}$, are obtained as best linear unbiased predictors (BLUPs).

If there are no data in particular areas we can still make predictions, if we assume the model holds for all areas.

Note: can add area level covariates to model.

Hierarchical Modeling of Survey Sample Data

A Horvitz-Thompson weighted estimator of the log-likelihood for binary data is

$$\sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ik}\left\{y_{ik}\log P_i + (1-y_{ik})\log(1-P_i)\right\} \tag{7}$$

(Binder, 1983) where y_{ik} is the binary outcome on person k in area i, with associated weight w_{ik}.

Method known as pseudo-likelihood.

Pseudo-likelihood (Skinner, 1989; Pfeffermann *et al.*, 1998) has been used within a hierarchical modeling framework with the scaling of the weights being a major issue (Potthoff *et al.*, 1992; Longford, 1996; Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006).

Congdon and Lloyd (2010) use a weighted likelihood to analyze BRFSS data and introduce residual spatial random effects at the state level.

Further References

Although there is a huge literature on small area estimation the spatial smoothing of survey data with complex weights is not routinely carried out.

In terms of spatial smoothing techniques, a number of authors allow for spatial correlation between areas, see for example Singh *et al.* (2005), Pratesi and Salvati (2008) and Pereira and Coelho (2010).

These models are subject to bias, however, since they do not adjust for the sampling scheme.

We shortly describe a relatively new approach based on the concept of "effective sample size" and "effective number of cases".

A related Bayesian model has recently been suggested by Ghitza and Gelman (2013), while a quite different approach, based on a penalized spline model, is described in Zheng and Little (2003) and Zheng and Little (2005).

Adjustment for non-response using hierarchical models

Suppose we have a binary response of interest, Y, and we wish to estimate (predict) the proportion and total in a region of interest, perhaps by domain (e.g. smaller geographical areas).

A model-based approach would assume

$$Y_j(\mathbf{x})|p_j(\mathbf{x}) \sim \text{Binomial}(n_j(\mathbf{x}), p_j(\mathbf{x})),$$

where:

- $\mathbf{x}$ is a vector of dummy variables that defines which group the unit lies within; typical strata include age, gender, race, ethnicity, geographical area; assume J groups in total.
- $n_j(\mathbf{x})$ is the number in the sample, with characteristics $\mathbf{x}$.
- $Y_j(\mathbf{x})$ is the number with $Y = 1$ with characteristics $\mathbf{x}$.
- $p_j(\mathbf{x}) = \Pr(Y = 1|\mathbf{x},\ \text{group } j)$.

Adjustment for non-response using hierarchical models

Hierarchical model:

$$\begin{aligned} p_j(\mathbf{x}_k) &= \text{expit}\left(\alpha_{i[k]} + \sum_{j=1}^{J} \beta_j \times 1(\text{group of unit } k = j)\right) \\ \alpha_{i[k]} &\sim \text{N}(0, \sigma_\alpha^2) \end{aligned}$$

where $\alpha_{i[k]}$ is the random effect associated with area i, within which unit k resides.

Once we have estimated the parameters of the model we can estimate the total in area i via

$$T_i = \sum_{j=1}^{J} N_{ij} p_j(\mathbf{x})$$

where N_{ij} are the known totals in group j and area i.

Adjustment for non-response using hierarchical models

Under this approach, which does not explicitly use the weights, it is assumed that the design variables are included in $\mathbf{x}$.

With many groups (i.e. large J), estimation is unstable and so hierarchical models are used (Gelman, 2007).

We describe the example in Gelman and Hill (2007, Chapter 14) which concerns combining pre-election opinion polls.

The outcome we consider is the probability that a survey respondent prefers the Republican candidate for president, using a set of seven CBS News polls conducted during the week before the 1988 presidential election.

State-level opinion polls

We first fit a simple model, to understand the parameters.

The model is fitted in Stan a computing environment that includes various possibilities for fitting, including an MCMC algorithm based on Hamiltonian Monte Carlo (the No U-Turns Sampler).

Model 1:

$$\begin{aligned} \Pr(y_i = 1|\mathbf{x}_i) &= \text{expit}(\alpha_{j[i]} + b_1 \times 1(\text{ black})_i + b_2 \times 1(\text{ female })_i) \\ \alpha_j|\sigma^2_{\text{state}} &\sim \text{N}(\mu_\alpha, \sigma^2_{\text{state}}), \end{aligned}$$

for $j = 1, \ldots, 51$ states and $i = 1, \ldots, 13,544$ survey responses and with $\mathbf{x}_i$ containing the information on the state, race and gender of individual i.

Notes:

- α_j is a state-level random effect and $\alpha_{j[i]}$ picks up the state for individual i,
- $1(\text{ black})_i$ and $1(\text{ female })_i$ are indicators for whether individual i is black or female, respectively.

State-level opinion polls

Parameter interpretation: note that

$$\frac{\Pr(Y_i = 1|\mathbf{x}_i)}{\Pr(Y = 0|\mathbf{x}_i)} = \exp(\alpha_{j[i]} + b_1 \times 1(\text{ black})_i + b_2 \times 1(\text{ female })_i).$$

Hence, for example,

$$\exp(b_2) = \frac{\frac{\Pr(Y_i|1(\text{ female })_i=1,\ \text{state }_{j[i]})}{\Pr(Y_i|1(\text{ female })_i=1,\ \text{state }_{j[i]})}}{\frac{\Pr(Y_i|1(\text{ male })_i=0,\ \text{state }_{j[i]})}{\Pr(Y_i|1(\text{ male })_i=0,\ \text{state }_{j[i]})}}$$

is the odds of being Republican for a female, as compared to a male of the same race and in the same state.

From the output below, the posterior mean is $\exp(-0.09) = 0.91$, so that females have a 9% reduction in the odds.

State-level opinion polls

Edited stan code (`election88.stan`):

```
transformed parameters {
  vector[N] y_hat;

  for (i in 1:N)
    y_hat[i] <- b[1] * black[i] + b[2] * female[i] + a[state[i]];
}
model {
  mu_a ~ normal(0, 1);
  a ~ normal (mu_a, sigma_a);
  b ~ normal (0, 100);
  y ~ bernoulli_logit(y_hat);
}
```

Summary:

```
        mean  se_mean   sd     2.5%     25%     50%     75%     97.5%  n_eff Rhat
b[1]   -1.76    0.01 0.21    -2.17    -1.89    -1.75    -1.61    -1.36  1440  1.00
b[2]   -0.09    0.00 0.10    -0.29    -0.16    -0.09    -0.03     0.09   640  1.00
mu_a    0.44    0.00 0.11     0.24     0.37     0.44     0.52     0.64   556  1.00
sigma_a 0.44    0.01 0.09     0.28     0.38     0.44     0.50     0.63   189  1.01
```

State-level opinion polls

Now for a more realistic model, which includes the demographics used by CBS in the survey weighting: age, sex, race and education.

Random effects are used for some of these, as data in some cells are small.

Model 2:

$$
\begin{aligned}
\Pr(y_i = 1|\mathbf{x}_i) &= \text{expit}(\beta_1 + d_{state[i]} + \beta_2 1(\text{ black})_i + \beta_3 1(\text{ female })_i \\
&+ \beta_4 1(\text{ black.female })_i + \beta_5 \text{vprev}_i + a_{age[i]} \\
&+ b_{edu[i]} + c_{age.edu[i]} + d_{state[i]} + e_{region[i]}) \qquad (8) \\
a_{age}|\sigma_a^2 &\sim \text{N}(0, \sigma_a^2) \\
b_{edu}|\sigma_b^2 &\sim \text{N}(0, \sigma_b^2) \\
c_{age.edu}|\sigma_c^2 &\sim \text{N}(0, \sigma_c^2) \\
d_{state}|\sigma_d^2 &\sim \text{N}(0, \sigma_d^2) \\
e_{region}|\sigma_e^2 &\sim \text{N}(0, \sigma_e^2).
\end{aligned}
$$

State-level opinion polls

Terms:

- vprev_j is a measure of the previous Republican vote in the state.
- *region* is a factor with 5 levels.
- We include gender and race, and the interaction of the two, as fixed effects.
- There are random effects for age (4 levels), education (4 levels), and state (51 levels).

For state j the estimated Republican vote is

$$\theta_j = \frac{\sum_{l \in j} N_l p_{jl}}{\sum_{l \in j} N_l},$$

where each summation is over the $4 \times 4 \times 4$ categories, and p_{jl} is the estimated probability of Republican in category l and state j.

Figure 8 plots

$$\Pr(y_i = 1 | \mathbf{x}_i) = \text{expit}(\text{linpred}_i + d_{state[i]}),$$

versus linpred_i as in (8).

State-level opinion polls

Stan code (`election88_full.stan`):

```
transformed parameters {
  vector[N] y_hat;

  for (i in 1:N)
    y_hat[i] <- beta[1] + beta[2] * black[i] + beta[3] * female[i]
                + beta[5] * female[i] * black[i]
                + beta[4] * v_prev_full[i] + a[age[i]] + b[edu[i]]
                + c[age_edu[i]] + d[state[i]] + e[region_full[i]];
}
model {
  a ~ normal (0, sigma_a);
  b ~ normal (0, sigma_b);
  c ~ normal (0, sigma_c);
  d ~ normal (0, sigma_d);
  e ~ normal (0, sigma_e);
  beta ~ normal(0, 100);
  y ~ bernoulli_logit(y_hat);
}
```

Summary:

```
         mean  se_mean    sd    2.5%    25%    50%      75%    97.5%   n_eff  Rhat
beta[1]  0.86    0.07  1.79    -2.77    -0.30     0.85     1.96     4.46   613 1.00
beta[2] -1.68    0.01  0.34    -2.34    -1.90    -1.67    -1.44    -1.04  1346 1.00
beta[3] -0.09    0.00  0.10    -0.28    -0.16    -0.09    -0.03     0.10  2000 1.00
beta[4] -0.59    0.12  3.02    -6.80    -2.49    -0.58     1.32     5.68   686 1.00
beta[5] -0.17    0.01  0.42    -0.99    -0.45    -0.17     0.12     0.63  1121 1.00
sigma_a  0.24    0.02  0.34     0.01     0.08     0.14     0.28     1.02   285 1.01
sigma_b  0.32    0.02  0.36     0.02     0.13     0.23     0.39     1.12   363 1.01
sigma_c  0.15    0.01  0.10     0.02     0.08     0.14     0.21     0.37    96 1.02
sigma_d  0.46    0.00  0.10     0.27     0.39     0.45     0.52     0.67   419 1.01
sigma_e  0.68    0.10  1.17     0.02     0.13     0.30     0.65     4.18   146 1.04
```

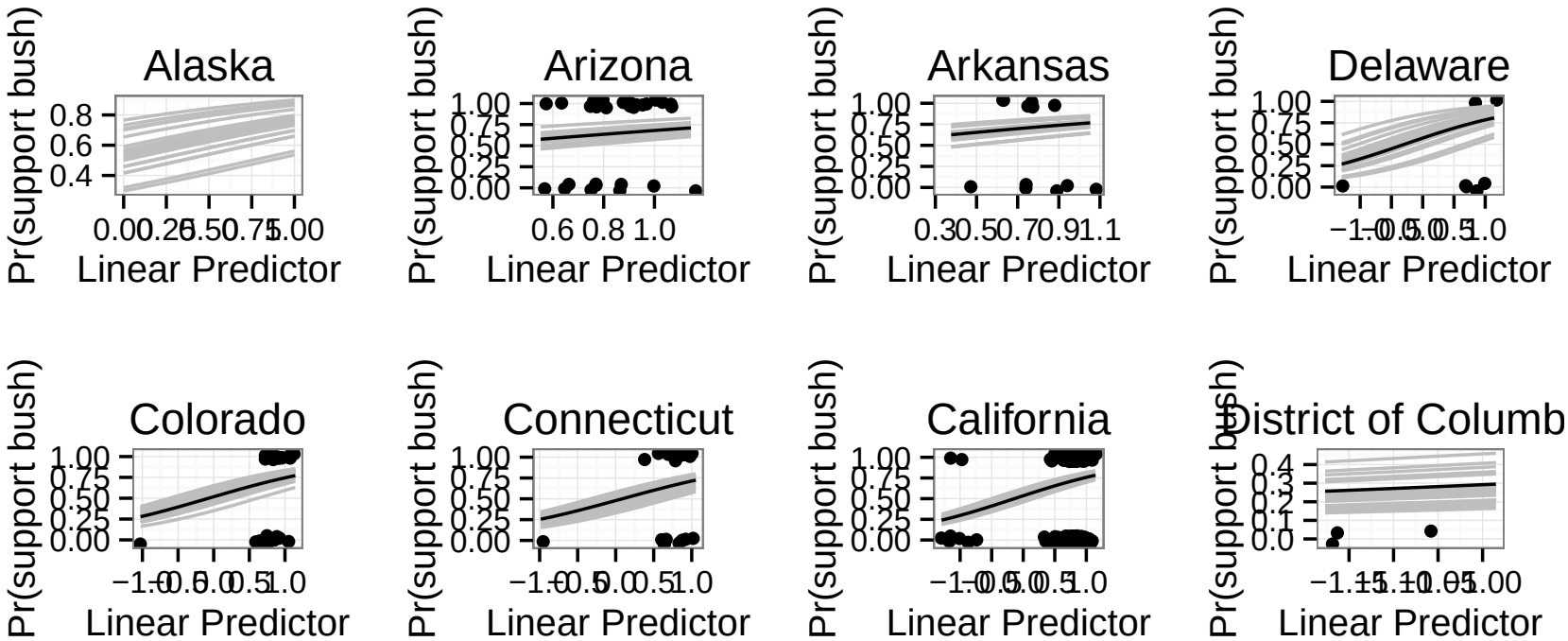

Figure 8 : Estimated probability of a survey respondent supporting Bush for president as a function of the linear predictor for the demographics, for 8 states. Dots show the respondents, the solid line the median and the light lines 20 draws from the posterior.

BRFSS Example

We analyze the HRA King County diabetes data using different models:

- Direct estimator[2] (no smoothing).
- A beta-binomial model with no adjustment for the design.
- Take as data the logit of the HT estimator:

$$y_i = \log\left[\frac{\widehat{P}_i^{\mathrm{HT}}}{1-\widehat{P}_i^{\mathrm{HT}}}\right] | P_i \sim \mathrm{N}\left(\log\left[\frac{P_i}{1-P_i}\right], \frac{\widehat{\mathrm{var}}(\widehat{P}_i^{\mathrm{HT}})}{(\widehat{P}^{\mathrm{HT}})_i^2(1-(\widehat{P}_i^{\mathrm{HT}}))^2}\right).$$

- That is, the first stage of the hierarchical model is $y_i|\eta_i \sim_{iid} \mathrm{N}(\eta, \widehat{V}_{\mathrm{DES},i})$, where $\widehat{V}_{\mathrm{DES},i}$ is the known design-based variance.
- Add independent area-specific normal random effects to the linear predictor.
- Logit normal with independent area-specific normal random effects and spatial (ICAR) random effects to the linear predictor, i.e., $\eta_i = \beta_0 + \epsilon_i + S_i$.
- The spatial model took less than one second to run using INLA.

[2] A direct estimator is based on data from the area *only*

BRFSS Example

Figure 9 shows the estimates under the different models.

The large variability in the non-hierarchical estimates is evident, and unrealistic.

The beta (empirical Bayes) and non-spatial normal models give very similar answers.

The weighting (adjustment) makes gives a small shift, but the variance reduction is the dominant factor for these (sparse) data.

Figure 10 shows the spatially smoothed estimates: large counts around Puget Sound.

Figure 11 shows the posterior standard deviation of the counts. Large uncertainty where the counts are high.

Weighting does not make a great deal of difference here, but spatial smoothing does.

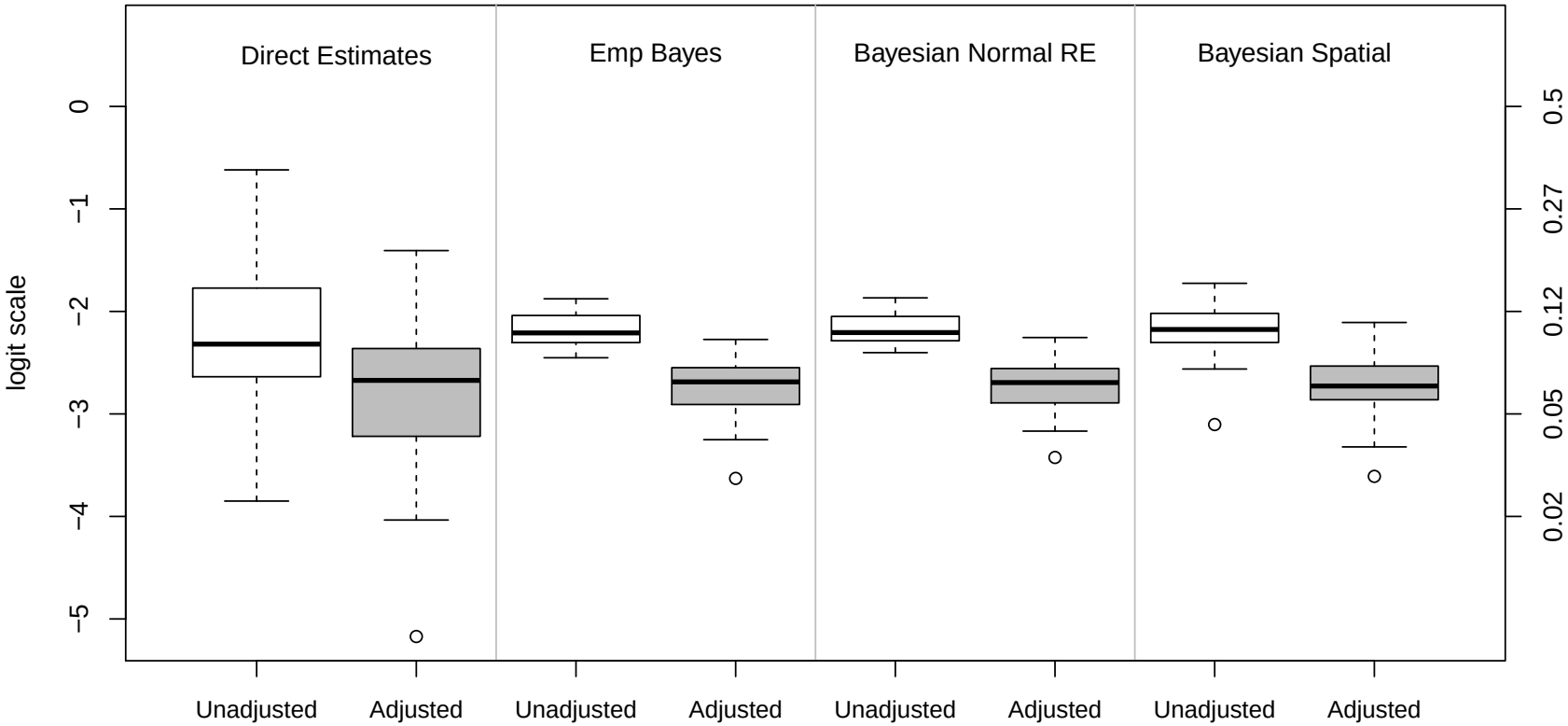

Figure 9 : Estimated diabetes prevalence by HRA: the left axis is on the logit scale and the right is on the [0,1] scale.

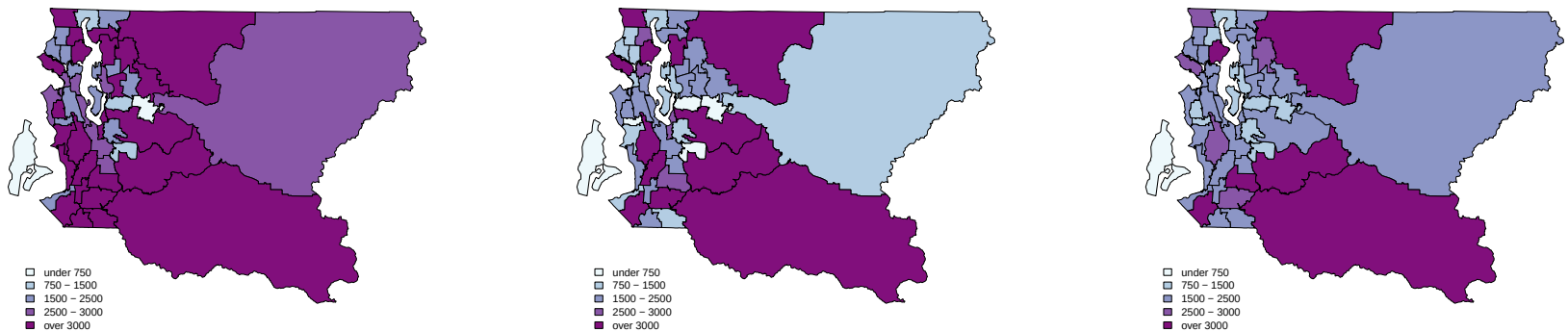

Figure 10 : Estimates of the total diabetes counts by HRA in King County under binomial, Horvitz-Thompson and weighted spatial models.

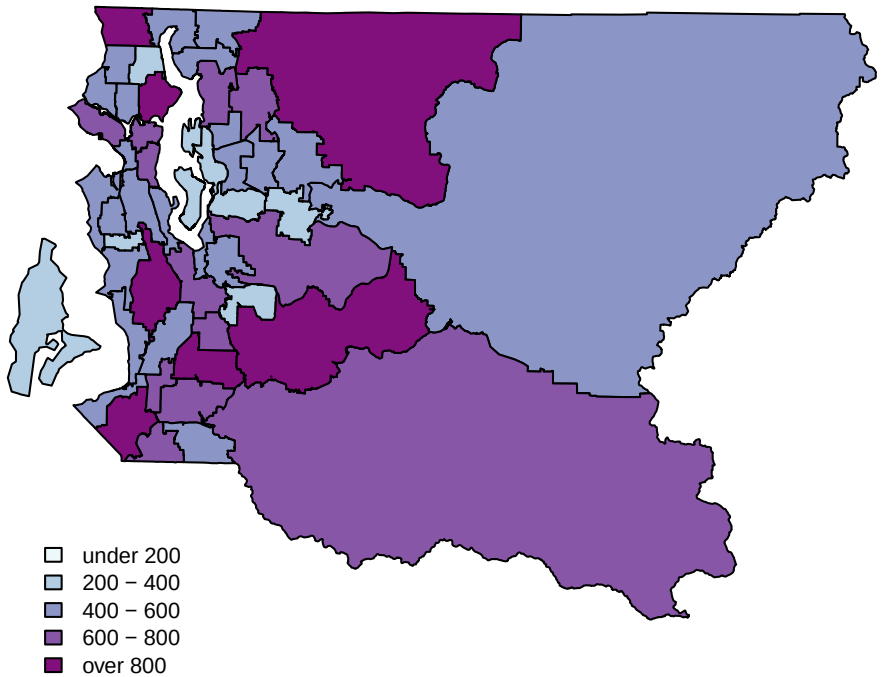

Figure 11 : Estimated uncertainty of the adjusted estimates of the total diabetes counts by HRA in King County under the spatial model.

Model Validation

To validate the model we examine a "large" area and take only a small sample of 20 individuals.

In particular, we treat the direct estimates for West Seattle in 2011 (with $m_i = 117$ samples) as the "gold standard".

Then repeatedly sample 20 individuals without replacement from this area and carry out prediction using different models.

The mean squared error of the estimates is then evaluated, by comparing the estimates with the gold standard.

Table 2 : Validation results for West Seattle.

	Direct		Bayesian Indept		Bayesian Spatial	
	Unadjust	Adjust	Unadjust	Adjust	Unadjust	Adjust
Bias^2	0.70	0.00	3.73	0.38	4.35	0.25
Variance	2.49	1.16	0.04	0.06	0.06	0.06
MSE	3.19	1.16	3.76	0.44	4.40	0.31

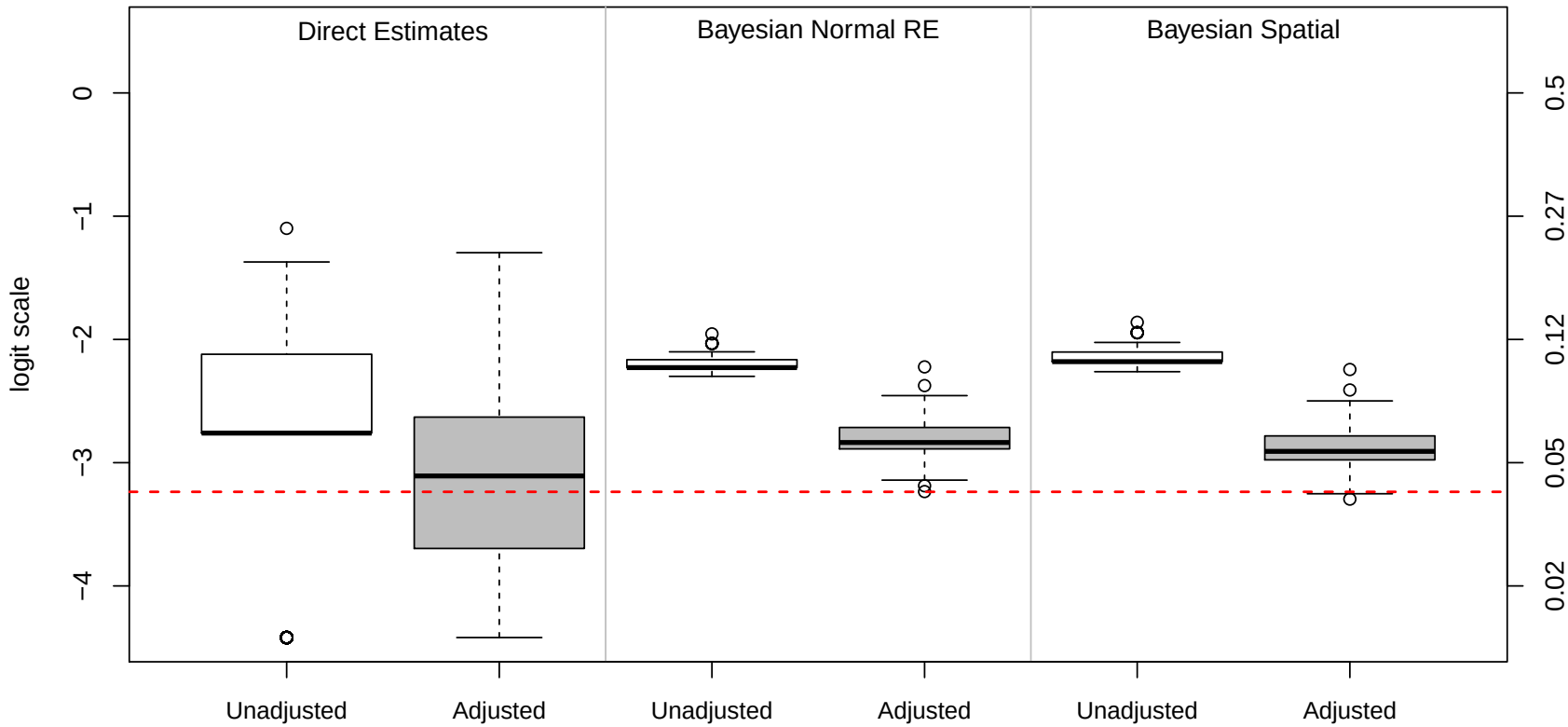

Figure 12 : Model validation: estimates of diabetes prevalence across simulations for West Seattle. The red line denotes the "truth".

Motivation for Tanzania study (Mercer *et al.*, 2016)

MDG4: Target 4.A: Reduce by two-thirds, between 1990 and 2015, the under-five mortality rate (U5MR).

How can we determine the U5MR? Vital registration is the ideal, but in many places this does not exist.

As an alternative, relevant information may be gathered in surveys, which often have complex designs.

National U5MR rates can obscure important subnational variability; highlighting areas with relatively high U5MR allows prioritizing of resources.

Here: analyze survey data on child mortality from 21 regions of Tanzania over 1980–2010.

Modeling: Use totality of data to smooth in space and time, to alleviate low power associated with direct estimates.

Available data

We focus on child mortality using data from:

- Five Tanzanian Demographic and Health Surveys (DHS).
- One Tanzania HIV and Malaria Indicator Survey (HMIS).
- Two health and demographic surveillance system (HDSS) sites in Ifakara and Rufiji.

Over the period 1980–2010 estimates of child mortality from the two types of data sources (surveys, surveillance sites) are generally similar but different in useful ways.

The HDSS estimates are accurate (low bias) and precise (small variance) measurements for comparatively small, geographically-defined populations, and the DHS/HMIS estimates are less accurate and much less precise but representative of large populations.

Demographic Household Surveys (DHS)

Full DHS surveys that collected data necessary for child mortality estimates were conducted in Tanzania in 2010, 2004–05, 1999, 1996, and 1991–92, in addition to the HMIS that included child mortality conducted in 2007–08.

The 2010 DHS, 2007–08 HMIS and 2004–05 DHS surveys used 2-stage cluster samples:

1. Clusters were sampled from enumeration areas (EAs) from the 2002 Tanzania census.
2. Sampling of households within each cluster was carried out.

First stage: 150–375 EAs sampled from ≈50,000 EAs.

Demographic Household Surveys (DHS)

The 1999 DHS, 1991–92 DHS and 1996 DHS used 3-stage cluster design:

1. Selecting wards and branches using the 1988 Tanzania Census as a sampling frame.
2. Using probability proportional to size sampling to select EAs from each selected ward or branch, and
3. Selecting households from a new list of all households in each selected EA.

Stratification by urban/rural and region was done at the first stage, with oversampling of Dar es Salaam, other urban areas, and Zanzibar (which is excluded in the results presented here, because of its non-comparability and data issues).

We focus on the 21 mainland regions of Tanzania (up to 2010).

Demographic Household Surveys (DHS)

All women age 15 to 49 who slept in the household the night before were interviewed in each selected household and response rates were high (above 95% for households in all surveys).

DHS provides sampling (design) weights, assigned to each individual in the dataset.

Limited information is provided for each survey concerning the calculation of survey weights, but the general explanation indicates that raw survey weights are the inverse of the product of the 2–3 probabilities of selection from each stage.

These raw weights were then adjusted to reflect household response and individual response rates.

Since the design variables are not available on the population, and the final weighting steps are mysterious, it is not straightforward to use a conventional model-based approach.

Health and Demographic Surveillance System (HDSS)

The Ifakara Health Institute (IHI), Tanzania runs a number of health and population research projects including two HDSS sites: Ifakara and Rufiji.

The HDSS data are generated through repeated household visits.

For the data we use, each household was visited three times per year at regular intervals.

During each visit a 'household roster' was updated and all new vital and migration events for all members of the household were recorded.

Because both HDSS sites are long-lived surveillance projects, there are many repeated observations on households and individuals, and all of these must be linked in order to conduct survival analysis.

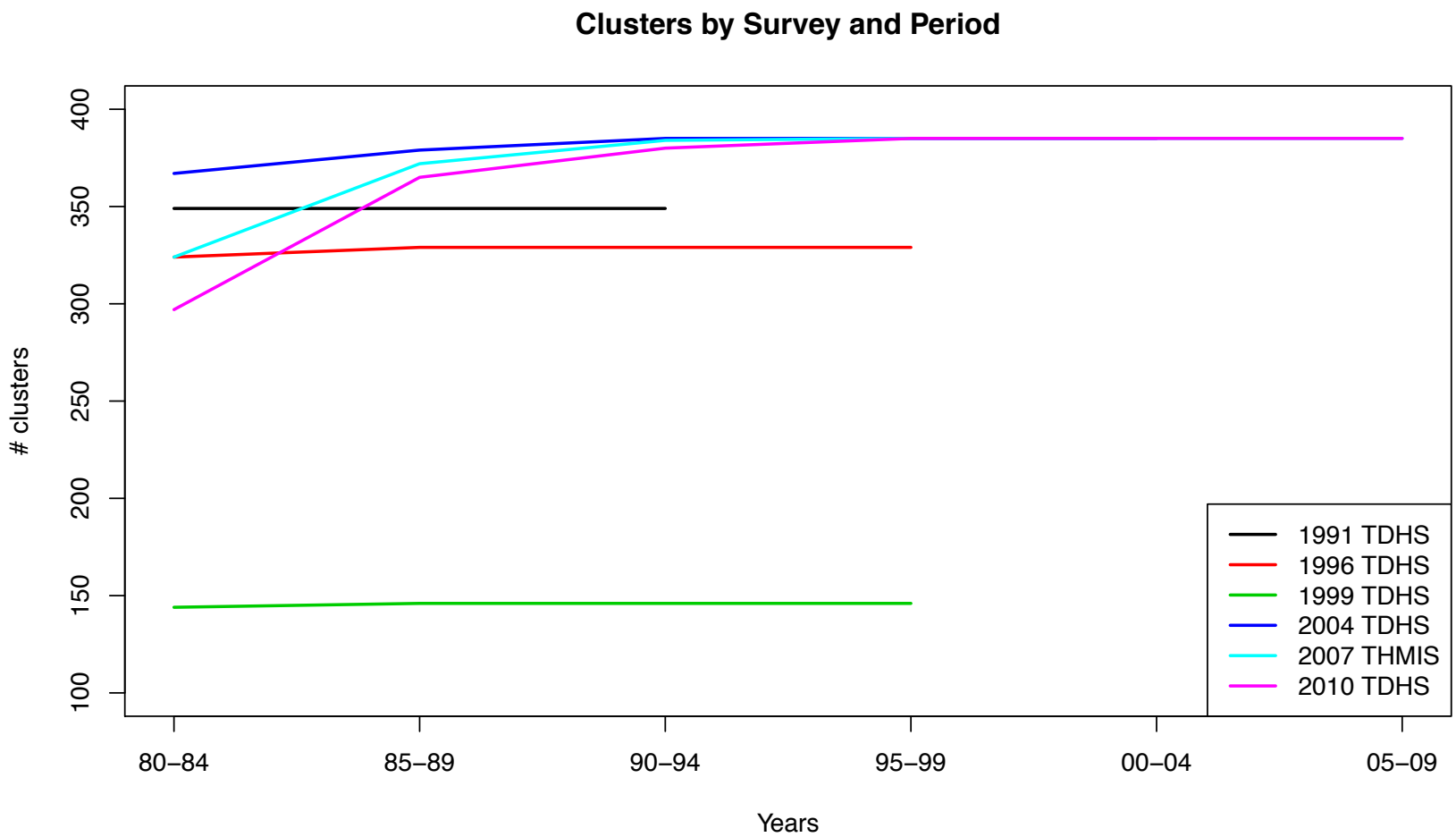

Figure 13 : Numbers of selected clusters (enumeration areas) by survey and period.

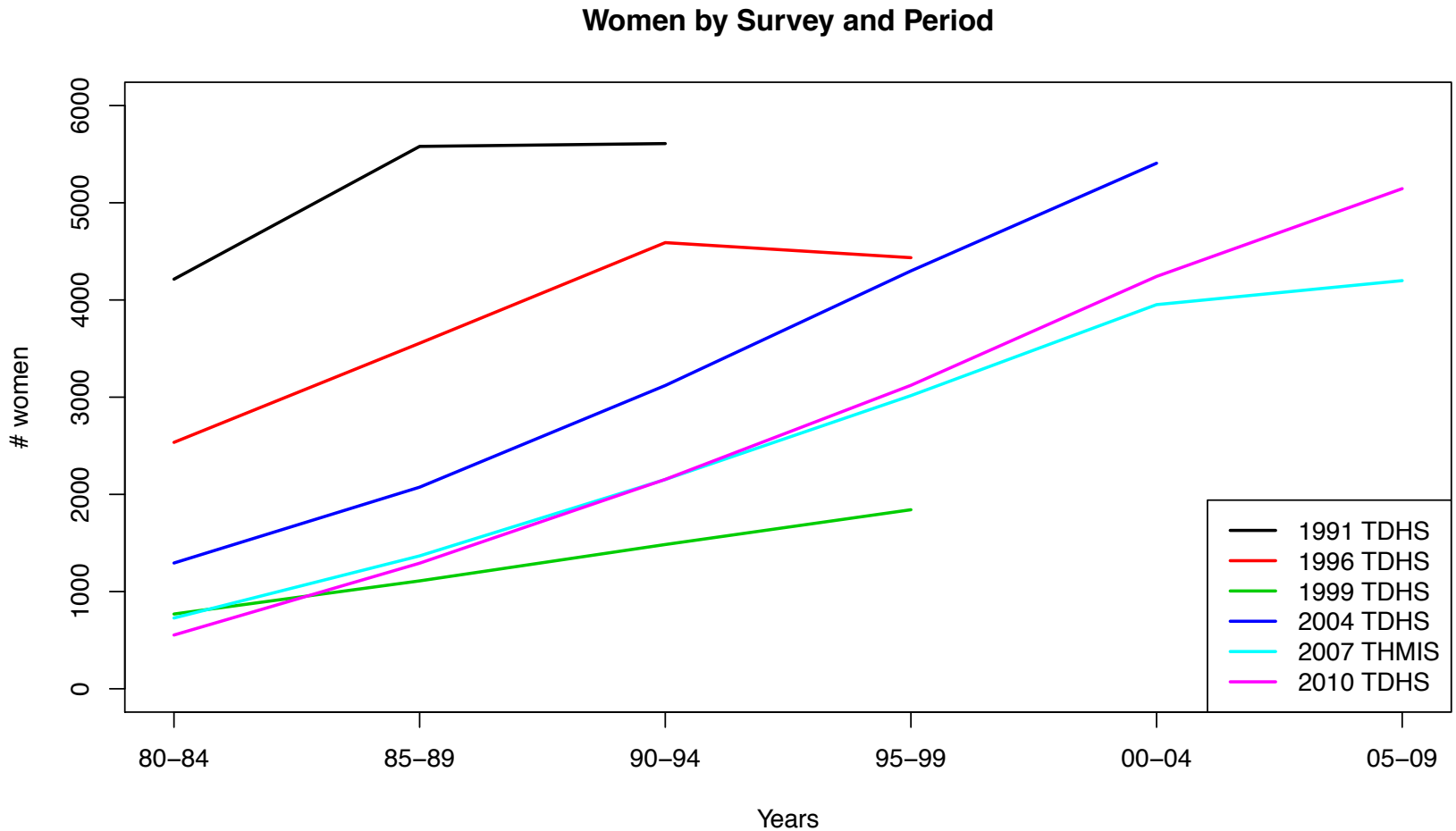

Figure 14 : Numbers of mothers by survey and period.

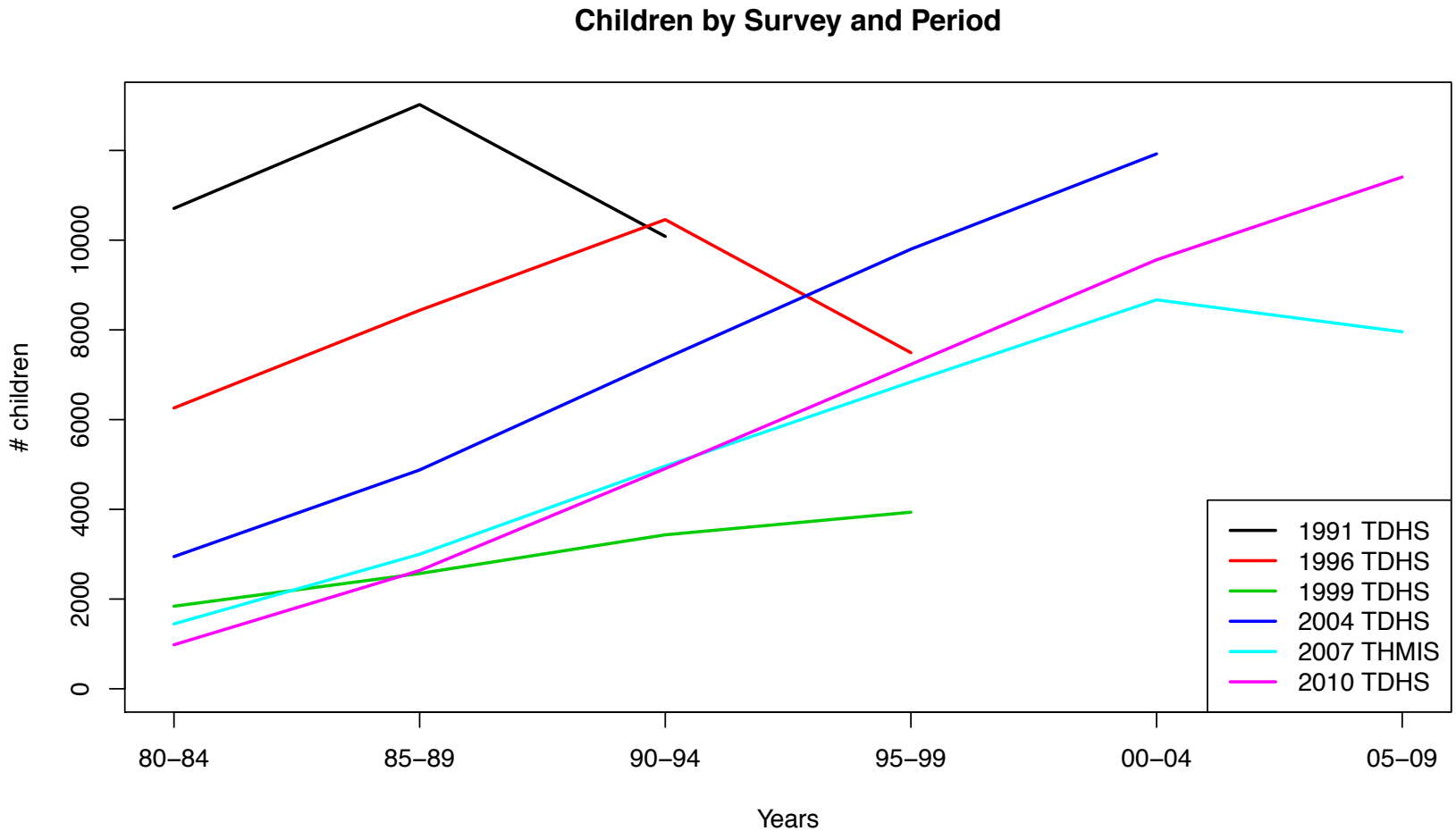

Figure 15 : Numbers of children, by survey and period.

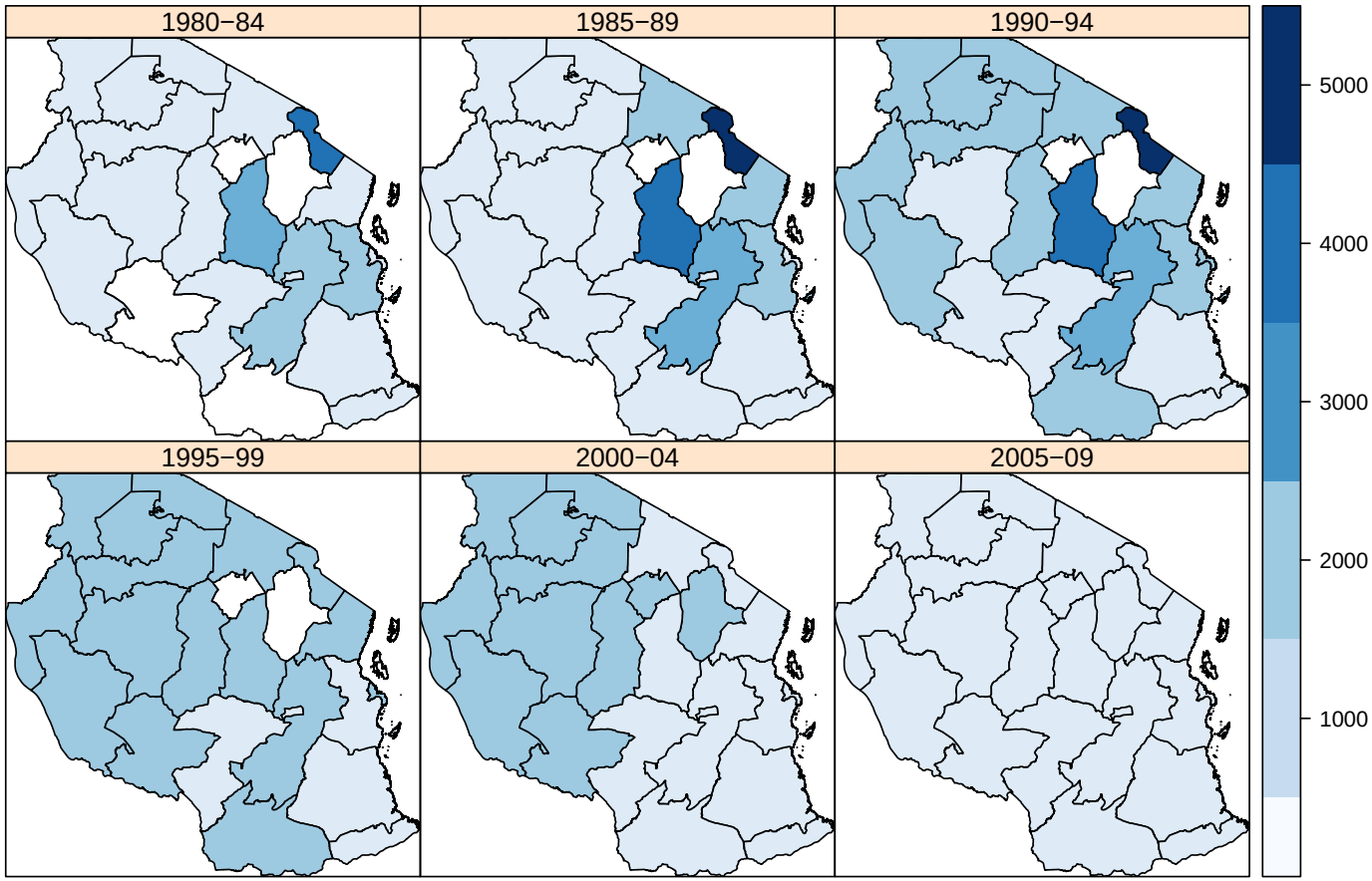

Figure 16 : Regional sample sizes (numbers of births) in DHSs.

Relevant substantive literature

Previously, United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME) produced national estimates (United Nations Inter-Agency Group for Child Mortality Estimation, 2011), with a local smoother in time.

IHME produced national estimates, summarized in Lozano *et al.* (2011) a complex method with many moving parts is used.

Alkema and You (2012) compared approaches and looked at reasons for discrepancies between estimates.

Alkema *et al.* (2014) and Alkema and New (2014) describe a Bayesian method, using B-splines, for national modeling — this has been adopted by the UN.

Relevant substantive literature

Dwyer-Lindgren *et al.* (2014) considered small area U5MR estimation with space-time smoothing:

- Extensive simulation study compared a variety of space-time models in terms of bias, variance and coverage.
- Survey acknowledged in point estimates, single unknown variance.
- Estimates (with uncertainty) obtained for districts in Zambia over 1980–2010 with careful examination of between district variability.

Discrete survival model

Define

$$ {}_nq_x = \Pr(\text{dying before } x+n \mid \text{lived until } x). $$

Split the [0,5) period into $J+1$ intervals

$$ [x_0, x_1), [x_1, x_2), \ldots, [x_J, x_{J+1}), $$

where $x_{j+1} = x_j + n_j$ so that n_j is the length of the interval beginning at x_j, $j = 0, \ldots, J$.

The probability of death in $[x_j, x_{j+1})$, given survival until x_j is ${}_{n_j}q_{x_j}$.

The U5MR is calculated as

$$ {}_5q_0 = 1 - \prod_{j=0}^{J} (1 - {}_{n_j}\, q_{x_j}). \tag{9} $$

Discrete survival model

The monthly probability of death for each interval,

$$p_j = \Pr(\text{dying in any month in the interval } x_j + n_j \mid \text{lived until } x_j),$$

may be estimated using a logistic generalized linear model (GLM):

$$\text{logit}\,(p_j) = \sum_{j=0}^{J} \beta_j I_j,$$

where I_j is the indicator for the $[x_j, x_{j+1})$ time interval.

In the complex survey context an important consideration is that the design weights must be acknowledged.

This is achieved by solving a (design) weighted score statistic (Binder, 1983): implemented in the `survey` package.

Results in parameter estimates $\widehat{\beta}$ and associated variance-covariance matrix, which accounts for the design.

Pseudo likelihood

Once estimates $\widehat{\beta}_j$ are estimated we can calculate

$$\widehat{p}_j = \frac{\exp(\widehat{\beta}_j)}{1 + \exp(\widehat{\beta}_j)}.$$

The complement of surviving each month of the interval $[x_j, x_j + n_j)$ is used to calculate

$${}_{n_j}\widehat{q}_{x_j} = 1 - (1 - \widehat{p}_j)^{n_j}$$

which may be substituted into (9) to give ${}_5\widehat{q}_0$.

How to determine the variance?

Pseudo likelihood

Transform to logit($_5\widehat{q}_0$) and use the delta method to find the asymptotic design-based variance, V_{DES}.

Let

$$\begin{aligned} y &= \log\left(\frac{{}_5\widehat{q}_0}{1-{}_5\,\widehat{q}_0}\right) \\ \eta &= \log\left(\frac{{}_5 q_0}{1-{}_5\,q_0}\right) \end{aligned}$$

Bottom line: We have a "working likelihood":

$$y|\eta \sim \mathsf{N}\left(\eta, V_{\text{DES}}\right).$$

Framework

Components of the model to distinguish sources of variability:

- Likelihood: Sampling variability, including which individual's are selected, i.e. hypergeometric sampling variability.
- Spatial Prior Model: Main effect of space: proxy for health care availability, disease burden (malaria and HIV), etc, in each area.
- Temporal Prior Model: Main effect of time, reflecting overall changes in risk factors.
- Spatio-temporal Model: Interaction: how are risk factors.
- Survey Model: Particular surveys may be systematically biased, and this bias may change over time and space. The HDSS only cover a small region within the areas we consider, and may not reflect the U5MR in the complete area and may be vulnerable to the Hawthorne effect.

How to define a likelihood when the design variables are not available?

Key Idea: Use the design-based (weighted estimator) and its sampling distribution as the "working likelihood".

Space-time-survey model

Let ${}_5\widehat{q}_{0its}$ represent the weighted estimate of U5MR from survey s in region i and in period t.

We summarize the data in area i at time point t from survey s via the asymptotic distribution of the estimator of the empirical logit:

$$y_{its} = \log\left(\frac{{}_5\widehat{q}_{0its}}{1 -_5 \widehat{q}_{0its}}\right).$$

We define the area, period and survey summary as

$$\eta_{its} = \log\left(\frac{{}_5q_{0its}}{1 -_5 q_{0its}}\right).$$

We take as working likelihood the asymptotic distribution

$$y_{its} \mid \eta_{its} \sim \mathrm{N}\left(\eta_{its}, \widehat{V}_{\mathrm{DES},its}\right) \tag{10}$$

which has been shown to perform well in the context of small area estimation from complex surveys (Mercer *et al.*, 2014).

Random Effect Models

Table 3 : Random effects models for time period t, region i and survey s.

Model	Linear Predictor η_{its}
I	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it}$
II	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s$
III	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{is}$
IV	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts}$
V	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is}$
VI	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is} + \nu_{its}$

- Time: Indept: $\alpha_t \sim_{iid} \mathsf{N}(0, \sigma^2_\alpha)$, Smooth: $\gamma_t \sim \mathsf{RW1}(\sigma^2_\gamma)$,
- Space: Indept: $\theta_i \sim_{iid} \mathsf{N}(0, \sigma^2_\theta)$, Smooth: $\phi_i \sim \mathsf{ICAR}(\sigma^2_\phi)$,
- Space-Time Interaction: $\delta_{it} \sim_{iid} \mathsf{N}(0, \sigma^2_\delta)$.
- Survey: Indept: $\nu_s \sim_{iid} \mathsf{N}(0, \sigma^2_{\nu 1})$, Space-Survey Interaction: $\nu_{is} \sim_{iid} \mathsf{N}(0, \sigma^2_{\nu 2})$, Time-Survey Interaction: $\nu_{ts} \sim_{iid} \mathsf{N}(0, \sigma^2_{\nu 3})$, Space-Time-Survey Interaction: $\nu_{its} \sim_{iid} \mathsf{N}(0, \sigma^2_{\nu 4})$.

Hyperpriors

Hyperpriors on variances:

$$\sigma^2_\alpha, \quad \sigma^2_\gamma, \quad \sigma^2_\theta, \quad \sigma^2_\phi, \quad \sigma^2_\delta, \quad \sigma^2_{\nu 1}, \quad \sigma^2_{\nu 2}, \quad \sigma^2_{\nu 3}, \quad \sigma^2_{\nu 4}$$

are difficult to specify in this setting.

We followed the procedure described in Wakefield (2009).

Briefly, one specifies a range of residual odds and a prior is assigned that results in this range.

The intrinsic models (RW1 and ICAR) need a fix up since their marginal variance does not exist.

Computation

Model fitting was carried out within the `R` computing environment.

Weighted logistic regressions were fit using the `svyglm()` function from the `survey` package (Lumley, 2004) from which the design-based variance was extracted.

The hierarchical Bayesian space-time models were fitted using the Integrated Nested Laplace Approximation (INLA) (Rue *et al.*, 2009) as implemented in the `INLA` package.

Model Comparison

Markov chain Monte Carlo in particular has allowed the fitting of more and more complex models, often hierarchical in nature with layers of random effects.

The search for a method to find the "best" of a set of candidate models has also grown.

Let $p(\mathbf{y}|\boldsymbol{\theta})$ represent a generic likelihood for $\mathbf{y} = [y_1, \ldots, y_n]$ and let

$$D(\boldsymbol{\theta}) = -2\log[p(\mathbf{y}|\boldsymbol{\theta})]$$

represent the deviance.

For example, in an iid $\mathrm{N}(\mu_i(\boldsymbol{\theta}), \sigma^2)$ normal the deviance is

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i - \mu_i(\boldsymbol{\theta})]^2.$$

Frequentist model comparison for nested models is often carried out using likelihood ratio statistics, which corresponds to the comparison of deviances in generalized linear models (GLMs), see for example McCullagh and Nelder (1989).

Model Comparison: AIC

One approach to model comparison is based on a model's ability to make good predictions.

Such an objective, and predicting the actual observed data, leads to Akaike's an information criterion (AIC), derived in Akaike (1973).

In AIC one tries to estimate the (Kullback-Leibler) distance between the true distribution of the data, and the modeled distribution of the data.

Model Comparison: AIC

AIC is given by

$$\text{AIC} = -2\log[p(y|\widehat{\boldsymbol{\theta}})] + 2k$$

where $\widehat{\theta}$ is the MLE and k is the number of parameters in the model, i.e. the size of $\boldsymbol{\theta}$.

Small values of the AIC are favored, since they suggest low prediction error.

The penalty term $2k$ penalizes the double use of the data.

In general for prediction: overly complex models are penalized since redundant parameters "use up" information in the data.

Model Comparison: BIC

Another approach is based on trying to identify the "true" model.

Schwarz (1978) developed the Bayesian Information Criterion (BIC) which is given by

$$\text{BIC} = -2\log[p(y|\widehat{\boldsymbol{\theta}})] + k\log n.$$

BIC approximates $-2\log p(\mathbf{y}|\boldsymbol{\theta})$ under a certain unit information prior (Kass and Wasserman, 1995).

BIC is consistent[3] for finding the true model, if that model lies in the set being compared.

AIC is not consistent for finding the true model, but recall is intended for prediction.

[3] meaning the BIC hones in on the true model as the sample size increases

Model Comparison: DIC

Spiegelhalter *et al.* (2002) introduced what has proved to be a very popular model comparison statistic, the deviance information criterion (DIC).

To define the DIC, define an "effective number of parameters" as

$$\begin{aligned} p_D &= E_{\theta|y}\{-2\log[p(\mathbf{y}|\boldsymbol{\theta})]\} + 2\log[p(\mathbf{y}|\overline{\boldsymbol{\theta}})] \\ &= \overline{D} + D(\overline{\boldsymbol{\theta}}) \end{aligned}$$

where $\overline{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\mathbf{y}]$ is the posterior mean, $D(\overline{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean and $\overline{D} = E[D|\mathbf{y}]$.

Hence, p_D is the

$$\text{posterior mean deviance} - \text{deviance of posterior means.}$$

Model Comparison: DIC

The DIC is given by

$$\begin{aligned} \text{DIC} &= D(\overline{\boldsymbol{\theta}}) + 2p_D \\ &= \overline{D} + p_D, \end{aligned}$$

so that we have a measure of goodness of fit + complexity.

DIC is straightforward to evaluate using MCMC or INLA.

Model Comparison: DIC

DIC has been heavily criticized (Spiegelhalter *et al.*, 2014):

- p_D is not invariant to parameterization.
- DIC is not consistent for choosing the correct model.
- DIC has a weak theoretical justification and is not universally applicable.
- DIC has been shown to under penalize complex models (Plummer, 2008; Ando, 2007).
- See Spiegelhalter *et al.* (2014) for an interesting discussion of the history of DIC, including a summary of attempts to improve DIC.
- According to Google Scholar, as of June 20th, 2014, Spiegelhalter *et al.* (2002) has 5251 citations...

Model Comparison: CPO

Another approach based on prediction uses the conditional predictive ordinate (CPO).

Let

$$\mathbf{y}_{-i} = [y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n]$$

represent the vector of data with the i-th observation removed.

The idea is to predict the density ordinate of the left-out observation, based on those that remain.

Specifically, the CPO for observation i is defined as:

$$\begin{aligned} \text{CPO}_i &= p(y_i|\mathbf{y}_{-i}) \\ &= \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{-i}) \, d\boldsymbol{\theta} \\ &= E_{\theta|y_{-i}} \left[p(y_i|\boldsymbol{\theta}) \right] \end{aligned}$$

Model Comparison: CPO

The CPOs can be used to look at local fit, or one can define an overall score for each model:

$$\log(\text{CPO}) = \sum_{i=1}^{n} \log \text{CPO}_i.$$

Good models will have relatively high values of $\log(\text{CPO})$.

See Held *et al.* (2010) for a discussion of shortcuts for estimation (i.e. avoidance of fitting the model n times) using MCMC and INLA.

Results

Table 4 : Model comparison: p_D is the effective degrees of freedom, as defined for the calculation of the deviance information criteria (DIC), which also uses the deviance evaluated at the posterior mean, $\overline{D}$; LCPO is defined as $\sum_{its} \log(CPO_{its})$. Also include the marginal distribution of the data.

Model	No Pars	$\log p(\mathbf{y})$	p_D	$\overline{D}$	DIC	LCPO
I	181	-311	75	409	484	-294
II	189	-305	80	384	464	-287
III	313	-258	119	222	341	-194
IV	223	-302	89	368	456	-283
V	347	**-255**	122	210	**332**	**-183**
VI	920	-255	135	199	334	-184

- $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the evidence in the data for a particular model.
- The ratio of marginal distrbutions under different models constitutes the Bayes factor.

Components of variability

Table 5 : Summaries of variance components. The proportion of variation is calculated as the contribution the relevant set of random effects makes to the total variation. In the case of the RW1 and ICAR models, the relevant contribution is evaluated empirically, since the variance parameter is conditional rather than marginal.

Variance	Interpretation	Median (95% Interval)	Percent Var
σ^2_α	Indept Time	0.003 (0.001, 0.035)	2.5
σ^2_γ	RW1 Time	0.038 (0.012, 0.146)	**43.3**
σ^2_θ	Indept Space	0.067 (0.033, 0.131)	**32.0**
σ^2_ϕ	ICAR Space	0.016 (0.002, 0.342)	5.0
σ^2_δ	Indept Space-Time Interaction	0.005 (0.001, 0.013)	2.4
$\sigma^2_{\nu_s}$	Indept Survey	0.002 (0.001, 0.013)	1.5
$\sigma^2_{\nu_{st}}$	Indept Survey-Time Interaction	0.004 (0.001, 0.012)	2.1
$\sigma^2_{\nu_{si}}$	Indept Survey-Space Interaction	0.024 (0.015, 0.038)	11.2

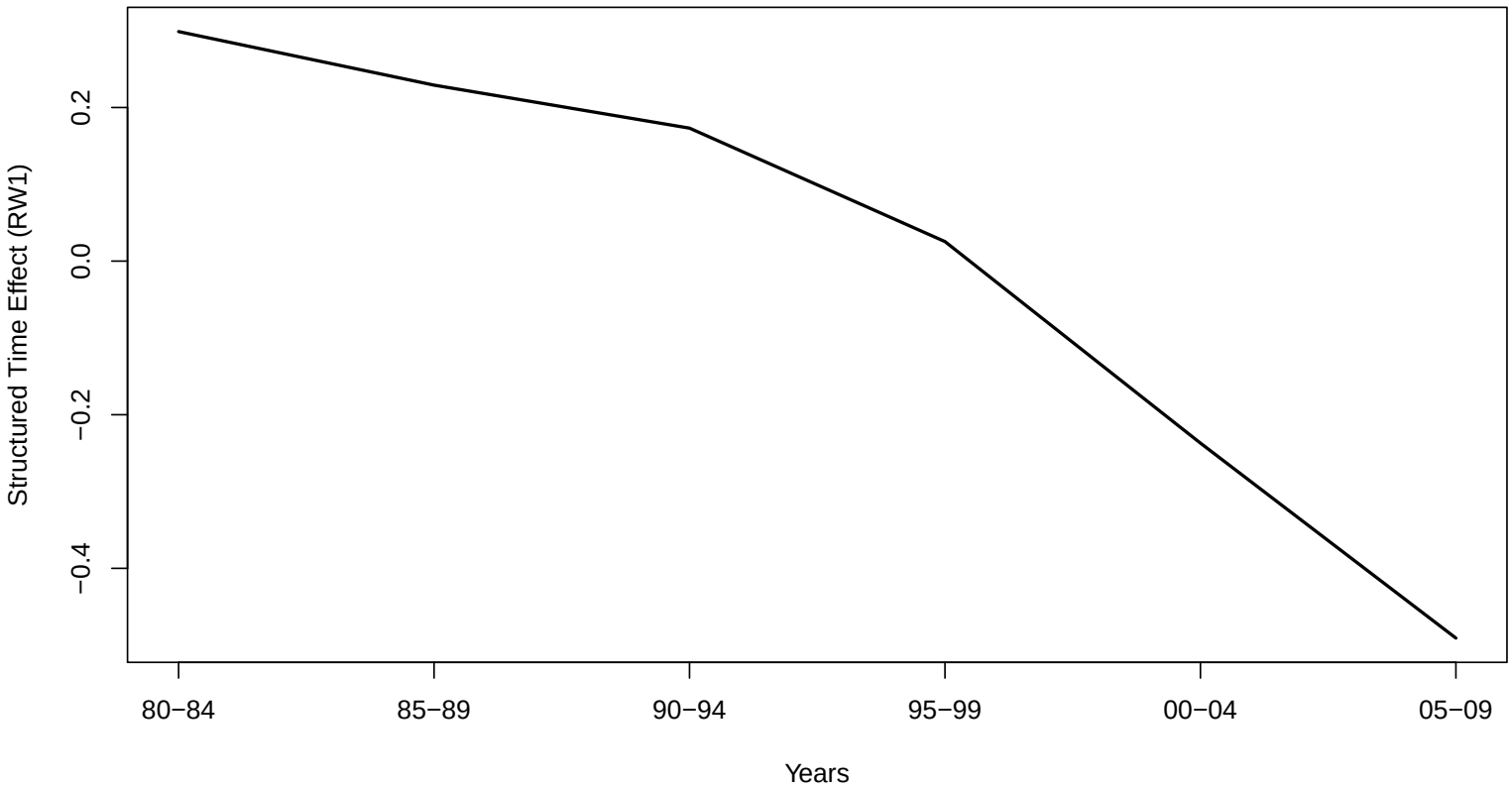

Figure 17 : Structured time RW1 random effects, (γ_t).

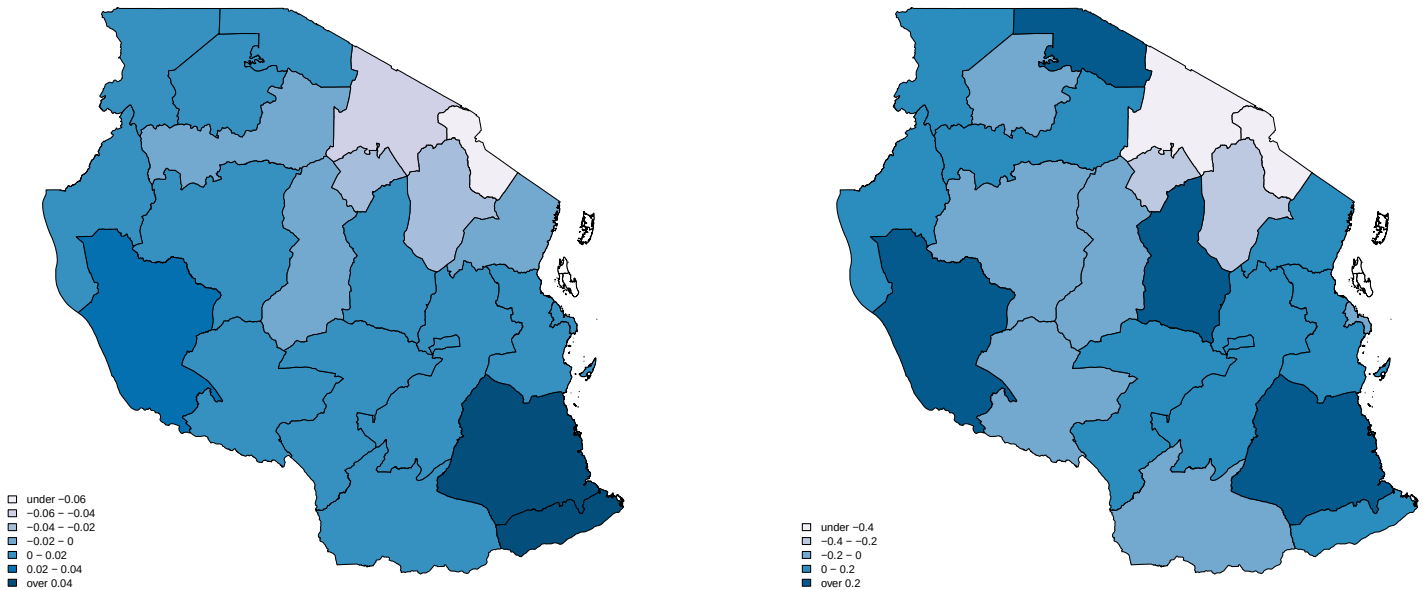

Figure 18 : ICAR random effects, ϕ_i (left) and unstructured spatial random effects, θ_i (right). Note the difference in the scales, and the heterogeneity in the right plot.

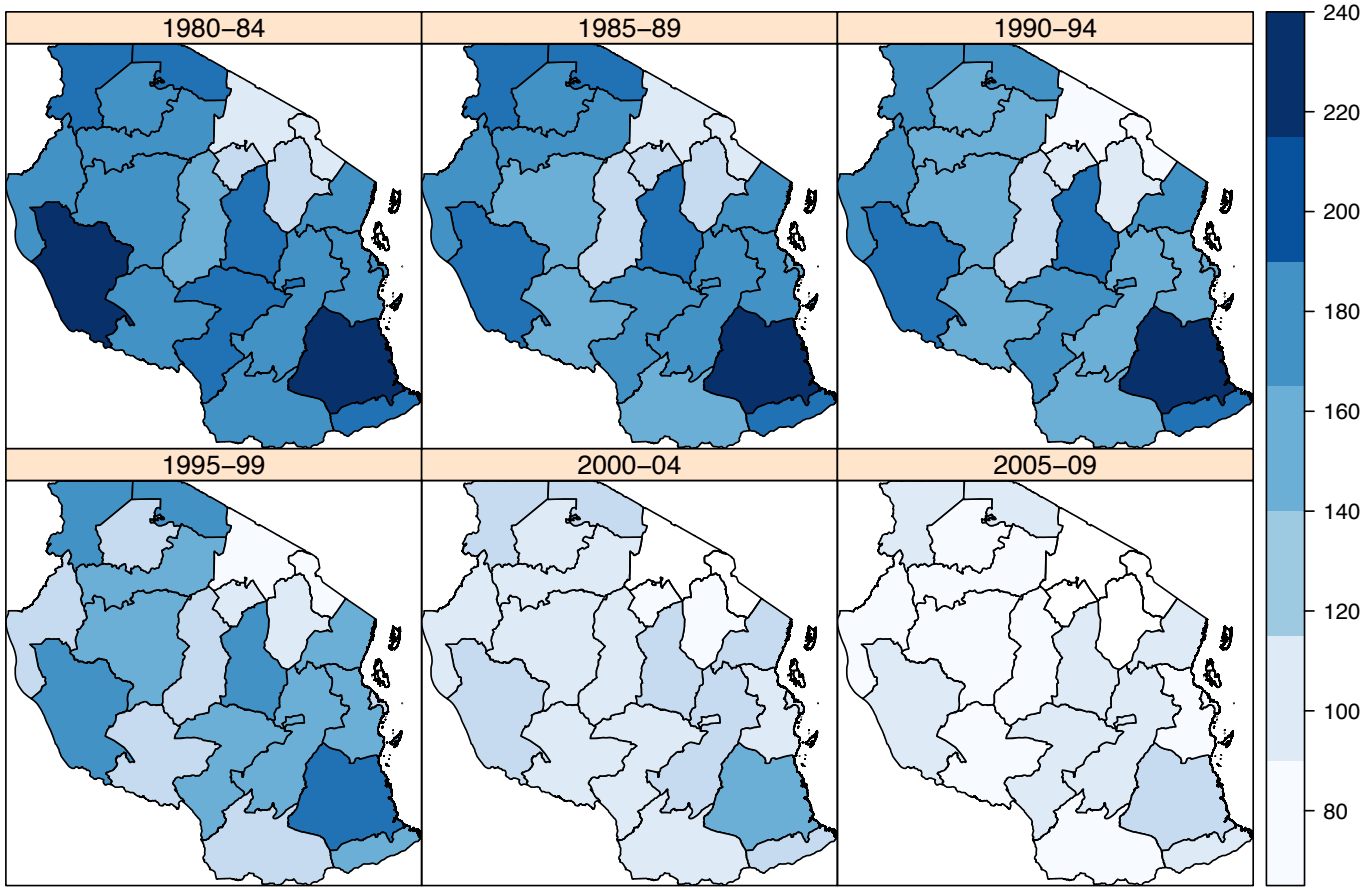

Figure 19 : Smoothed regional estimates of child mortality (per 1000 births).

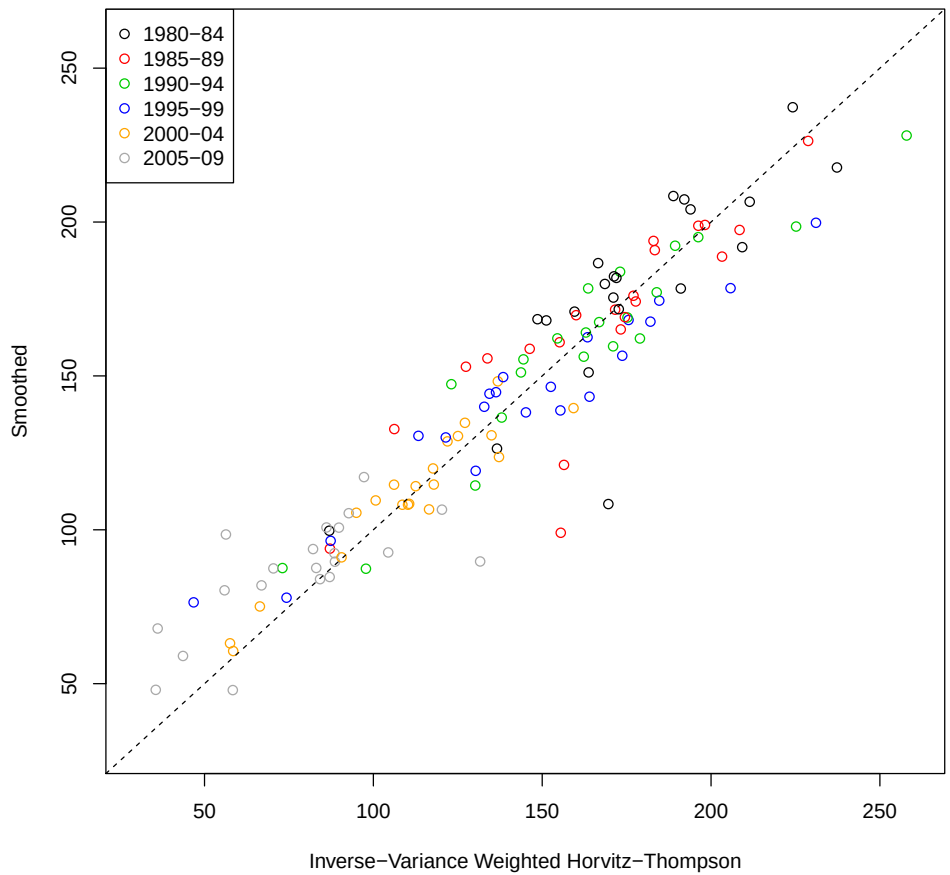

Figure 20 : Smoothed and inverse-variance weighted Horvitz-Thompson regional estimates of child mortality (per 1000 births).

Prior Sensitivity

We consider three ranges of residual odds: [0.5,2], [0.2,5], [0.1,10].

We see sensitivity for the spatial random effects, though the total spatial random effects contributions remain relatively constant since as the structured random effects increase, the unstructured random effects decrease.

The structured temporal random effects are robust, which is reassuring since these provide the largest contribution to the overall variability; these are well-estimated since the trend is strong.

Similarly, the unstructured survey-area random effects are robust, but all of the standard deviations of the remaining independent random effects show modest increases as the prior moves further from zero.

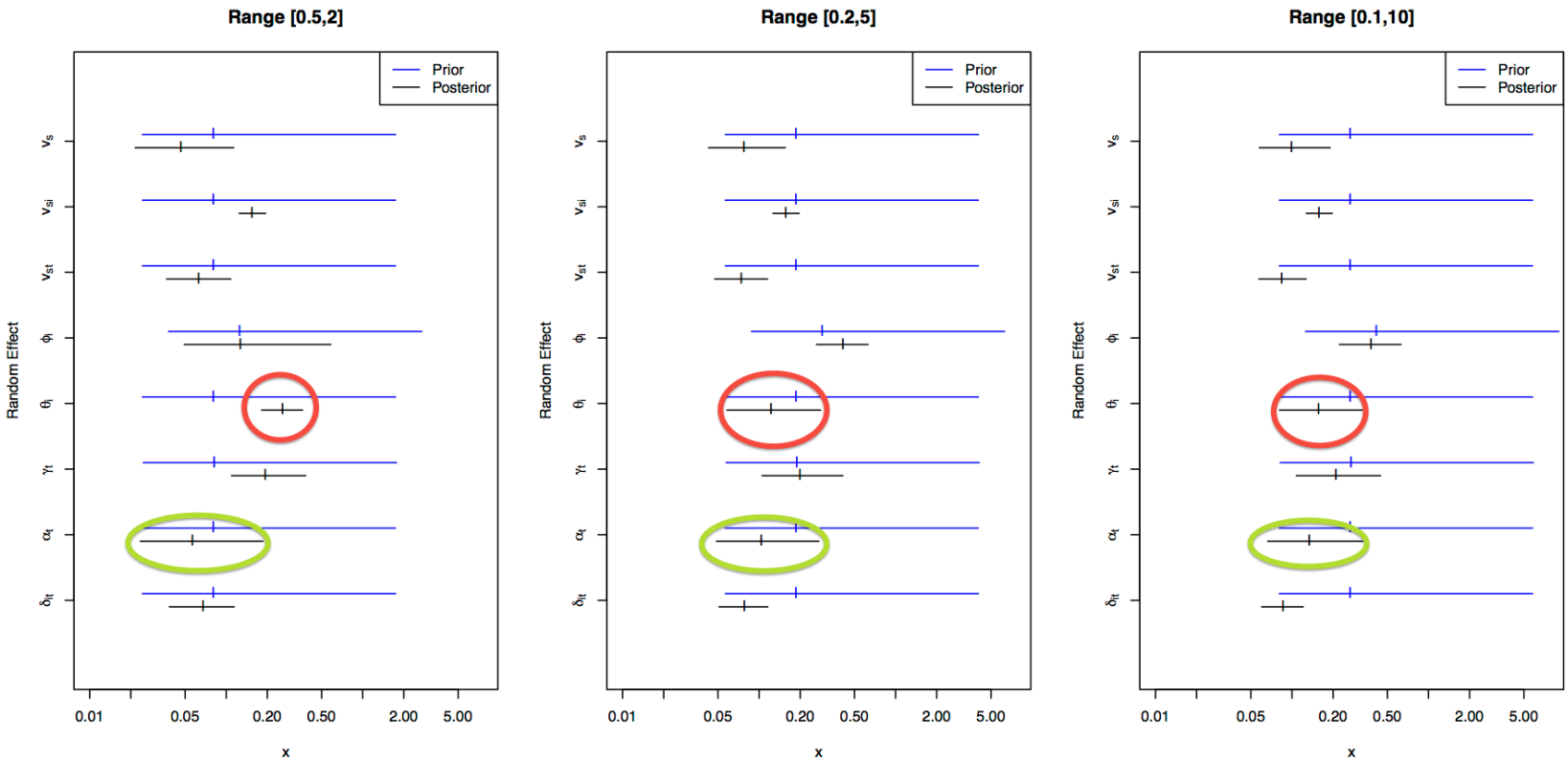

Figure 21 : Prior sensitivity of the standard deviations of the eight random effects in the model. The three priors are based on 95% prior intervals on the residual odds of [0.5,2], [0.2,5], [0.1,10].

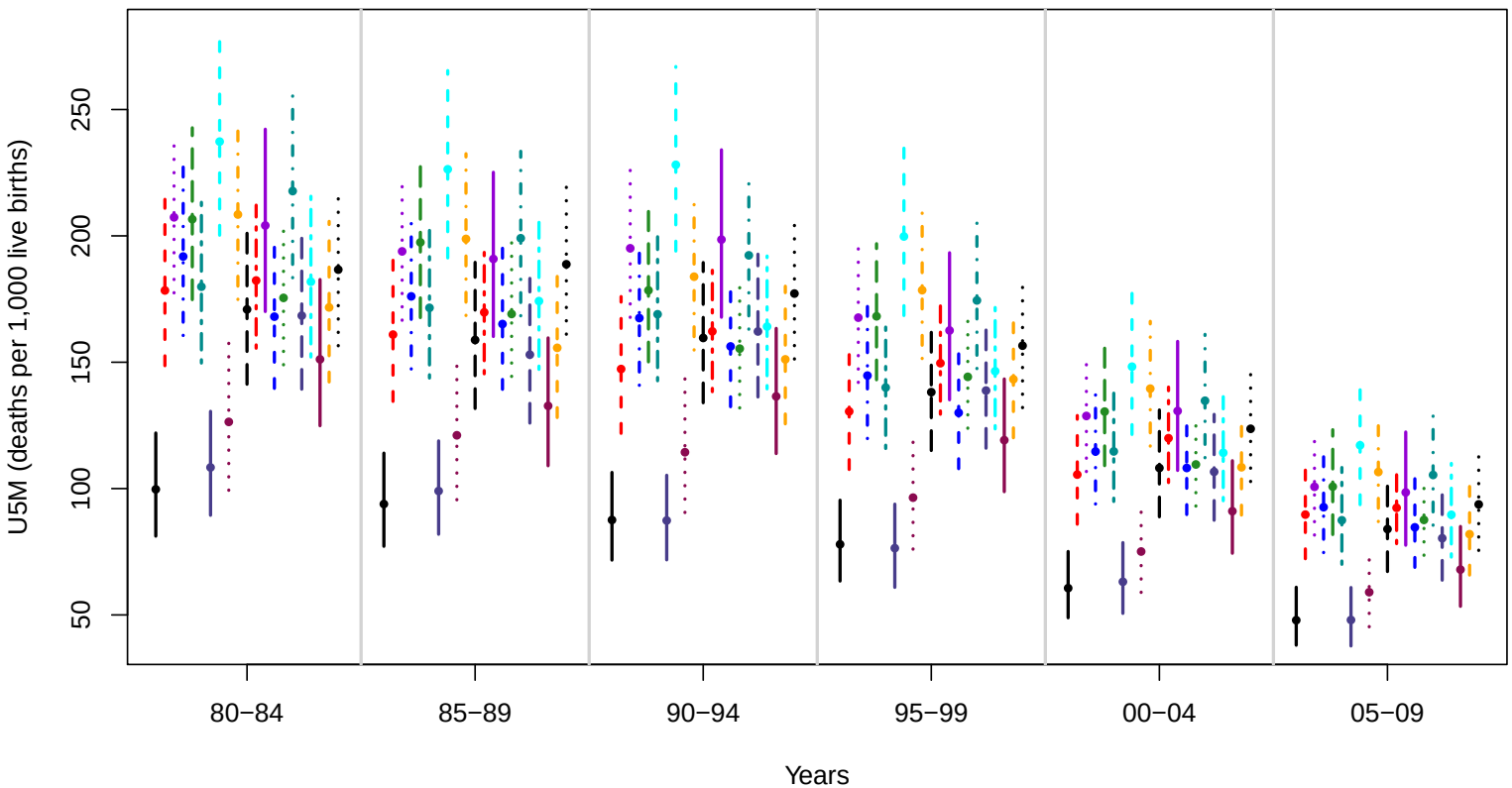

Figure 22 : Posterior distributions (2.5, 50, 97.5% points) of the U5MR for each of the regions, as a function of period.

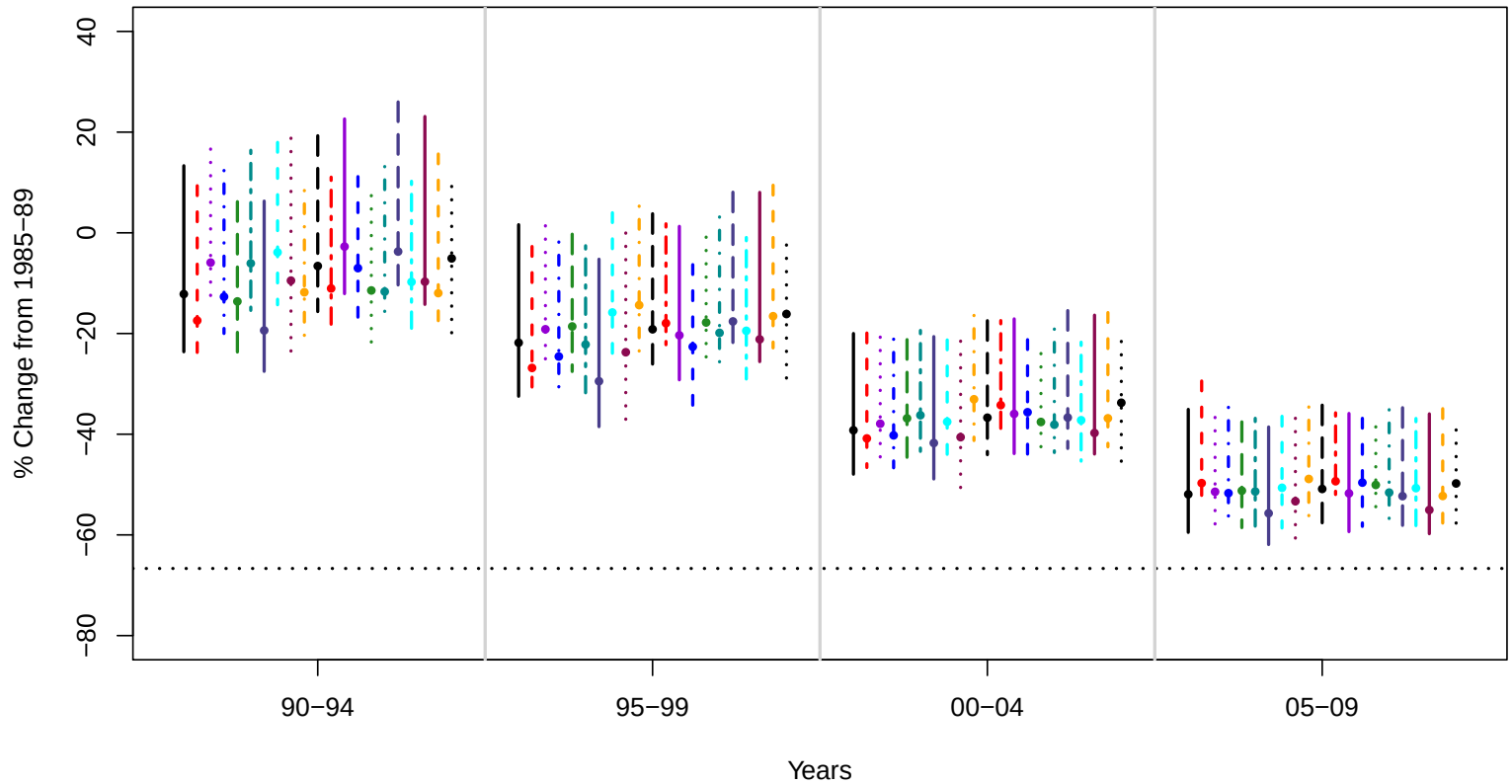

Figure 23 : Posterior distributions (2.5, 50, 97.5% points) of the changes in U5MR from posterior medians in 1985–1989, for each of the regions, as a function of period.

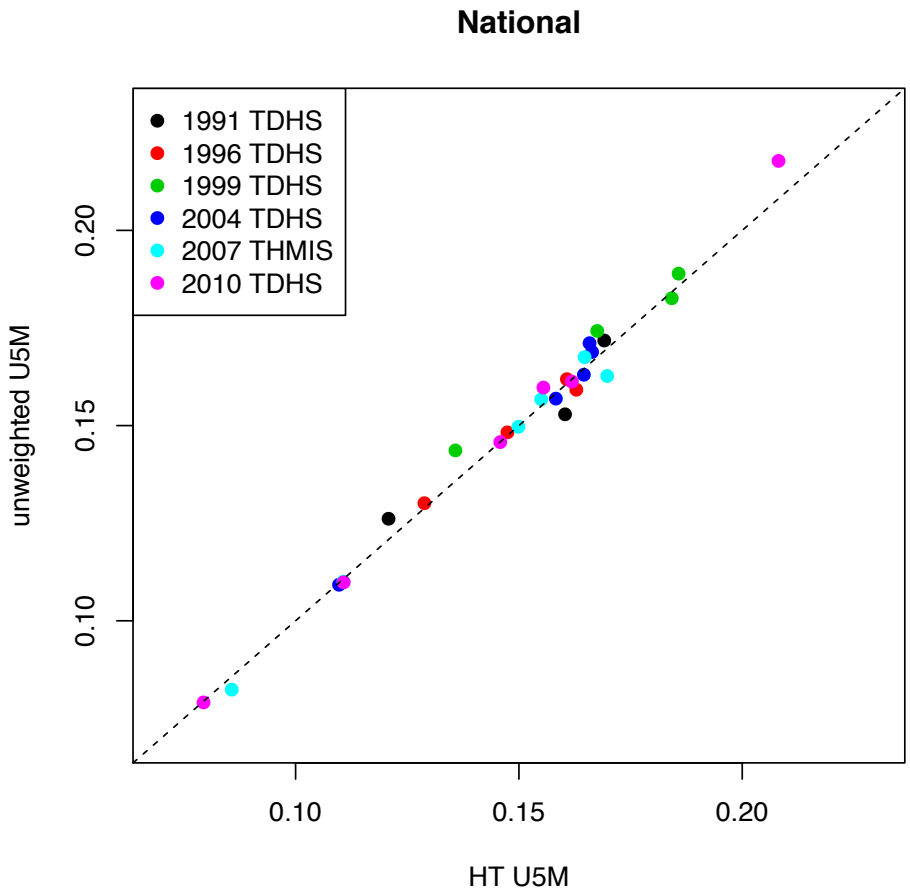

Figure 24 : Effect of weighting: national estimates.

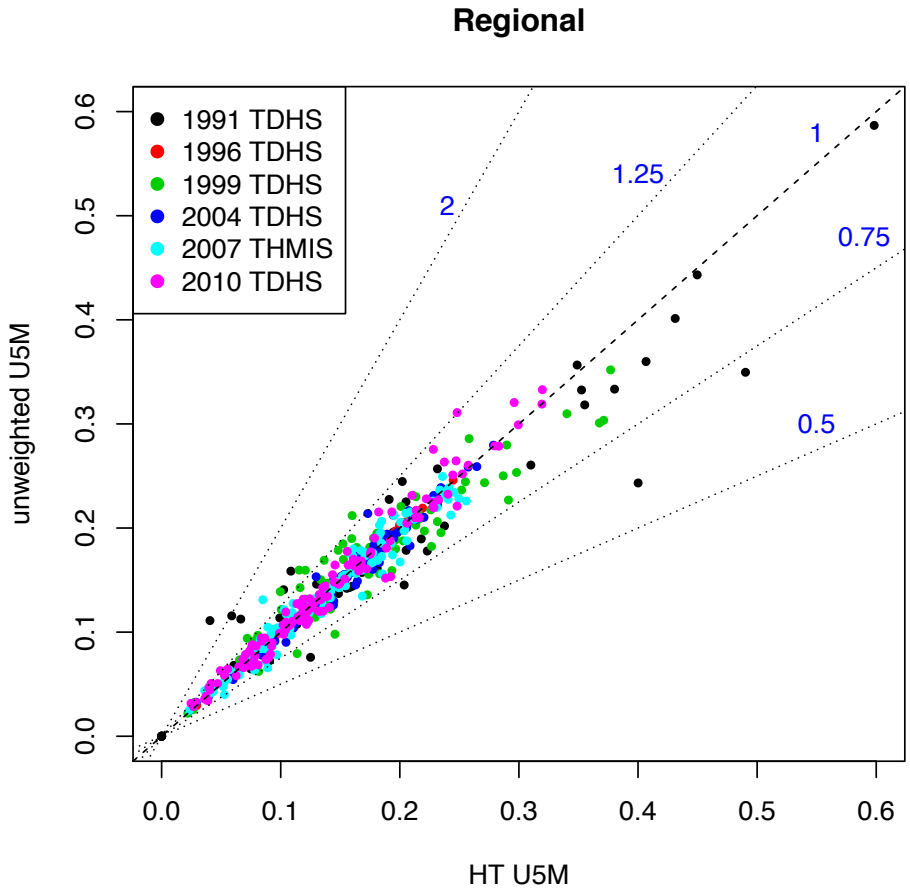

Figure 25 : Effect of weighting: regional estimates.

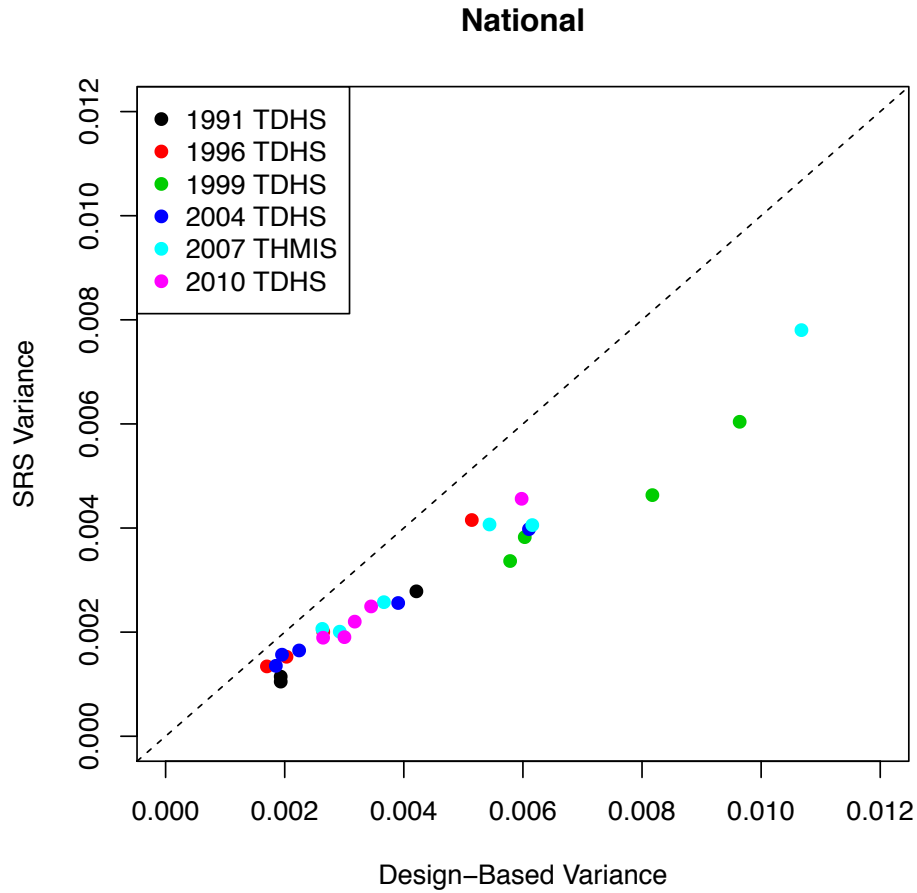

Figure 26 : Effect of weighting: national variances.

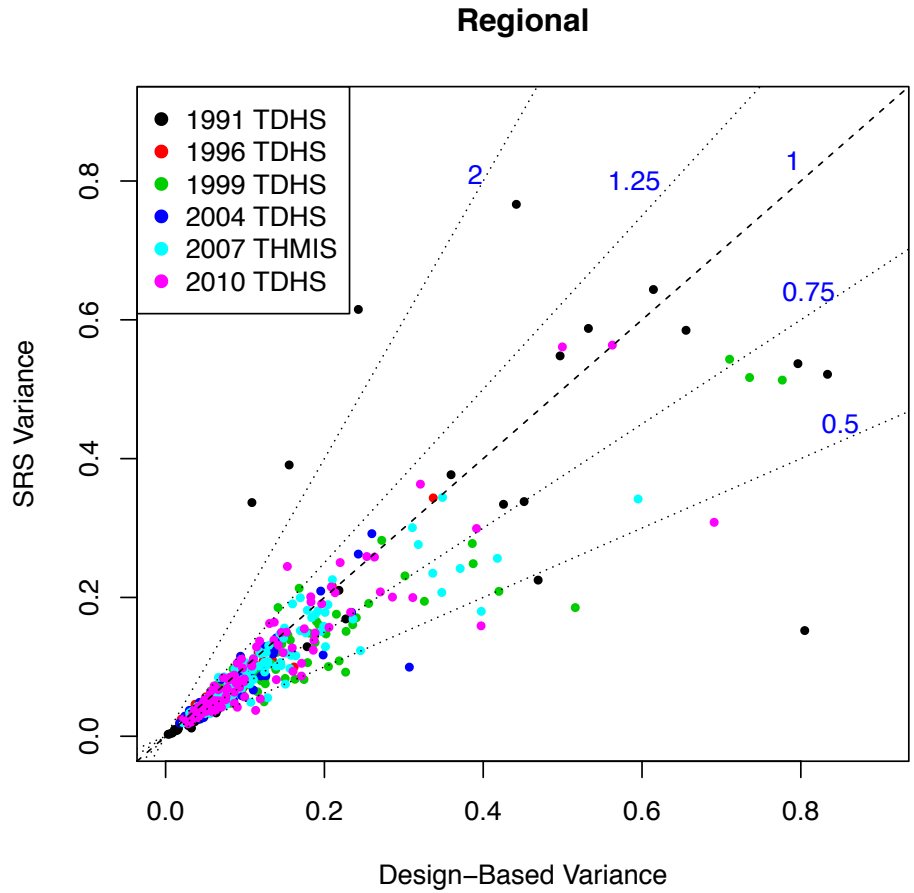

Figure 27 : Effect of weighting: regional variances.

Conclusions: Hierarchical Modeling

Hierarchical models allow complex dependencies within data to be modeled.

Prior specification for variance components (in particular) is not straightforward, and sensitivity analysis is a good idea.

No universally agreed upon approach to carrying out model comparison.

Model checking is difficult.

Conclusions: Spatial Modeling of Survey Data

No widely accepted approach to spatial smoothing that adjusts for the sampling scheme.

Whether to weight or not is contentious, it depends on how bad the bias is, since weighting in general increases the variance.

If the design variables upon which sampling is based are available then a model-based hierarchical approach is available.

Gelman (2007) and discussants provide a range of views.

Traditional SAE Approaches: Notation

We use notation more consistent with the survey sampling literature in this section.

We assume the study region can be partitioned into $d = 1, \ldots, D$ sub-regions (domains) with N_d being the population of the domain, which may or not be known.

A survey is carried out and n_d is the sample size in domain d; if the survey was not designed to fix the sample size n_d for domain d then it is a random variable with respect to the randomization distribution and we need to consider ratio estimation.

Let U_d and S_d, $d = 1, \ldots, D$, be the index sets for the units of the population and the sample respectively in domain d with $U = U_1 \cup \cdots \cup U_D$ and $S = S_1 \cup \cdots \cup S_D$.

Notation

The population mean in domain d is

$$\overline{y}_{U_d} = \frac{\sum_{k \in U_d} y_k}{N_d}.$$

We define

$$\begin{aligned} z_{dk} &= \begin{cases} 1 & \text{if } k \in U_d \\ 0 & \text{if } k \notin U_d \end{cases} \\ u_{dk} = y_k z_{dk} &= \begin{cases} y_k & \text{if } k \in U_d \\ 0 & \text{if } k \notin U_d \end{cases} \end{aligned}$$

so that z_{dk} is just an indicator of whether unit k lies within domain d and u_{dk} is the value of y for such units.

Two key quantities:

- $t_d = \sum_{k=1}^{N} u_{dk}$ is the population total in domain d.
- $N_d = \sum_{k=1}^{N} z_{dk}$ is the population size in domain d.

Direct estimation without auxiliary information

A direct estimator is one in which response data y from the domain only are used.

An indirect estimator uses responses from other domains.

The population average in domain d can be written as a ratio of totals:

$$\overline{y}_{U_d} = \frac{t_d}{N_d} = B.$$

If N_d is unknown, this suggests ratio estimation as a way forward.

Direct domain estimation

Under a general design π_k is the probability of selection for unit k and $w_k = \pi_k^{-1}$ is the associated design weight.

The domain sample sizes are

$$n_d = \sum_{k \in U} z_{dk} I_k = \sum_{k \in U_d} I_k,$$

where I_k are the usual sample membership indicators.

Note that

$$\mathrm{E}[n_d] = \sum_{k \in U} z_{dk} \pi_k = \sum_{k \in U_d} \pi_k,$$

so that the choice of π_k indicates whether we are likely to have a sufficiently large sample in domain d.

Direct domain estimation

For general sampling we refer to Särndal et al (1992, Section 10.3).

Whether N_d is known or unknown, Särndal et al (1992, p. 391) recommend the domain mean (Hajek) as:

$$\widehat{\overline{y}}_d = \frac{1}{\widehat{N}_d} \sum_{k \in S_d} w_k y_k = \frac{1}{\widehat{N}_d} \sum_{k \in S} z_k w_k y_k,$$

where $w_k = 1/\pi_k$ and

$$\widehat{N}_d = \sum_{k \in S_d} w_k = \sum_{k \in S} z_k w_k.$$

Direct domain estimation

To estimate the domain total when N_d is unknown

$$\widehat{t}_d = \sum_{k \in S_d} w_k y_k.$$

To estimate the domain total when N_d is known

$$\widetilde{t}_d = N_d \times \widehat{\overline{y}}_d = \frac{N_d}{\widehat{N}_d} \sum_{k \in S_d} w_k y_k.$$

Variance estimators are given in Särndal et al (1992, Section 10.3).

Direct domain GREG with study auxiliary information

The problem with using the direct ratio estimators is that the variance may be large in areas with low n_d.

When auxiliary variable is available, this may be used to define a new estimator; suppose we have a single variable x for which the total is known, across all domains, t_x, and we have a HT estimator $\widehat{t}_x$.

In general, GREG with multiple x values and a linear regression model may be utilized; we describe some special cases.

A ratio estimator is

$$\widehat{t}_d^{\,\text{dir,rat1}} = \widehat{t}_d \times \frac{t_x}{\widehat{t}_x}, \tag{11}$$

where $\widehat{t}_d$ is the usual HT estimator.

This is a direct domain estimator because y values only from the domain are used, though the total x, t_x from all domains are used.

Direct domain GREG with study auxiliary information

This estimator is approximately unbiased, if the overall sample size n is large (because $\widehat{t}_x \rightarrow t_x$) , and design consistency occurs as the domain sample size n_d increases.

See Rao and Molina (2015, Section 2.4.2) describe GREG estimators, and give a number of special cases including (11).

Notice that the same adjustment is made to every area.

Example: Smoking by county in Washington State

As an example suppose we wish to estimate the number of current smokers across the 40 counties of Washington State, based on a survey.

For each individual in the survey, information is collected on $y_k, k \in S$, a binary indicator of current smoking status, along with x_k, the income and the basic demographics (age and gender).

Suppose we know the total income of residents in Washington State, t_x; then (11) may be directly applied with $\widehat{t}_x = \sum_{k \in s} w_k x_k$ being the estimated total income from the sample.

Direct domain GREG with study auxiliary information

Suppose now we have auxiliary information on the population sample sizes across (usually demographic) groups g, $g = 1, \ldots, G$.

A post-stratified estimator (Rao and Molina 2015, Section 2.4.2) is

$$\widehat{t}_d^{\,\text{dir,ps1}} = \sum_{g=1}^{G} \frac{N_{.g}}{\widehat{N}_{.g}} \sum_{k \in S_{dg}} w_k y_k = \sum_{g=1}^{G} \frac{N_{.g}}{\widehat{N}_{.g}} \widehat{t}_{dg}. \tag{12}$$

where

- S_{dg} is the set of samples falling in post-stratification group g of domain d,
- $\widehat{t}_{dg}$ is the estimate of the total for y in domain d and group g (note that $\widehat{t}_d = \sum_g \widehat{t}_{dg}$), and
- $\widehat{N}_{.g} = \sum_{k \in S_{.g}} w_k$.

This estimator is approximately unbiased (and is design consistent) but the variance can be large since the adjustments between domain d and the whole region may be large.

Example: Smoking by county in Washington State

Suppose we know the population totals for Washington State by 18 age bands and gender, $N_{.g}, g = 1, \ldots, G = 36$.

In county d, to use (12), we would estimate:

- the total number of smokers by stratum g, $\widehat{t}_{dg} = \sum_{k \in s_{dg}} w_k y_k$, this may have high variability as $|s_{dg}| = n_{dg}$ may be small,
- the population total by group g, across the state, $\widehat{N}_{.g} = \sum_{k \in s_{.g}} w_k$.

Direct domain GREG with domain auxiliary information

To reduce the bias we may use domain-specific auxiliary information, as described in (Rao and Molina 2015, Section 2.4.3).

A ratio estimator is

$$\widehat{t}_d^{\,\text{dir,rat2}} = \widehat{t}_d \times \frac{t_{xd}}{\widehat{t}_{xd}}, \tag{13}$$

where $\widehat{t}_d$ is the HT estimator.

This gives an area-specific adjustment.

This is a direct domain estimator since it uses y (and x values) only from the domain.

Direct domain GREG with domain auxiliary information

A post-stratified estimator is

$$\widehat{t}_d^{\,\text{dir,ps2}} = \sum_{g=1}^{G} \frac{N_{dg}}{\widehat{N}_{dg}} \widehat{t}_{dg}, \tag{14}$$

where $\widehat{N}_{dg} = \sum_{k \in S_{dg}} w_k$.

Example: Smoking by county in Washington State

For (13), suppose we know the income totals by county (from the census, for example), t_{xd}, and we then estimate $\widehat{t}_{xd} = \sum_{k \in s_d} w_k x_k$.

For (14) ,suppose we know the population totals for Washington State by 18 age bands and gender and by domain, $N_{dg}, g = 1, \ldots, G = 36$.

In county d we would then estimate:

- the total number of smokers by stratum g, $\widehat{t}_{dg} = \sum_{k \in s_{dg}} w_k y_k$,
- the population total by group g, across the state, $\widehat{N}_{.g} = \sum_{k \in s_{.g}} w_k$.

Synthetic estimation

Now we consider indirect estimators, and begin with synthetic estimation, as described in (Rao and Molina 2015, Section 3.2).

The simplest synthetic estimator of a domain mean for area d does not use auxiliary information and is

$$\widehat{\overline{y}}_d^{\text{syn,basic1}} = \frac{\widehat{t}_y}{\widehat{N}}, \tag{15}$$

which is the mean over the complete study region.

Large bias will result in domains within which the means deviate from the overall mean, i.e. in which it is not true that $\overline{y}_{U_d} \approx \overline{y}_U$.

The variance of the estimator will be very small, however.

Synthetic estimation

One possibility is to consider a larger region r that contains d rather than the complete study region and use

$$\widehat{\overline{y}}_d^{\text{syn,basic2}} = \frac{\widehat{t}_y(r)}{\widehat{N}(r)}, \tag{16}$$

which is approximately design unbiased if $\overline{y}_{U_d} \approx \overline{y}_{U(r)}$.

These estimators are not design consistent, though the MSE may be relatively small, if the regional sample size is large.

This is a very basic form of spatial smoothing.

Example: Smoking by county in Washington State

For (15), we would estimate the total number of smokers in Washington State, $\widehat{t}_y = \sum_{k \in s} w_k y_k$, and the total population size $\widehat{N} = \sum_{k \in s} w_k$.

For (16), we could split Washington State into (say) contiguous regions based on predictors of smoking.

For example, we could group together contiguous urban and rural counties (this categorization could be based on population density, or percent of farmland,...).

We would estimate the total number of smokers in region r, $\widehat{t}_y(r) = \sum_{k \in s(r)} w_k y_k$, where $s(r)$ is the set of indices of samples in region r and the total population size $\widehat{N}(r) = \sum_{k \in s(r)} w_k$.

Synthetic estimation

With auxiliary information consisting of known totals in domain d, $\mathbf{x}_d$, the synthetic estimator is

$$\widehat{t}_d^{\,\text{syn,reg}} = \mathbf{x}_d^{\mathsf{T}} \widehat{\mathbf{B}}, \tag{17}$$

where

$$\widehat{\mathbf{B}} = \left(\sum_{d=1}^{D} \sum_{k \in s_d} w_{dk} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{x}_{dk} \right)^{-1} \sum_{d=1}^{D} \sum_{k \in s_d} w_{dk} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{y}_{dk}, \tag{18}$$

is the WLS estimator over all of the units who provide responses, and w_{dk} are the design weights.

This estimator is not design unbiased.

Synthetic estimation

The design bias of $\widehat{t}_d^{\,\text{syn,reg}}$ is approximately $\mathbf{x}_d^{\mathsf{T}}\mathbf{B} - t_d$, where

$$\mathbf{B} = \left(\sum_{d=1}^{D} \sum_{k \in U_d} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{x}_{dk} \right)^{-1} \sum_{d=1}^{D} \sum_{k \in U_d} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{y}_{dk}, \tag{19}$$

is the population regression coefficient.

The bias will be small if the domain specific regression coefficient

$$\mathbf{B}_d = \left(\sum_{k \in U_d} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{x}_{dk} \right)^{-1} \sum_{k \in U_d} \mathbf{x}_{dk}^{\mathsf{T}} \mathbf{y}_{dk},$$

is close to $\mathbf{B}$.

Synthetic estimation

A special case is the ratio estimator

$$\widehat{t}_d^{\,\text{syn,rat}} = \widehat{t}_y \times \frac{t_{xd}}{\widehat{t}_x}. \tag{20}$$

Another special case is the post-stratification estimator

$$\widehat{t}_d^{\,\text{syn,ps}} = \sum_{g=1}^{G} \frac{N_{dg}}{\widehat{N}_{.g}} \widehat{t}_{.g}. \tag{21}$$

These estimators have low variance since information from all domains is used, but the bias may be large.

Notice that, in contrast to the direct GREG estimators described previously, these forms are adjusting a global response estimate, using domain specific auxiliary information.

Example: Smoking by county in Washington State

For (17), suppose we have domain-specific totals on income $\mathbf{x}_d = [1, t_{xd}]^\top$, along with individual income levels in the sample; the latter are used to estimate the population regression coefficient (18).

To examine whether $\mathbf{B}_d$ is close to $\mathbf{B}$ we could calculate

$$\widehat{\mathbf{B}}_d = \left(\sum_{k \in s_d} w_k \mathbf{x}_{dk}^\top \mathbf{x}_{dk} \right)^{-1} \sum_{k \in s_d} w_k \mathbf{x}_{dk}^\top \mathbf{y}_{dk},$$

and see how close these estimates are to $\widehat{\mathbf{B}}$ (though the former may have large uncertainty).

Example: Smoking by county in Washington State

For (20), we use the total incomes in domain d and the estimated income across the whole state $\widehat{t}_x = \sum_{k \in s} w_k x_k$, along with the estimated total smokers across the state $\widehat{t}_y = \sum_{k \in s} w_k y_k$.

For (21), we use the total population in domain d and stratum g, N_{dg} and the estimated stratum g population across the whole state $\widehat{N}_{.g} = \sum_{k \in s} w_k$, along with the estimated total stratum g population across the state $\widehat{t}_{.g} = \sum_{k \in s_g} w_k$.

Composite estimators

The direct estimator is approximately unbiased but will have large variance if n_d is small, while the synthetic estimator may have large bias but has small variance.

This suggests the composite estimator:

$$\overline{y}_d^{\text{comp}} = \phi_d \times \overline{y}_d^{\text{dir}} + (1 - \phi_d) \times \overline{y}_d^{\text{syn}}.$$

Rao and Molina (2015, Section 3.3) discusses how ϕ_d may be estimated, by attempting to minimize the MSE of $\overline{y}_d^{\text{comp}}$.

Next we will consider model-based approaches in which a formal method is used to balance using data from domain d, and the totality of data.

Model-based approaches

Indirect methods provide a link, often with an implicit model, between different areas; we now turn to explicit models that provide such a link.

The models we describe are hierarchical (or mixed-effects) models which aim to accurately describe between domain (area) differences.

Such models offer several advantages:

- Models can be tuned to the application, building on the existing theory and practical experience of mixed models, including non-linear models, such as logistic mixed models.
- Domain-(Area)-specific measures of uncertainty are produced.
- Estimates for areas with no data can be formed.
- One can attempt to check assumptions using diagnostics.
- A variety of area-specific random effects, including spatial, are available.

Model-based approaches

The use of explicit models has not been carried out greatly in survey sampling, where design-based inference is historically the norm.

The design consistency of model-based estimators is therefore of interest.

Disadvantages of mixed models:

- How to incorporate the design weights/acknowledge the design?
- It is often difficult to check modeling assumptions.
- Computation can be demanding, though this is improving.

Models can be specified at the level of the area or the unit

Fay-Herriot area-level model

Fay and Herriot (1979) introduced a model which has been highly influential.

Let $\theta_d = g(\overline{y}_{U_d})$ be a domain-specific quantity of interest, and $\mathbf{x}_d$ be a vector of domain-specific covariates.

Specify the linear model

$$\theta_d = \mathbf{x}_d^{\mathsf{T}}\boldsymbol{\beta} + U_d, \tag{22}$$

for $d = 1, \ldots, D$ with random effects $\mathrm{E}[U_d] = 0$ and $\mathrm{var}(U_d) = \sigma_u^2$.

Now assume we have direct estimators $\widehat{\theta}_d$ with associated design-based estimated variances $\widehat{V}_d$. Assume

$$\widehat{\theta}_d = \theta_d + \epsilon_d, \tag{23}$$

with $\mathrm{E}[\epsilon_d] = 0$ and $\mathrm{var}(\epsilon_d) = \widehat{V}_d$, i.e., $\widehat{\theta}_d \sim \mathrm{N}(0, \widehat{V}_d)$.

Fay-Herriot area-level model

Combining (22) and (23) gives the model:

$$\widehat{\theta}_d \quad = \quad \mathbf{x}_d^{\mathsf{T}}\boldsymbol{\beta} + U_d + \epsilon_d. \tag{24}$$

Often it is assumed that:

$$\begin{aligned} U_d|\sigma_v^2 &\sim_{iid} \mathsf{N}(0, \sigma_u^2) \\ \epsilon_d|\widehat{V}_d &\sim_{iid} \mathsf{N}(0, \widehat{V}_d). \end{aligned}$$

Example: Smoking by county in Washington State

To fit the FH model, we can take $g(\cdot)$ to be the logit transform of the proportion of smokers, so that $\widehat{\theta}_d$ is the design-weighted logit estimator in domain d and $\widehat{V}_d$ is the associated design variance.

We could take $\mathbf{x}_d = [1\ x_d]^\mathsf{T}$ where x_d is the median income in domain d.

Unit-level model

For a continuous response, y_{dk}, a conventional hierarchical model would typically have first stage:

$$y_{dk} = \mathbf{x}_{dk}^{\mathsf{T}}\boldsymbol{\beta} + U_d + \epsilon_{dk}$$

where

- $\mathbf{x}_{dk}$ are unit-level covariates, for example age, gender, race, with associated regression parameters $\boldsymbol{\beta}$.
- Area-specific random effects:

  $$U_d|\sigma_u^2 \sim_{iid} \mathrm{N}(0, \sigma_u^2),$$

  which forms the second stage of the model.
- Unit-level errors:

  $$\epsilon_{dk}|\sigma_\epsilon^2 \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2).$$

- The error terms ϵ_{dk}, U_d are assumed independent.

No mention of weights! The required assumption is that the selection probabilities in domain d do not depend on y_{dk}, but may depend on $\mathbf{x}_{dk}$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In P. B.N. and C. F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademia Kiadó, Budapest.

Alkema, L. and New, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction method. *Annals of Applied Statistics*. To appear.

Alkema, L. and You, D. (2012). Child mortality estimation: a comparison of un igme and ihme estimates of levels and trends in under-five mortality rates and deaths. *PLoS Medicine*, **9**(8), e1001288.

Alkema, L., New, J. R., Pedersen, J., You, D., *et al.* (2014). Child mortality estimation 2013: an overview of updates in estimation methods by the united nations inter-agency group for child mortality estimation. *PloS ONE*, **9**, e101112.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, **94**, 443–458.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, **35**, 439–460.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Congdon, P. and Lloyd, P. (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science*, **8**, 235–252.

Dwyer-Lindgren, L., Kakungu, F., Hangoma, P., Ng, M., Wang, H., Flaxman, A., Masiye, F., and Gakidou, E. (2014). Estimation of district-level under-5 mortality in Zambia using birth history data, 1980–2010. *Spatial and Spatio-temporal epidemiology*, **11**, 89–107.

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.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153–164.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Ghitza, Y. and Gelman, A. (2013). Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, **57**, 762–776.

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib and

G. Tutz, editors, *Statistical Modeling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 91–110. Physica-Verlag.

Kass, E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothesis and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**(431), 928–934.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys*. John Wiley and Sons, New York.

Lohr, S. (2010). *Sampling: Design and Analysis, Second Edition*. Brooks/Cole Cengage Learning, Boston.

Longford, N. (1996). Model-based variance estimation in surveys with stratified clustered design. *Australian Journal of Statistics*, **38**, 333–352.

Lozano, R., Wang, H., Foreman, K. J., Rajaratnam, J. K., Naghavi, M., Marcus, J. R., Dwyer-Lindgren, L., Lofgren, K. T., Phillips, D., Atkinson, C., *et al.* (2011). Progress towards millennium development goals 4 and 5 on maternal and child mortality: an updated systematic analysis. *The Lancet*, **378**, 1139–1165.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, **9**.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

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. (2016). Small area estimation of childhood of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*. To appear.

Pereira, L. N. and Coelho, P. (2010). Small area estimation of mean price of habitation transaction using timeseries and cross-sectional area-level models. *Journal of Applied Statistics*, **37**, 651–666.

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

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.

Potthoff, R., Woodbury, M., and Manton, K. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference

using survey weights under superpopulation models. *Journal of the American Statistical Association*, **87**, 383–396.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17**, 113–141.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A*, **169**, 805–827.

Rao, J. (2003). *Small Area Estimation*. John Wiley, New York.

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

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.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Singh, B., Shukla, G., and Kundu, D. (2005). Spatio-temporal models in small-area estimation. *Survey Methodology*, **31**, 183–195.

Skinner, C. (1989). Domain means, regression and multivariate analysis. In C. Skinner, D. Holt, and T. Smith, editors, *Analysis of Complex Surveys*, pages 59–87. Wiley, Chichester.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B*, **64**, 485–493.

United Nations Inter-Agency Group for Child Mortality Estimation (2011). Levels & trends in child mortality: Report 2012.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*. John Wiley.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology*, **38**, 330–336.

Zheng, H. and Little, R. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–17.

Zheng, H. and Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a

penalized spline nonparametric model. *Journal of Official Statistics*, **21**, 1–20.