Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 1-30

Bayesian Small Area Inference Using a Non-probability Sample

Balgobin Nandram
Department of Mathematical Sciences, Worcester Polytechnic Institute
100 Institute Road. Worcester, MA 01609

Received: 20 April 2024; Revised: 24 May 2024; Accepted: 28 May 2024

Abstract

We show how to use supplemental information from a small probability sample (ps) to do Bayesian predictive inference for finite population means of small areas using a relatively larger non-probability sample (nps). We focus on the most practical situation when there are common covariates in the nps and ps, where the nps has the study variable but no survey weights and the ps has survey weights but no study variable. We assume that the population model is correct and any functional relation between the study variable and the covariates is unspecified. Data preparation is necessary, and there are three steps, which are a double mass imputation, stratification of the population and allocation of the sample to the strata (domains), and creating a spatial structure to accommodate the covariates. Our main Bayesian analysis uses the conditional auto-regressive model, which helps to accommodate the covariates without incorporating them into the model, thereby avoiding a functional relation between the study variable and the covariates. However, the actual small areas are not part of the model, but we need to keep track of them, and the strata are modeled as the "small areas". Our procedure allows a small area (not a stratum) to participate in several strata, and this helps to mitigate over-shrinkage, which is common in small area models. Using an illustrative example on body mass index data, our method appears to work better than a standard method with a linear regression of the study variable on the covariates. Our new framework allows several extensions and it avoids an approximation used in survey design analysis.

Key words: BHF baseline model; Finite population mean; Gibbs sampler; Inverse probability weighting; Mass imputation; Robustness; Stratification; Surrogate samples.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

We have data from a non-probability sample, nps (1), and a probability sample, ps (2), relatively much smaller. The nps and ps have common covariates, $\underline{x}$, a p-vector including an intercept. The nps has the study variable (response), y, but no survey weights, W, and the

Corresponding Author: Balgobin Nandram
Email: balnan@wpi.edu & balgobin.nandram@usda.gov

ps has survey weights, W, but no study variable, y. We know the small areas (*e.g.,* counties) in the nps, but we do not need to know these small areas in the ps. The population size, N, may be unknown and the nonsampled covariates from the nps (or the ps) are unknown. These two quantities can be constructed from the ps sampled data. The population has $(\underline{x}_i, y_i), i = 1, \ldots, N$. Letting $y_{ij}, j = 1, \ldots, N_i, i = 1, \ldots, \ell$, denote the values of the study variable from the ℓ small areas, we want to make inference about the finite population mean of the i^{th} area,

$$\bar{Y}_i = \frac{1}{N_i} \sum_{i=1}^{N_i} y_i, i = 1, \ldots, \ell.$$

Again the N_i may be unknown, and these can be constructed from the ps.

Non-probability sampling has become important over the past decade. Many surveys are done without proper survey designs. Observational studies, which are very useful, have no probabilistic structure. A probability sample is the gold standard, but many data collection procedures are done by walking around a mall or standing at the door of a library. Clearly, these data are non-probability samples, and they are generally biased. On the other hand the response rates for many probability samples have declined to situations, where the sample can more reasonably be called a non-probability sample. Therefore, it is important to do research into the topic of non-probability sampling; indeed, this is a "hot" topic and it is difficult.

From the nps, we have $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, which are respectively the covariates and the study variable (response) but we do not have survey weights, $W_i, i = 1, \ldots, n_1$. From the ps, we have $(W_{2i}, \underline{x}_{2i}), i = 1, \ldots, n_2$, but not the study variable, $y_{2i}, i = 1, \ldots, n_2$. This is a fairly general and practical set up; see Li *et al.* (2020).

In our new approach, no participation (selection mechanism) model is needed. However, our standard assumption is that the population model for the study variable is correct and the sample model is derived from it. Therefore, robustness to its assumptions plays a key role in this paper.

In Table 1, we show different scenarios of massively missing data and what action can be taken, where all the data are missing for a specific variable, and mass imputation is needed; see Kim *et al.* (2021). For example, for the nps, all survey weights are missing.

Sakshaug *et al.* (2019), Wisniowski *et al.* (2020) and Salvatore *et al.* (2023) worked with Scenario 2a to provide a full Bayesian approach. They supplemented a probability sample with a much larger non-probability sample. They mainly studied superpopulation parameters, not finite populations. There is no penalty on the non-probability sample, except for Salvatore *et al.* (2023) where they used the idea of Nandram and Rao (2021, 2023) on discounting historical data for binary data via logistic regression (*e.g.*, see Ibrahim and Chen, 2002, for the power prior). Except for Wisniowski *et al.* (2020), who used survey weights as a covariate which is a bit controversial, the others did not do so. Finite population parameter estimation was not done, and inference was not required about small areas.

Chen *et al.* (2020), henceforth sometimes CLW, studied Scenario 1 with propensity scores. They presented a doubly robust inference with non-probability samples mainly within the design-based approach. They estimated propensity scores via logistic regression (parametric modeling for the participation variable) using the probability sample without

Table 1: Data scenarios with massively missing data

	nps (1)			ps (2)			
Scenarios	w	$\underline{x}$	y	w	$\underline{x}$	y	Action
1	–	+	+	+	+	–	Supplement nps by ps
2 a	–	+	+	+	+	+	Supplement ps by nps
2 b	–	+	+	+	+	+	Data integration
3	–	+	–	+	–	+	Supplement ps by nps

NOTE: Here "+" means observed and "–" means missing.

study variable, and inverse probability weighting (IPW) for the finite population mean.

Based on the method of Chen, Li and Wu (2020) to estimate propensity scores, Nandram and Rao (2021, 2023) studied Scenario 2b. They argued that the nps should be used to construct the prior for the parameters (the study variable is observed in both nps and ps), a less practical situation. In addition, thinking of the nps as historical data, the prior of the parameters, obtained from it, should be partially discounted. For Bayesians, normalized weighted density should be used to model the study variable. Nandram, Choi and Liu (2021) discussed some mixed analyzes.

Rafei, Elliot and Flannagan (2022) and Marellla (2023) used Scenario 2b. Rafei *et al.* (2022) used Bayesian additive regression trees (BART), which has its own problems. Marella (2023) used empirical likelihood, and the study variable for the ps is manufactured and assumed real.

There has also been some activities in small area estimation for non-probability samples. Beaumont (2020) used Scenario 2a and Beaumont and Rao (2021) used covariates from nps in ps (*e.g.*, Fay-Herriot model) in Scenario 3; Nandram and Rao (2024) used Scenario 2a with unit level data and used the method of CLW to estimate propensity scores for the nps; they also discussed Scenario 2b. Rao (2021) discussed many scenarios and how the ideas of probability samples can be used to study non-probability samples. See Elliott and Valliant (2017) for an earlier review, where they discussed quasi-randomization, used in CLW and others.

Following CLW, in our current work under Scenario (1) without using propensity scores, we do not have a participation model for the selection mechanism, and the non-Bayesian notion of double robustness is null and void; in our case only the model for the study variable (response) is needed. However, this situation may not be fully practical because even though a nps is available, we may still need to plan and field a small probability sample (an unnecessary burden). Instead it is possible to obtain the necessary information in our set up using web-scraping. For example, it is possible to obtain population sizes and

population total covariates from the web for many variables.

In this paper, we have a nps (1) and a ps (2) from California (NHANES III). Body mass index (BMI) is the study variable, and for adults the normal range of BMI is $[20, 25]$; see Nandram and Choi (2010) for more details about the survey design and a discussion about a much larger BMI data set. The population size is about 5000 times the sample size (.02% sampling). The nps has about 80% of the data and the ps has about 20%. Our data are of the form, $(W_{sj}, \underline{x}_{sj}, y_{sj}), s = 1, 2, j = 1, \ldots, n_s$, where W_{sj} are original survey weights, $\underline{x}_{sj}$, a p-vector of covariates including an intercept, and y_{sj} is the study variable; but W_{1j} and y_{2j} are unknown. It is worth noting some of the features of the data:

a. There are three covaariates :Age (20-90 years' old), race (white, non-white), sex (male, female) are covariates but interactions are not significant;

b. The data are partitioned into eight (8) counties (small areas), an area in the ps may not have data, and interest is on the finite population mean of each area (county);

c. The population is stratified and the data from the nps and ps are allocated to 56 strata (areas) with some discretization: age (20-24, 25-29,. . . , 85-90), race $(0, 1)$, sex $(0, 1)$, where other covariates are redacted or missing, and the partition of age is normally used at the National Center for Health Statistics;

d. The study variable (response) and covariates in the nps, $(\underline{x}, y)$, provide the spatial relation (both $\underline{x}$ and $\underline{y}$ are used);

e. A double mass imputation is used to get the survey weights in the nps (W, $\underline{x}$ and y are used), and the weights are trimmed, calibrated to the population size, and adjusted to the effective sample size.

Our procedure differs from all others in the literature. We stratify the population into distinct values of the covariates in the nps, each distinct covariate vector represents a stratum. Note that these strata are obtained using basic knowledge about the population. We may need to discretize some covariates; in survey sampling many of the covariates are usually discrete.

In our example on BMI, age, race and sex are covariates; age (20-90 years old), race (white, non-white) and sex (male, female). So there are $71 \times 2 \times 2 = 484$ possible strata. The nps and ps data are then allocated to the strata, of course, after the sample data are observed. Then some strata will be empty and some discretization and regular imputation needs to be done. The covariates and study variable from the nps will be used to obtain a spatial structure among the strata via an incidence matrix, V. Then, we construct Table 2 for BMI data with some discretization on age to get 56 strata and 8 counties to avoid sparseness and empty strata. These are the data we actually analyze. Note that an area may have data in different strata, and each stratum will have a different parameter in our models. That is, a small area (county) may have several parameters associated with it, and this helps to mitigate over shrinkage, which is very common in small area estimation. In Section 3, we will describe how to prepare Table 2 because our new Bayesian analysis is based on it.

As a summary, we are using the nps data to make inference about the finite population mean of each small area (county for the BMI data). We have data from the nps and some information from the ps that we used to supplement the nps. This paper has six sections, including the current one. In Section 2, we review the main ideas, relevant to our current work, Nandram and Rao (2021, 2023) and Chen *et al.* (2020). In Section 3, we show how to prepare the data for a robust analysis. Specifically, we discuss stratification of the population by distinct covariate values, mass imputation to construct the survey weights for the nps, and a spatial structure, which we use to replace any functional relation between the study variable and the covariates, and Bayesian predictive inference for the finite population mean of each area (county); see Table 2. In Section 4, we discuss the hierarchical Bayesian models, where robustness is mostly based on a two-component mixture model. These models are primarily based on the Scott-Smith model; see Scott and Smith (1969) and Nandram *et al.* (2011). This section also has detailed discussion of an example of body mass index (BMI) in parallel to the rest of the technical discussion. In Section 5, we discuss some improvements and possible extensions. In Section 6, we present concluding remarks.

Table 2: Structurally complete BMI data for nps with $G = 56$ strata and $\ell = 8$ counties

Stratum	nps	Size
$\underline{x}_1$	$(W_{1j}, y_{1j}), j = 1, \ldots, n_1$	N_1
.	.	.
.	.	.
$\underline{x}_g$	$(W_{gj}, y_{gj}), j = 1, \ldots, n_g$	N_g
.	.	.
.	.	.
$\underline{x}_G$	$(W_{Gj}, y_{Gj}), j = 1, \ldots, n_G$	N_G

NOTE: There are G distinct values of $\underline{x}_g, g = 1, \ldots, G$; $G = 56$ for the BMI data. In addition, for the spatial analysis, we have the **incidence matrix**, V, among the strata. Weights may be equal or unequal and adjusted weights are used in the sampling process (study variable). An actual area (county) may be represented in several strata. The normal range of BMI for adults should be $[20, 25]$.

2. Review and background information

In this section, a review of Nandram and Rao (2021, 2023, 2024) is presented. We also present the method of Chen, Li and Wu (2020) to construct the propensity scores, and

some general comments are made about this method for propensity scores to highlight and justify our new approach.

2.1. Review of Nandram and Rao (2021, 2023, 2024)

First, we describe the normalized weighted density. Nandram and Rao (2021, 2023, 2024) considered a single population, not sub-populations (*e.g.,* small areas), and they showed how to make inference for finite population mean. They assumed that sample data are available from a nps (1) and a ps (2).

We recall here that the original survey weights are $W_{si}, i = 1, \ldots, n_s, s = 1, 2$, and W_{si} is the number of units the i^{th} sampled individual represents in the finite population of size, N. Since we assume that both samples are drawn from the same population, $\sum_{i=1}^{n_s} W_{si} = N, s = 1, 2$.

Nandram and Rao (2021, 2023) defined adjusted weights,

$$w_{si} = \hat{n}_s \frac{W_{si}}{\sum_{j=1}^{n_s} W_{sj}}, i = 1, \ldots, n_s, s = 1, 2, \tag{1}$$

where the y_{si} are assumed to be independent, and the effective sample size is

$$\hat{n}_s = \frac{(\sum_{j=1}^{n_s} W_{sj})^2}{\sum_{j=1}^{n_s} W_{sj}^2}.$$

The adjusted weights in (1) is needed to construct the likelihood functions. They also assumed an estimator of the population total, based on the ps, is $\hat{N} = \sum_{i=1}^{n_2} W_{2i}$. This is actually a Horvitz-Thompson estimator, but this interpretation is not necessary. Also, a Horvitz-Thompson estimator of the total covariates, based on the ps, is

$$\widehat{\sum_{i=1}^{N} \underline{x}_{2i}} = \sum_{i=1}^{n_2} W_{2i} \underline{x}_{2i}.$$

Of course, these two estimators are based on inverse probability weighting (IPW), where the first one being natural.

In their approach, they assumed the population model, $f(y \mid \underline{x}, \underline{\theta})$, is correct, and so robustness is a serious consideration. The participation model (for selection indicators) must also be robust; in our new approach there is no participation model, a huge gain. For the sample distribution, they used the weighted density,

$$f(y \mid \underline{x}, \underline{\theta}, \underline{w}) = \frac{\{f(y \mid \underline{x}, \underline{\theta})\}^w}{\int \{f(y \mid \underline{x}, \underline{\theta})\}^w dy\}}. \tag{2}$$

In the Bayesian paradigm, the sampling process must also be proper. Because the normalization constant may be a function of $\underline{\theta}$, it must be included, and therefore, the general specification in (2) is needed.

Next, we describe Bayesian predictive inference. We have estimated the survey weights; we use Chen, Li and Wu (2020) to get nps weights. For the nps and ps we now have

$$(W_{si}, \underline{x}_{si}, y_{si}), i = 1, \ldots, \ell, s = 1, 2.$$

Therefore, the population model is

$$y_i \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_i'\underline{\beta}, \sigma^2), i = 1, \ldots, N.$$

The sample model is

$$f(y_i \mid \underline{\beta}, \sigma^2) \propto \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y_i - \underline{x}_i'\underline{\beta})^2\}\right]^{w_i}$$

and normalizing, we have

$$y_{1i} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_{1i}'\underline{\beta}, \frac{\sigma^2}{w_{1i}}), i = 1, \ldots, n.$$

The prior distribution is

$$\pi(\underline{\beta}, \sigma^2) \propto \sigma^{-2}.$$

Once the sample model is fit, we use the ps to guess the population size and covariate total and we use surrogate sampling (Nandram, 2007) to sample population (prediction).

The finite population mean is

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

and

$$\pi(\bar{Y} \mid \underline{y}_s) = \int f(\bar{Y} \mid \underline{\beta}, \sigma^2)\pi(\underline{\beta}, \sigma^2 \mid \underline{y}_s)\, d\underline{\beta}\, d\sigma^2.$$

For Bayesian predictive inference, Nandram and Rao (2021, 2023, 2024) used surrogate sampling,

$$\bar{Y} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{i=1}^{n_2} W_{2i}x_{2i}'}{\sum_{i=1}^{n_2} W_{2i}}\underline{\beta}, \frac{\sigma^2}{\sum_{i=1}^{n_2} W_{2i}}\right).$$

We do not split the finite population as $\bar{Y} = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = \frac{n}{N}$ is the sample fraction, $\bar{y}_s$ is the sample mean and $\bar{y}_{ns}$ is the nonsample mean, because both $\bar{y}_s$ and $\bar{y}_{ns}$ are corrupted (biased). We will use a similar procedure in our new method.

To perform Bayesian predictive inference for small areas, Nandram and Rao (2024) assumed

$$(W_{sij}, \underline{x}_{sij}, y_{sij}), j = 1, \ldots, n_{si}, i = 1, \ldots, \ell, s = 1, 2,$$

are available,

$$y_{ij} \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_{ij}'\underline{\beta} + \nu_i, \sigma^2), j = 1, \ldots, N_i, i = 1, \ldots, \ell$$

where $\Omega = (\underline{\nu}, \underline{\beta}, \sigma^2)$. We note that the population is too large to sample completely, population sizes and covariates are unknown. We use surrogate sampling (Nandram, 2007) again.

The finite population means are

$$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}, \quad \pi(\bar{Y}_i \mid \underline{y}_s) = \int f(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2)\pi(\nu_i, \underline{\beta}, \sigma^2, \mid \underline{y}_s)\, d\nu_i\, d\underline{\beta}\, d\sigma^2, i = 1, \ldots, \ell.$$

Again, Nandram and Rao (2024) used surrogate sampling for small areas.

1. For $i = 1, \ldots, \ell$,

$$\bar{Y}_i \mid \Omega \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{j=1}^{n_{2i}} W_{2ij} x'_{2ij}}{\sum_{j=1}^{n_{2i}} W_{2ij}} \underline{\beta} + \nu_i, \frac{\sigma^2}{\sum_{j=1}^{n_{2i}} W_{2ij}}\right); \bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s$$

2. For the h^{th} iterate from the Gibbs sampler,

$$\bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{j=1}^{n_{2i}} W_{2ij} x'_{2ij}}{\sum_{j=1}^{n_{2i}} W_{2ij}} \underline{\beta}^{(h)} + \nu_i^{(h)}, \frac{\sigma^{2(h)}}{\sum_{j=1}^{n_{2i}} W_{2ij}}\right).$$

We note that $\sum_{j=1}^{n_{2i}} W_{2ij}$ is very large, and in this case the variance can be very small with $\bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s$ essentially a point mass.

2.2. Propensity scores (Chen, Li and Wu, 2020)

We review the method of Chen *et al.* (2020), recall CLW, and we make some commentaries about their method; see also Wu (2022).

Denote the common covariates by $\underline{z}_i, i = 1, \ldots, N$, with nps, $\underline{z}_{1i}, i = 1, \ldots, n_1$, and ps, $\underline{z}_{2i}, i = 1, \ldots, n_2$. Again we note that nonsampled covariates are unknown.

Let $R_i, i = 1, \ldots, N$, where $R_i = 1$ if unit i is sampled and $R_i = 0$ if unit i is not sampled. With $r_i = 0, 1$, they used the parametric assumption,

$$\pi_i = P(R_i = 1 \mid \underline{z}_i) = \pi(\underline{z}_i; \underline{\theta}),$$

$$p(\underline{r} \mid \underline{\theta}) = \prod_{i=1}^{N} \{\pi(\underline{z}_i; \underline{\theta})\}^{r_i} \{1 - \pi(\underline{z}_i; \underline{\theta}\}^{1-r_i}$$

$$= \prod_{i=1}^{n} \frac{\pi(\underline{z}_i; \underline{\theta})}{1 - \pi(\underline{z}_i; \underline{\theta})} \prod_{i=1}^{N} \{1 - \pi(\underline{z}_i; \underline{\theta}\}$$

with independence over i.

Then, the population log-likelihood is

$$\ell(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{z}_{1i}; \underline{\theta})}{1 - \pi(\underline{z}_{1i}; \underline{\theta})}\right\} + \sum_{i=1}^{N} \log\{1 - \pi(\underline{z}_i; \underline{\theta})\},$$

and the pseudo-log-likelihood is

$$\ell_1(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{z}_{1i}; \underline{\theta})}{1 - \pi(\underline{z}_{1i}; \underline{\theta})}\right\} + \sum_{i=1}^{n_2} W_{2i} \log\{1 - \pi(\underline{z}_{2i}; \underline{\theta})\}.$$

The propensity scores for the nps are then $\pi(\underline{z}_{1i}; \hat{\underline{\theta}}), i = 1, \ldots, n_1$, where $\hat{\underline{\theta}}$ is the MLE of $\underline{\theta}$

Now, we make some comments about the CLW Approach.

1. Horvitz-Thompson estimator may be sub-optimal with survey weights because the ratio of the study variable and the selection probabilities may not be close to a constant;

2. Propensity scores depend only on $\underline{z}$ (common variables only), and there may be other important variables;

3. Propensity scores are not selection probabilities because the entire population should be taken into consideration, and a proper quasi-randomization (see Elliot and Valiant, 2017) cannot be executed;

4. Logistic regression is not robust against its assumptions (*e.g.*, the linearity assumption on logit);

5. They assume ignorable selection, but nonignorable selection is preferred;

6. There are difficulties in optimization (convergence) especially for small samples;

7. In the Bayesian paradigm, the uncertainty in the estimation of the propensity scores must be taken into consideration, and this is a difficult problem, but a bootstrap procedure shows a 50% increase in standard deviation;

8. Need more robust participation models, perhaps a mixture of several link functions (*e.g.*, t_4, t_8, *etc.*).

Our new method has the following features: (a) We avoid using propensity scores because they are not selection probabilities in the CLW method; (b) We avoid direct link between the study variable and covariates (robustness, spatial model); (c) We do not need non-sample covariates for prediction; (d) We avoid using the Horvitz-Thompson estimator for prediction because it is sub-optimal with survey weights; (e) We consider a robust population model in which a two-component mixture model is used to accommodate outliers and non-normality.

We are particularly interested in inference about the finite population mean of each small area. The small areas (counties) are not modeled in our procedure, but rather the strata are modeled as small areas. Therefore, the phrase "small areas" is used in two ways, one for the constructed strata and one for counties. Another approach for non-probability samples with small areas is given by Nandram and Rao (2023b), where the actual small areas are modeled directly, and inference is about finite population means and percentiles (*e.g.*, 85^{th} and 95^{th} percentiles for BMI are useful). Our new approach avoids many difficult problems, except one of them is to incorporate the uncertainty in the estimated propensity scores, but we are still working on this problem.

3. Data preparation

In this section, we show how to prepare the data to construct our procedures and methods. We show how to obtain survey weights in the nps, form the strata, allocate the sample units to the strata, and obtain the the neighborhood structure among the strata. As we stated already, the strata are formed by distinct covariates. This is done for the BMI data; see Table 2 for the data we analyze in this paper.

We note the following steps in our procedure.

a. A few strata may be empty, and we can donate one unit from a stratum with at least two units (with all variables) to a "nearby" empty stratum separately for the nps (1) and the ps (2). For the nps, there was one empty stratum and for the ps there were three empty strata when the BMI data is processed.

b. The $W_{1ij}, j = 1, \ldots, n_{1i}$, are unknown, but it is true that $\sum_j^{n_{1i}} W_{1ij} = \sum_j^{n_{2i}} W_{2ij} = N_i, i = 1, \ldots, G$.

c. Assume all W_{1ij} are the same with the same $\underline{x}_i$ (i^{th} stratum); under simple random sampling (SRS) without replacement $W_{1ij} = N_i/n_{1i}$ (multiple-level regression and post-stratification (MRP) uses equal weights).

d. We have also obtained unequal weights, $W_{1ij}, j = 1, \ldots, n_{1i}$, and this can be accomplished using a double (reverse) mass imputation.

e. All weights are trimmed to mitigate the effects of outlying weights. Outliers decrease the effective sample size. Trimmed weights are calibrated to the population size and adjusted to the effective sample size for modeling, but too much trimming can lead to a false sense of security (decreased variance but increased bias).

We use double mass imputation to provide a structurally complete nps data set; see Kim *et al.* 2021. In mass imputation with two data sets, an entire variable may be missing from one data set, and both data sets are used to impute the missing data. For nps, we have $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, and for the ps we have $(W_{2i}, \underline{x}_i), i = 1, \ldots, n_2$ (no intercept). The procedure is straightforward:

a. Use $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, to fill in $y_{2i}, i = 1, \ldots, n_2$. This is done using the Mahalanobis distance among the $\underline{x}_{1i}$ and $\underline{x}_{2i}$ via nearest neighbors.

b. Stack the y_{1i} under the $\underline{x}_{1i}$ to create a new vector, $\underline{x}^*_{1i}$. Similarly, stack y_{2i} under the $\underline{x}_{2i}$ to create a new vector, $\underline{x}^*_{2i}$.

c. Use $(W_{2i}, \underline{x}^*_{2i}), i = 1, \ldots, n_2$, to fill in $W_{1i}, i = 1, \ldots, n_1$. This is done using the Mahalanobis distance among the $\underline{x}^*_{1i}$ and $\underline{x}^*_{2i}$ via nearest neighbor again.

We compare two situations, which are equal weights within strata (different strata can have different weights) and unequal weights within strata (small areas can have different weights). Actually the situation of equal weights correspond to ignorable selection and the situation of unequal weights corresponds to the situation of non-ignorable selection. See Nandram and Choi (2010) and more recently Nandram (2022).

Next, we show how to use the nps to get the spatial structure, which is used as a surrogate of the covariates to avoid the specification of any functional relation between the study variable and the covariates. Recall robustness is very important because the population model is assumed to be correct.

We follow the steps below:

a. Use a surrogate for any functional relation between the study variable and the covariates (Lockwood 2023, PhD Dissertation). The spatial approach can provide this surrogate.

b. Use ordinary least squares to find the $G \times G$ incidence matrix, V, which has zeros every where except when two strata are neighbors. We use a distribution-free procedure, where we assume that

$$y_{gj} = \underline{x}_g' \underline{\beta} + \nu_g, j = 1, \ldots, n_g,$$
$$\hat{\nu}_g = \bar{y}_g - \underline{x}_g' \hat{\underline{\beta}}, g = 1, \ldots, G,$$
$$\bar{y}_g = n_g^{-1} \sum_{j=1}^{n_g} y_{gj}, \ \hat{\underline{\beta}} = \left(\sum_{g=1}^{G} n_g \underline{x}_g \underline{x}_g' \right)^{-1} \sum_{g=1}^{G} n_g \underline{x}_g \bar{y}_g.$$

The neighbors of stratum g are $\mathcal{N}_g = \{h : |\hat{\nu}_g - \hat{\nu}_h| < t_o\}, h = 1, \ldots, G$.

c. Choose t_o to make the Moran's I correlation coefficient strong.

d. Once the strata are obtained, we never need the covariates again. Specifically, we do not need to estimate the nonsampled covariates, a huge saving, and the uncertain Horvitz-Thompson estimator of the population total covariate is not needed.

We note that there are difficulties using the Mahalanobis distance to find the incidence matrix. The distinct covariates should not be used, and there is virtually no control over this procedure.

Finally, we show how to do Bayesian predictive inference for the finite population means of the small areas using surrogate sampling as described in the review. Again, the population model is

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}(\mu_g, \sigma^2), j = 1, \ldots, N_g, g = 1, \ldots, G.$$

Let $\mathcal{C}_{ig}, g = 1, \ldots, G, i = 1, \ldots, \ell$, denote the set of sampled units of the i^{th} area in g^{th} stratum. Define the sum of the weights associated with i^{th} area in the g^{th} stratum as

$$N_{ig} = \sum_{j \in \mathcal{C}_{ig}} W_{gj}.$$

Note that many of these N_{ig} are zeros. Then, for $N_{ig} > 0$,

$$\bar{Y}_{ig} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\mu_g, \frac{\sigma^2}{N_{ig}}\}$$

and

$$\bar{Y}_i = \frac{\sum_{g=1}^{G} N_{ig} \bar{Y}_{ig}}{\sum_{g=1}^{G} N_{ig}}, i = 1, \ldots, \ell.$$

For posterior inference, we have $(\mu_g^{(h)}, \sigma^{2(h)}), h = 1, \ldots, M, g = 1, \ldots, G$, from the sample model. Then, we have $\bar{Y}_i^{(h)}, h = 1, \ldots, M, i = 1, \ldots, \ell$.

4. Hierarchical Bayesian models and a numerical example

This is the main section of the paper in which we present the hierarchical Bayesian models to analyze the new data. In parallel, we present our numerical example on body mass index (BMI) data.

First, we fit a small area model with covariates. This is the basic model of Battese *et al.* (1988); henceforth the BHF model. A full Bayesian approach of the BHF model is given by Toto and Nandram (2011) and Molina *et al.* (2014). However, this model is notoriously non-robust to its assumptions (normality, linearity, outliers). We use the BHF model as the baseline model for comparison.

Second, we use the Scott-Smith model (Scott and Smith, 1969), which does not have covariates; this model was intended for cluster sampling, but nowadays we have been using it for small area estimation; see Nandram *et al.* (2011), henceforth SS model. Note that the SS model is a special case of the BHF model. For example, we will construct a spatial model to accommodate the covariates and we will add robustness through mostly a two-component mixture model (Chakraborty *et al.*, 2019, Goyal *et al.*, 2020) and one slightly different model with a stick-breaking prior (Ishwaran and James, 2001). So we are using the SS model to drop the linearity assumption of the study variable and the covariates, to robustify the assumption of normality, and accommodating outliers. Note all SS models have a spatial component to accommodate the covariates.

We will compare the models with (a) equal weights and (b) unequal weights using posterior inference about the finite population means by actual small areas (counties in the application on BMI). For comparison, we use posterior mean (PM), posterior standard deviation (PSD), numerical standard error (NSE), posterior coefficient of variation (PCV) and 95% highest posterior density interval (HPDI). All our data analyses will be done on the BMI data.

All computations were done on WPI's Solar Cluster. The Gibbs sampler was used for almost all models. The computations used more time (increasing complex models) as we go from Table 3 to Table 7 because longer runs are needed to ensure strong mixing in the Gibbs samplers. Apart from the data preparation, which took just a few minutes, the computational time for all models (equal & unequal weights) took just about one hour.

Baseline (BHF) model

In the BHF model we use the original 8 counties. The stratum table with the weights is reverted to the small areas (not strata). We note that BHF model is notoriously not robust to its assumptions.

The population model is

$$y_{ij} \mid \nu_i, \underline{\beta}, \sigma^2 \stackrel{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, \sigma^2\}, j = 1, \ldots, N_i, i = 1, \ldots, \ell.$$

The sample model is

$$y_{ij} \mid \nu_i, \underline{\beta} \stackrel{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{ij}}\}, j = 1, \ldots, n_i,$$

$$\nu_i \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(0, \frac{\rho}{1-\rho}\sigma^2), i = 1, \ldots, \ell,$$

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}. \tag{3}$$

The joint posterior density for the sample model described in (3) is proper if the design matrix is full rank and $\ell \geq 2$. It can be sampled using a random sampler; a noisy Gibbs sampler is not necessary.

Bayesian predictive inference is now standard using surrogate sampling,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\underline{\bar{X}}_i'\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}\},$$

$$\pi(\bar{Y}_i \mid \underline{y}_s) = \int f(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2)\pi(\nu_i, \underline{\beta}, \sigma^2, \mid \underline{y}_s)d\nu_i \; d\underline{\beta} \; d\sigma^2, i = 1, \ldots, \ell.$$

In Table 3 we compare the BHF models, with equal weights and unequal weights. We observe that the PMs are rougher, PSDs are larger and PMs are closer to 25 (but not as close as we want) under the model with unequal weights than the model with equal weights. Note that counties 4 & 8 are mostly different from the others. But can we do better on these three measures?

SS model with spatial effects

We next delete the covariates and replace them with the spatial covariance matrix. This is the SS model with spatial effects.

Now the population model is

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\mu_g, \sigma^2\}, j = 1, \ldots, N_g, g = 1, \ldots, G.$$

The sample model is

$$y_{gj} \mid \mu_g, \rho \overset{ind}{\sim} \text{Normal}\{\mu_g, \frac{\sigma^2}{w_{gj}}\}, j = 1, \ldots, n_g,$$

$$\mu_g \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(\theta, \frac{\rho}{1-\rho}\sigma^2), g = 1, \ldots, G,$$

$$\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}. \tag{4}$$

The model, described in (4) and the few lines above it, is the basic SS model, which can be fit using a random sampler.

However, we need to add the spatial structure and we use the simultaneous conditional auto-regressive (CAR) model (*e.g.*, He and Sun, 2000; Chung and Datta, 2022). The sample model is now given by

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\left\{\mu_g, \frac{\sigma^2}{w_{gj}}\right\}, j = 1, \ldots, n_g, g = 1, \ldots, G,$$

Table 3: BHF model is fit directly to the small areas

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.629	0.365	0.004	0.013	(26.905, 28.351)
2	138	28.277	28.025	0.398	0.004	0.014	(27.293, 28.847)
3	667	27.340	27.629	0.192	0.002	0.007	(27.248, 27.995)
4	133	25.980	27.051	0.405	0.004	0.015	(26.221, 27.814)
5	96	27.075	27.547	0.406	0.004	0.015	(26.767, 28.368)
6	119	27.313	27.318	0.395	0.004	0.014	(26.538, 28.108)
7	100	27.518	27.881	0.449	0.005	0.016	(27.002, 28.777)
8	137	26.698	27.470	0.370	0.004	0.013	(26.741, 28.205)
b. Unequal weights							
1	140	27.324	27.234	0.440	0.004	0.016	(26.359, 28.085)
2	138	28.277	28.594	0.436	0.005	0.015	(27.728, 29.442)
3	667	27.340	27.193	0.197	0.002	0.007	(26.821, 27.590)
4	133	25.980	25.901	0.463	0.005	0.018	(25.002, 26.794)
5	96	27.075	27.855	0.497	0.005	0.018	(26.838, 28.790)
6	119	27.313	27.372	0.457	0.005	0.017	(26.489, 28.302)
7	100	27.518	27.394	0.514	0.005	0.019	(26.443, 28.439)
8	137	26.698	26.266	0.460	0.005	0.017	(25.355, 27.136)

NOTE: $\bar{y}_1 = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}, i = 1, \ldots, \ell$, is the ordinary sample average. Posterior inference is based on 1,000 iterates that provide posterior mean (PM), posterior standard deviation (PSD), numerical standard error (NSE), posterior coefficient of variation (PCV), and 95% highest posterior density interval (HPDI). Only random samplers, no Gibbs samplers, are used.

$$\underline{\mu} \mid \theta, \rho, \psi, \sigma^2 \sim \text{Normal}\left\{\underline{j}\theta, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\right\},$$

($\underline{j}$ is a vector of ones),

$$\pi(\theta, \rho, \psi, \sigma^2) \propto \frac{1}{\sigma^2},$$

$$0 < \rho < 1, \frac{1}{\lambda_1} < \psi < \frac{1}{\lambda_G},$$

where $\lambda_1 < \ldots < \lambda_G$, are eigenvalues of $R^{-1}V$, and $R = \text{diagonal}(r_g, g = 1, \ldots, G)$ with r_g the g^{th} row (column) sum of V. We note that the joint posterior density is proper and it can be fit using the Gibbs sampler. This is our first SS model.

In Table 4 we notice that the PSDs for some of the areas are smaller under unequal weights (not desirable). The PMs are smoother under the unequal weights, but closer to 25.

We need to improve this model, which can be done by robustification of either the sampling process or the area means or both. All these models require the use of the Gibbs sampler. Details of number of iterations used are described in the notes of the tables. To come up with those numbers, we have used the Geweke test of stationarity and the effective sample, and in all cases the Gibbs sampler mixed strongly.

Table 4: Scott-Smith model is fit to the strata with spatial effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.197	0.314	0.009	0.012	(26.638, 27.832)
2	138	28.277	26.741	0.258	0.008	0.010	(26.233, 27.276)
3	667	27.340	27.104	0.204	0.007	0.008	(26.709, 27.485)
4	133	25.980	27.050	0.269	0.010	0.010	(26.602, 27.591)
5	96	27.075	26.829	0.300	0.009	0.011	(26.256, 27.435)
6	119	27.313	27.082	0.356	0.012	0.013	(26.378, 27.812)
7	100	27.518	27.125	0.411	0.013	0.015	(26.358, 27.867)
8	137	26.698	27.076	0.279	0.008	0.010	(26.592, 27.632)
b. Unequal weights							
1	140	27.324	26.640	0.299	0.008	0.011	(26.113, 27.229)
2	138	28.277	26.730	0.253	0.006	0.009	(26.232, 27.208)
3	667	27.340	26.546	0.178	0.005	0.007	(26.225, 26.895)
4	133	25.980	26.687	0.424	0.016	0.016	(25.831, 27.375)
5	96	27.075	26.149	0.310	0.008	0.012	(25.575, 26.758)
6	119	27.313	26.412	0.293	0.010	0.011	(25.779, 26.954)
7	100	27.518	25.877	0.372	0.012	0.014	(25.257, 26.632)
8	137	26.698	25.951	0.289	0.010	0.011	(25.399, 26.538)

NOTE: The Gibbs sampler is run $11,000$ times with a "burn-in" of $1,000$ and a systematic sample of every tenth is taken.

SS model with spatial effects and robust study variable

The population model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma < 1$.

The sample model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}),$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$

$$\underline{\mu} \overset{ind}{\sim} \text{Normal}\{\theta \underline{j}, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\},$$

$$\pi(\theta, \sigma^2, p, \rho, \gamma, \psi) \propto \frac{1}{\sigma^2}, \frac{1}{\lambda_1} < \psi < \frac{1}{\lambda_G}.$$

The joint posterior density is

$$\pi(\underline{z}, \underline{\mu}, \theta, \sigma^2, \gamma, \rho, \psi \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2} \prod_{g=1}^{G} \prod_{j=1}^{n_g} [(1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}})]^{1-z_{gj}} [p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}})]^{z_{gj}}$$

$$\times \text{Normal}\{\theta \underline{j}, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\},$$

and this posterior can be sampled using the Gibbs sampler; see note to Table 5. The results are better than those in BHF model and the SS model with only spatial effects.

SS Model with spatial effects, robustness on study variable and random effects

The population model is now

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma_0 < 1$.

The sample model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{5}$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$,

$$\mu_g \overset{ind}{\sim} (1-q)\text{Normal}(\nu_g, \gamma_1\frac{\rho_1}{1-\rho_1}\sigma^2) + q\text{Normal}(\nu_g, \frac{\rho_1}{1-\rho_1}\sigma^2), \tag{6}$$

$0 < q < 1/2, 0 < \gamma_1 < 1$,

$$\underline{\nu} \sim \text{Normal}\{\theta \underline{j}, \frac{\rho_2}{1-\rho_2}\sigma^2(R - \psi V)^{-1}\},$$

$$\pi(\theta, \sigma^2, p, q, \rho_1, \rho_2, \psi, \gamma_0, \gamma_1) \propto \frac{1}{\sigma^2}.$$

The assumptions in (5) and (6) express a form of Bayesian double robustness.

The joint posterior density is

$$\pi(\underline{z}, \underline{t}, \underline{\mu}, \underline{\nu}, \theta, \sigma^2, p, q, \rho_1, \rho_2, \psi, \gamma_0, \gamma_1 \mid \underline{y}) \propto$$

Table 5: Scott-Smith model is fit to the strata with spatial effects and robust study variable

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.257	0.309	0.008	0.011	(26.701, 27.839)
2	138	28.277	27.100	0.248	0.010	0.009	(26.579, 27.541)
3	667	27.340	27.315	0.202	0.006	0.007	(26.969, 27.667)
4	133	25.980	27.206	0.252	0.007	0.009	(26.715, 27.674)
5	96	27.075	26.956	0.358	0.011	0.013	(26.341, 27.538)
6	119	27.313	27.286	0.233	0.008	0.009	(26.821, 27.712)
7	100	27.518	27.040	0.425	0.014	0.016	(26.251, 27.731)
8	137	26.698	27.235	0.238	0.008	0.009	(26.838, 27.628)
b. Unequal weights							
1	140	27.324	27.021	0.345	0.011	0.013	(26.349, 27.668)
2	138	28.277	27.084	0.311	0.009	0.011	(26.464, 27.584)
3	667	27.340	26.835	0.186	0.005	0.007	(26.493, 27.203)
4	133	25.980	26.752	0.381	0.011	0.014	(26.095, 27.410)
5	96	27.075	26.503	0.367	0.011	0.014	(25.784, 27.121)
6	119	27.313	27.003	0.341	0.011	0.013	(26.340, 27.619)
7	100	27.518	26.163	0.457	0.016	0.017	(25.311, 26.969)
8	137	26.698	26.417	0.281	0.010	0.011	(25.823, 26.902)

NOTE: The Gibbs sampler is run $40,000$ times with a "burn-in" of $10,000$ and a systematic sample of every thirtieth.

$$\frac{1}{\sigma^2}\prod_{g=1}^{G}\prod_{j=1}^{n_g}[(1-p)\text{Normal}_{y_{gj}}(\mu_g,\gamma_0\frac{\sigma^2}{w_{gj}})]^{1-z_{gj}}[p\text{Normal}_{y_{gj}}(\mu_g,\frac{\sigma^2}{w_{gj}})]^{z_{gj}}$$
$$\times\prod_{g=1}^{G}[(1-q)\text{Normal}_{\mu_g}(\nu_g,\gamma_1\frac{\rho_1}{1-\rho_1}\sigma^2)]^{1-t_g}[q\text{Normal}_{\mu_g}(\nu_g,\frac{\rho_1}{1-\rho_1}\sigma^2)]^{t_g}$$
$$\times\left(\frac{1-\rho_2}{\rho_2\sigma^2}\right)^{G/2}|R-\psi V|^{1/2}\exp\left\{-\frac{1-\rho_2}{2\rho_2\sigma^2}(\underline{\nu}-\theta\underline{j})'(R-\psi V)(\underline{\nu}-\theta\underline{j})\right\},$$

and this can be sampled using the Gibbs sampler; see note to Table 6. Again the results look better than the previous ones. This model appears to be the best: The PSDs for unequal weights are larger than those for equal weights, and the PMs for unequal weights are smaller than those for equal weights (therefore closer 25).

Table 6: Scott-Smith model is fit to the strata with spatial effects and robust study variable and robust random effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	26.826	0.268	0.008	0.010	(26.283, 27.321)
2	138	28.277	26.798	0.270	0.007	0.010	(26.305, 27.324)
3	667	27.340	27.096	0.221	0.006	0.008	(26.679, 27.523)
4	133	25.980	27.166	0.198	0.006	0.007	(26.778, 27.550)
5	96	27.075	26.722	0.239	0.007	0.009	(26.280, 27.172)
6	119	27.313	26.997	0.238	0.006	0.009	(26.539, 27.451)
7	100	27.518	26.814	0.227	0.007	0.008	(26.379, 27.238)
8	137	26.698	27.185	0.202	0.006	0.007	(26.815, 27.597)
b. Unequal weights							
1	140	27.324	26.228	0.355	0.010	0.014	(25.479, 26.839)
2	138	28.277	26.523	0.333	0.010	0.013	(25.889, 27.167)
3	667	27.340	26.367	0.270	0.008	0.010	(25.851, 26.899)
4	133	25.980	26.460	0.236	0.007	0.009	(26.000, 26.915)
5	96	27.075	25.949	0.283	0.008	0.011	(25.389, 26.477)
6	119	27.313	26.432	0.317	0.010	0.012	(25.746, 26.953)
7	100	27.518	25.917	0.274	0.008	0.011	(25.376, 26.412)
8	137	26.698	26.171	0.245	0.008	0.009	(25.709, 26.653)

NOTE: The Gibbs sampler is run $60,000$ times with a "burn-in" of $15,000$ and a systematic sample of every forty-fifth.

SS model with spatial effects, robust study variable and stick-breaking priors on random effects

The population model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma_0 < 1$.

The sampling model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{7}$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$,

$$\underline{\mu} \sim \text{Normal}\{\theta\underline{j} + \underline{\eta}, \frac{\rho_1}{1-\rho_1}\sigma^2(R - \psi V)^{-1}\},$$

where $\underline{j}$ is a vector of ones, and the Pitman-Yor two-parameter process is

$$\eta_g \mid \underline{t} \overset{ind}{\sim} \sum_{s=1}^{G_0} p_s \text{Normal}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2), G_0 \leq G, g = 1, \ldots, G, \tag{8}$$

$$p_1 = \nu_1, p_2 = \nu_2(1-\nu_1), \ldots, p_{G_0} = \prod_{s=1}^{G_0-1}(1-\nu_s),$$

$$\nu_s \mid \delta_1, \delta_2 \overset{ind}{\sim} \text{Beta}\{1-\delta_1, \frac{1-\delta_2}{\delta_2} + (s-1)\delta_1\}, s = 1, \ldots, G_0,$$

$$t_s \overset{ind}{\sim} \text{Normal}(0, \frac{\rho_3}{1-\rho_3}\sigma^2), s = 1, \ldots, G_0.$$

The assumptions in (7) and (8) are a form of Bayesian double robustness; these are more flexible than the assumptions in (5) and (6). We have used the prior,

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}, \theta \sim \text{Normal}(\theta_o, \sigma_o^2),$$

where θ_o and σ_o^2 must be specified. Also, for computational stability, we have kept $0 < \delta_1 < \frac{1}{2} < \delta_2 < 1$ with uniform priors on δ_1, δ_2, $\gamma_o, p, \psi, \rho_1, \rho_2, \rho_3$.

In Table 7, this model is similar to the model in which there is robustness on both the study variable and the area effects. But there are some aberrations as some of the PSDs under the stick-breaking prior are smaller with unequal weights, which is a bit concerning.

As a summary, we compare all the models in Table 8. We use the summaries,

$$RD = \{1 - \frac{AVG_{UW} - 25}{AVG_{EW} - 25}\}100\%, \quad RP = \{\frac{GM_{UW}}{GM_{EW}} - 1\}100\%, \tag{9}$$

where AVG_{UW} and AVG_{EW} are the arithmetic means of the 8 PMs for respectively the unequal weights (UW) case and the equal weights (EW) case, and GM_{UW} and GM_{EW} are the geometric means of the 8 PSDs for respectively the unequal weights (UW) case and the equal weights (EW) case. Here RD is the percent AVG_{UW} is closer to 25 than AVG_{EW} (expected to be positive), and RP is the percent increase of GM_{UW} over GM_{EW} (expected to be positive).

Under the RD measure, three models stand out with robust study variable. Under the RP measure, three models stand out. For both measures the model that wins is the model in which study variable and the area effects are both robust; indeed, this is a novel model. While the stick-breaking of the area effects is robust, it is a bit concerning that RP is negative, but it is possible to overcome this problem. Detailed results, like the other models, were not shown for the first and second SS model in Table 8. These results show that robustness on the study variable is important, because for the first SS model, without robustness on the study variable, there are artificially low PSDs relative to the BHF model and the second SS model with robust study variable.

Table 7: Scott-Smith model is fit to the strata with spatial effects and robust study variable and stick-breaking priors on random effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.286	0.347	0.012	0.013	(26.708, 28.092)
2	138	28.277	26.674	0.330	0.010	0.012	(25.989, 27.256)
3	667	27.340	27.172	0.214	0.006	0.008	(26.762, 27.580)
4	133	25.980	27.105	0.303	0.009	0.011	(26.573, 27.689)
5	96	27.075	26.880	0.367	0.010	0.014	(26.152, 27.533)
6	119	27.313	27.226	0.575	0.021	0.021	(26.194, 28.372)
7	100	27.518	27.174	0.378	0.010	0.014	(26.403, 27.824)
8	137	26.698	27.133	0.311	0.009	0.011	(26.534, 27.739)
b. Unequal weights							
1	140	27.324	26.738	0.347	0.009	0.013	(26.144, 27.448)
2	138	28.277	26.738	0.282	0.008	0.011	(26.193, 27.296)
3	667	27.340	26.581	0.201	0.006	0.008	(26.213, 26.943)
4	133	25.980	26.715	0.424	0.012	0.016	(25.872, 27.429)
5	96	27.075	26.076	0.310	0.011	0.012	(25.522, 26.747)
6	119	27.313	26.453	0.410	0.011	0.016	(25.642, 27.223)
7	100	27.518	25.856	0.372	0.009	0.014	(25.189, 26.543)
8	137	26.698	25.914	0.324	0.009	0.013	(25.283, 26.503)

NOTE: The Gibbs sampler is run $100,000$ times with a "burn-in" of $25,000$ with a systematic sample of every seventy-fifth, and it took twenty-five minutes.

5. Improvements and extensions

In this section, we show what improvements can be made to our new procedure and possible extensions. We also show how to make inference about all propensity scores, not just those associated with the non-probability sample, but its non-sampled part of the population as well.

5.1. Improvements, Mahalanobis distance

We discuss how to replace the Mahalanobis distance because it is not appropriate with discrete variables. Leon and Carriere (2005) introduced a generalized Mahalanobis distance for mixed data but this method is cumbersome; so we seek a simpler mathod that avoids the Mahalanobis distance completely. Basically we use the nps (1) and the ps (2) to get propensity scores, and matching to get surrogates for $y_{2j}, j = 1, \ldots, n_2$. Then, we use the nps (1) and ps (2) with the surrogates, to get propensity scores again, and matching to

Table 8: Comparison of the six models by summaries of the PMs and PSDs over the eight counties

	PM			PSD		
Models	EW	UW	RD (%)	EW	UW	RP (%)
Covariates (BHF)	27.569	27.226	15.391	0.363	0.418	15.295
NR, NS-RE (SS)*	27.540	27.174	16.840	0.143	0.152	5.844
RS, NS-RE (SS)*	27.186	26.732	26.243	0.265	0.329	24.381
RS, S-RE (SS)	27.034	26.372	48.264	0.294	0.298	1.200
RS & S-R-RE (SS)	26.917	26.215	57.703	0.233	0.284	22.009
RS & S-SB-RE (SS)	27.081	26.384	50.402	0.341	0.326	-4.423

NOTE: Scott-Smith (SS) spatial models have the CAR prior and the non-spatial models* replace $R - \psi V$ by the identity matrix. The summaries are respectively the arithmetic mean of the eight PMs and geometric mean of the eight PSDs. The first three models are non-spatial (NS), and last four models have robust study variable (RS).

get surrogates for $W_{1j}, j = 1, \ldots, n_1$. The y_{2j} are not used for further analysis.

Our procedure has two steps. First, we massively impute $y_{2i}, i = 1, \ldots, n_2$, using propensity scores (i.e., matching via nearest neighbors). Second, we match (nearest neighbor) propensity scores conditional on the nps and ps data now available.

In the first step, we define $I_i = 0$ if $i = 1, \ldots, n_1$ (nps) and $I_i = 1$ if $i = n_1+1, \ldots, n_1+n_2 = n$ (ps); $\underline{x}_i = \underline{x}_{1i}, i = 1, \ldots, n_1$ and $\underline{x}_{n_1+i} = \underline{x}_{2i}, i = 1, \ldots, n_2$. We then assume logistic regression model,

$$I_i \overset{ind}{\sim} \text{Bernoulli}(\frac{e^{\underline{x}_i'\underline{\beta}}}{1 + e^{\underline{x}_i'\underline{\beta}}}), i = 1, \ldots, n = n_1 + n_2.$$

Let $\hat{\underline{\beta}}$ denote the maximum likelihood estimator of $\underline{\beta}$, we have propensity scores,

$$\pi_i = \frac{e^{\underline{x}_i'\hat{\underline{\beta}}}}{1 + e^{\underline{x}_i'\hat{\underline{\beta}}}}, i = 1, \ldots, n.$$

We fill in the missing $y_{2i}, i = 1, \ldots, n_2$, using matching on the π_i. For each $i = n_1+1, \ldots, n_1+n_2$, we find which $j, j = 1, \ldots, n_1$, minimizes $|\pi_i - \pi_j|$, say j^*; j^* may not be unique. Then, the value of the study variable given to unit i is $y_{2j^*}, i = n_1 + 1, \ldots, n_1 + n_2$.

In the second step, we define $\underline{x}_i = \underline{x}_{1i}, i = 1, \ldots, n_1, \underline{x}_i = \underline{x}_{2i}, i = 1, \ldots, n_2$. Similarly, we define $y_i = y_{1i}, i = 1, \ldots, n_1, y_i = y_{2i}, i = 1, \ldots, n_2$. Also, define $I_i = 1, i = 1, \ldots, n_1$, for the nps (1) and $I_i = 0, i = n_1 + 1, \ldots, n_1 + n_2 = n$, for the ps (2). Note that $\underline{x}_i$ has p components, including an intercept. For the nonignorable model, we assume logistic regression,

$$I_i \mid \underline{\beta}, y_i \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{(\underline{x}_i'\underline{\beta}_{(p)} + y_i\beta_{p+1})}}{1 + e^{(\underline{x}_i'\underline{\beta}_{(p)} + y_i\beta_{p+1})}}\right\}, i = 1, \ldots, n.$$

We optimize the likelihood function to obtain the maximum likelihood estimator of $\underline{\beta}$, which we now denote by $\hat{\underline{\beta}}$. The propensity scores are then

$$\pi_i = \frac{e^{\underline{x}_i'\hat{\underline{\beta}}+y_i\hat{\beta}_{p+1}}}{1+e^{\underline{x}_i'\hat{\underline{\beta}}+y_i\hat{\beta}_{p+1}}}, i = 1, \ldots, n.$$

For each $i = 1, \ldots, n_1$, we find which $j, j = n_1 + 1, \ldots, n$, minimizes $|\pi_i - \pi_j|$, say j^*; j^* may not be unique. Then, the weight given to unit i is $W_{2j^*}, i = 1, \ldots, n_1$. Denote these weights by $W_{1i}, i = 1, \ldots, n_1$. Letting $N = \sum_{i=1}^{n_2} W_{2i}$, the design estimate of the population size, our final weights for the nps (1) are

$$W_{1i} \equiv N\frac{W_{1i}}{\sum_{i=1}^{n_1} W_{1i}}, i = 1, \ldots, n_1.$$

Note that the y_{2i} are discarded and are not used in any further analysis.

5.2. Extensions

Unfortunately, in our new procedure sample sizes and the sub-population sizes of the strata (domains) are random variables. The uncertainty in their values should be taken into consideration.

Let $\underline{j} = (1, \ldots, 1)'$, a vector of ones, and $\underline{O} = (0, \ldots, 0)'$, a vector of zeros, $N_g >> n_g, g = 1, \ldots, G$ denote G-vectors.

a. Population sizes

Letting $\underline{N} = (N_1, \ldots, N_G)'$, domain sizes from ps (1), we assume

$$\underline{N} - \underline{j} \sim \text{Multinomial}(N - G, \underline{P}), \quad \underline{P} \sim \text{Dirichlet}(\underline{O}),$$

$$\underline{P} \mid \underline{N} \sim \text{Dirichlet}(\underline{N} - \underline{j}), \quad \underline{T} - \underline{j} \mid \underline{P}, \underline{N} \sim \text{Multinomial}(N - G, \underline{P})$$

b. Sample sizes

Letting $\underline{n} = (n_1, \ldots, n_G)'$, observed domain sizes, we assume

$$\underline{n} - \underline{j} \sim \text{Multinomial}(n - G, \underline{p}), \quad \underline{p} \sim \text{Dirichlet}(\underline{O}),$$

$$\underline{p} \mid \underline{n} \sim \text{Dirichlet}(\underline{n} - \underline{J}), \quad \underline{t} - \underline{j} \mid \underline{p}, \underline{n} \sim \text{Multinomial}(n - G, \underline{p}).$$

c. Posterior inference

Using Bayes' theorem, the joint posterior density is

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}),$$

where $\underline{z}$ is a vector of latent variables; $z_{gj}, j = 1, \ldots, t_g, g = 1, \ldots, G$. There are three cases, which we must consider,

a. If $t_g < n_g$, take a simple random sample without replacement from $y_{gj}, j = 1, \ldots, n_g$ (carry on W_{gj});

b. If $t_g = n_g$, retain $(y_{gj}, W_{gj}), j = 1, \ldots, n_g$;

c. If $t_g > n_g$, take all $(y_{gj}, W_{gj}), j = 1, \ldots, n_g$ and draw a simple random sample with replacement from them to get the others.

We use the following decomposition of the joint posterior density,

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}) =$$

$$\pi_1(\Omega \mid \underline{z}, \underline{t}, \underline{T}, \underline{y}, \underline{W}, \underline{n}, \underline{N})\ \pi_2(\underline{z} \mid \underline{t}, \underline{T}, \underline{y}, \underline{W}, \underline{n}, \underline{N})\ \pi_3(\underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}).$$

We assume

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}) =$$

$$\pi_1(\Omega \mid \underline{z}, \underline{t}, \underline{y}, \underline{W}, \underline{n})\ \pi_2(\underline{z} \mid \underline{t}, \underline{y}, \underline{W}, \underline{n})\ \pi_{30}(\underline{t} \mid \underline{n})\ \pi_{31}(\underline{T} \mid \underline{N}).$$

5.3. Inference for all propensity scores

As usual, there are two cases, ignorable selection and non-ignorable selection. For ignorable selection in our new procedure, logistic regression is null and void. Basically, we have a simple random sample without replacement from each stratum. So the selection probabilities for the g^{th} stratum are $n_g/N_g, j = 1, \ldots, N_g$; obviously these vary with the sizes of the strata/domains. It does not affect the modeling of the study variable. It is trivial to deal with random sample sizes of the strata/domains; see Section 5.2. Now the method of Chen, Li and Wu (2020) is completely useless.

For nonignorable selection, we do not need the Horvitz-Thompson estimator (pseudo-likelihood). However, we will obtain the propensity scores for the entire population of N individuals, and therefore these propensity scores can be interpreted as selection probabilities. We use the following logistic regression model,

$$\prod_{g=1}^{G} \left[\prod_{j=1}^{n_g} \frac{e^{\underline{x}_g'\underline{\beta} + y_{gj}\beta_{p+1}}}{1 + e^{\underline{x}_g'\underline{\beta} y_{gj}\beta_{p+1}}} \left\{ \frac{1}{e^{\underline{x}_g'\underline{\beta} + y_{gj}\beta_{p+1}}} \right\}^{N_g/n_g - 1} \right].$$

Note that we are assuming each individual in the g^{th} stratum in the sample is reproduced N_g/n_g with the same value of the study variable. Maximum likelihood estimators, $\hat{\underline{\beta}}$, can be obtained for $\underline{\beta}$. Hence, the propensity scores, π_{gs}, are given by

$$\pi_{gs} = \frac{e^{\underline{x}_g'\hat{\underline{\beta}} + y_{gs}\hat{\beta}_{p+1}}}{1 + e^{\underline{x}_g'\hat{\underline{\beta}} y_{gs}\hat{\beta}_{p+1}}}, s = (j-1)\frac{N_g}{n_g} + 1, \ldots, j\frac{N_g}{n_g},\ j = 1, \ldots, n_g.$$

Some comments are in order.

(a) It is possible to provide a full Bayesian method to obtain the propensity scores.

(b) The π_{gs} are very variable, and they will not add up to n, and we can rake them up to n. String out the π_{gs} as $\pi_i, i = 1, \ldots, N$.

(c) The raking procedure can be a bit problematic because $0 < \pi_i < 1, i = 1, \ldots, N$ and $\sum_{i=1}^{N} \pi_i = n$ for sampling without replacement. Of course, if $\sum_{i=1}^{n} \pi_i \geq n$, raking does not cause any problems. If $\sum_{i=1}^{n} \pi_i < n$, there are difficulties because raking up will make some of the raked $\pi_i > 1$, which we do not want. Keep the π_i in the fourth quartile unchanged. Suppose $\sum_{i \in Q_4} \pi_i = n_o$, where Q_4 is the fourth quartile; then rake up the π_i in the first three quartiles to $n - n_o$.

(d) It is not much more difficult to make inference about the study variable with random sample sizes of the strata/domains.

6. Concluding remarks

A structurally complete probability sample is obtained from the non-probability sample using double mass imputation with supplemental data from a relatively small probability sample (no study variable). The population is stratified by distinct covariates and the nps and ps are allocated to the strata. The study variable and the covariates are used to construct an incidence matrix (spatial structure), which is used to accommodate the covariates. The covariates are never used in the proposed models.

We have used the Scott-Smith model to avoid specifying the uncertain relationship between the study variable and covariates for unit level data without consideration of the participation variable. Robust models are specified for both the study variable and the random effects, and in one model, the Pitman-Yor two-parameter stick-breaking process is used. This is needed because the population model, used for prediction, is assumed to be correct. The nps data are used to construct an incidence matrix with the neighboring strata.

The small areas are not modeled directly, rather the model is placed on the strata with the non-probability samples. Inference about the actual small areas is obtained in the output analysis. This helps with over-shrinkage.

A simulation study to assess the predictive power of the proposed models will be useful. Sensitivity of Bayesian predictive inference of the finite population mean to the size of the ps sample needs to be investigated. Further robustification can be done using the stick-breaking process for the sampling process (study variable) instead of the two component mixture model for the study variable in the Scott-Smith model. An alternative approach, using structural error models, is presented by Nandram (2023).

The method is promising, and the non-Bayesian notion of double robustness is null and void. The probability sample plays a fairly minor role, and it can be eliminated if the required population information can be obtained using web-scraping, an emerging science. There is a good comparison with the BHF (baseline) model. Future work will be focused in this direction (stratification and matching).

At one of my talks, a participant asked if the means of the two components in mixture model can be different. Of course, they can be different; the mixture model in the stick-breaking process has a different mean for each cluster. The model of Goyal *et al.* (2020) is limited, and so in Appendix B, we describe how one might use different means for a single area; thereby adding flexibility to this two-component mixture model.

Acknowledgements

This work was presented as invited lectures at the IISA 2023 Conference, Colorado, USA, at the ISI 2023 Conference, Ottawa, Canada, and in India, Febrauary 2024, as a Special Invited Lecture at the SSCA Conference, Banasthali Vidyapith, Rajasthan, and in March 2024, at the Indian Institute of Management (IIM), Ahmedabad, as a more extended lecture. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). Balgobin also thanks five of his PhD students (Dr. Ashley Lockwood, Dr. Lingli Yang, Dr. Dilli Bhatta, Yang Liu and Zihang Xu) for a final reading of the paper.

Appendices

Appendix A: Extension of the Scott-Smith model with stick breaking process

We extend the Scott-Smith model to have a robust model on the study variable, spatial effects and random effects, where the random effects have a stick-breaking prior (Ishwaran and James, 2001).

For $g = 1, \ldots, G$, $j = 1, \ldots, N_g$, again we have the mixture population model (robustness),

$$y_{gj} \mid \mu_g \sim (1-p)\text{Normal}(\mu_g, \gamma\sigma^2) + p\text{Normal}(\mu_g, \sigma^2), \tag{A.1}$$

where $0 < p < \frac{1}{2}, 0 < \gamma < 1$.

For $g = 1, \ldots, G$, $j = 1, \ldots, n_g$, again we have the sample model,

$$y_{gj} \mid \mu_g \sim (1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{A.2}$$

where $0 < p < \frac{1}{2}, 0 < \gamma < 1$.

Then,

$$\underline{\mu} \mid \underline{\eta} \sim \text{Normal}\{\underline{j}\theta + \underline{\eta}, \frac{\rho_1}{1-\rho_1}\sigma^2(R - \psi V)^{-1}\}. \tag{A.3}$$

We consider the Pitman-Yor two parameter stick-breaking process for $\eta_g, g = 1, \ldots, G$.

The stick-breaking process is

$$\pi(\eta_g \mid \underline{t}, etc.) = \sum_{s=1}^{G_o} p_s \text{Normal}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2),$$

where G_o is the number of distinct clusters with independence over $g = 1, \ldots, G$. Using the latent variables, d_g, it can be expressed in a computationally convenient form,

$$\pi(\eta_g, d_g \mid \underline{t}, etc.) = \prod_{s=1}^{G}[p_{d_g}\text{Normal}_{\eta_g}\{t_{d_g}, \frac{\rho_2}{1-\rho_2}\sigma^2\}]^{I(d_g=s)}$$

with independence over $(\eta_g, d_g), g = 1, \ldots, G$. The number of clusters, G_o, is the number of distinct d_g and d_g informs which cluster η_g belongs. Note that the limit of the product is ℓ

and it cannot be larger; this is different from a Dirichlet process, where the upper limit in the product goes to infinity. Here $p_1 = \nu_1, p_2 = \nu_2(1-\nu_1), \ldots, p_{G_o} = \prod_{s=1}^{G_o-1}(1-\nu_s)$ and

$$\nu_s \overset{ind}{\sim} \text{Beta}\{1-\delta_1, \frac{1-\delta_2}{\delta_2} + (s-1)\delta_1\}, s = 1, \ldots, G,\ 0 < \delta_1, \delta_2 < 1.$$

Then, for $\underline{\eta}$, we have

$$\underline{\eta} \mid \underline{t}, \underline{d} \sim \text{Normal}\{P\underline{t}, \frac{\rho_2}{1-\rho_2}\sigma^2 I\}, \tag{A.4}$$

where P is an incidence (partition) matrix (i.e., it consists of zeros and ones), mapping the areas to the clusters. Finally, we assume

$$t_s \overset{ind}{\sim} \text{Normal}\{0, \frac{\rho_3}{1-\rho_3}\sigma^2\}, s = 1, \ldots, G_o. \tag{A.5}$$

Note that $\underline{\eta}$ and $\underline{t}$ are G_o-vectors.

Also, it is clear that

$$Pr(d_g = s \mid \underline{t}, etc.) \propto p_s \text{Normal}_{\eta_g}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2), s = 1, \ldots, G, g = 1, \ldots, G.$$

Once $d_g, g = 1, \ldots, G$, are sampled, the incidence matrix, P is obtained, and all other parameters can be sampled. That is, draw the d_g first, and all other parameters can be sampled easily. Note at this moment, if we do not have enough z_s or ν_s, we must sample their priors.

We need to specify a proper prior for θ. Actually, $\mu_g = \phi_g + \eta_g, g = 1, \ldots, G$, where the ϕ_g are spatial effects and the η_g are clustering effects. Therefore, there is weak identifiability and we must take care of this issue. Now, $\text{E}(\mu_g \mid \theta, \eta_g) = \text{E}(\phi_g \mid \theta, \eta_g) + \eta_g$, and $\text{E}(\phi_g \mid \theta, \eta_g) = \theta$. Also, note that $\text{E}(\underline{\eta}) = \text{E}\{\text{E}(\underline{\eta} \mid \underline{t})\} = \text{E}(P\underline{t}) = \underline{0}$. Therefore, we have centered the μ_g on θ by taking $\text{E}(\phi_g \mid \eta_g) = \theta$ and $\text{E}(\underline{\eta}) = \underline{0}$ This centering together with proper diffused priors on θ can overcome the weak identifiability in this hierarchical Bayesian model. It is possible to improve this model further.

However, there are some additional problems that are likely to occur with stick breaking. First, there can be a single cluster, but at least three clusters are needed in any partition. One can use a random grouping with three clusters if this happens (this is rare). Second, to allow a relatively larger number of clusters, we take $1 - \delta_1 > \frac{1}{2}$, better than Jeffrey's prior for the first parameter of the beta density, and $\frac{1-\delta_2}{\delta_2} < 1$, thereby assisting the second beta parameter from getting too large a priori. These two conditions give $0 < \delta_1 < \frac{1}{2}$ and $\frac{1}{2} < \delta_2 < 1$. Third, an informative prior for θ is required. This can be obtained by using a small sub-sample of the data to avoid double using all the data. These three things help to obtain a more efficient Gibbs sampler.

To obtain the prior for θ, we can take a random sample of 10% of the data, $\bar{y}_g, g = 1, \ldots, G$. Now calculate the average, θ_o, and the variance, σ_o^2. Then, we take $\theta \sim \text{Normal}(\theta_o, \sigma_o^2)$. This is like a proper diffused prior, which avoids the uncertainty in inflating σ_o^2 near to

vagueness. Admittedly it double uses the data, but only 10% of the data, not all the data. Otherwise, we would need a prior for θ from an independent source. For example, if we also have the study variable from a small probability sample, we can use that as we have done here. Alternatively, we can take a small percent (*e.g.*, 5%) of the nps data, and do the same, and we can use the remaining data for the analysis. So that we do not need the ps for the construction of the prior.

The CPD of θ consists of two pieces, one from the proper prior and other from the model. It is easy to show that the contribution from the model is

$$\theta \sim \text{Normal}\left\{\frac{\sum_{g=1}^{G} r_g \mu_g}{\sum_{g=1}^{G} r_g}, \frac{\rho_1}{1-\rho_1}\sigma^2 \frac{1}{(1-\psi)\sum_{g=1}^{G} r_g}\right\},$$

where r_g are the row (column) sums of V. It is now a standard calculation to combine the two pieces.

Starting values for the Gibbs sampler can be obtained by first doing three things.

a. Find the sample averages, $\bar{y}_g = \frac{1}{n_g}\sum_{j=1}^{n_g} y_{gj}, g = 1, \ldots, G$.

b. Find clusters in these G stratum means, say 10 clusters, and this will give initial values of the d_g.

c. Form the partition matrix, P.

With this set up, we can generate starting values for all pertinent parameters. We can set $\rho_s = \frac{1}{2}, s = 1, 2, 3$, and $\delta_1 = \delta_2 = \frac{1}{2}$. Also, $t_s, s = 1, \ldots, 10$, can be obtained by averaging the appropriate y_g. Then, we can now sample the $\eta_g, g = 1, \ldots, G$. At the first iterate of the griddy Gibbs sampler, we can easily sample the d_g followed by all the parameters.

Appendix B: Unequal means in the two-component mixture model

We consider a single area (or stratum) to show how we can proceed. We assume

$$y_i \mid p, \mu, \sigma^2, \rho, \gamma \overset{ind}{\sim} (1-p)\text{Normal}_{y_i}(\mu - \gamma, \rho\sigma^2) + p\text{Normal}_{y_i}(\mu, \sigma^2), i = 1, \ldots, n,$$

where the first component has smaller mean and variance, and a priori,

$$\pi(p, \mu, \sigma^2, \rho, \gamma) \propto \frac{1}{\sigma^2}, 0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0.$$

Continuous survey data are typically skewed to the right, and so it is safe to take $\gamma > 0$. Of course, we can do regression (with covariates) in a similar manner.

The joint posterior density is

$$\pi(p, \mu, \sigma^2, \rho, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2}\prod_{i=1}^{n}\left\{\frac{1-p}{\sqrt{\rho\sigma^2}}\phi\{(y_i - \mu + \gamma)/\sqrt{\rho\sigma^2}\} + \frac{p}{\sqrt{\sigma^2}}\phi\{(y_i - \mu)/\sqrt{\sigma^2}\}\right\}, \tag{B.1}$$

where $\phi(\cdot)$ is the standard normal density and $0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0$.

We can simplify drawing samples from the joint posterior density by introducing latent variables, $z_i, i = 1, \ldots, n$, where $z_i = 0$ if an observation comes from the first component and $z_i = 1$ if an observation comes from the second component. Then, the augmented joint posterior density, starting with (B.1), is now

$$\pi(\underline{z}, p, \mu, \sigma^2, \rho, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2} \prod_{i=1}^{n} \left\{ \left[\frac{1-p}{\sqrt{\rho\sigma^2}} \phi\{(y_i - \mu + \gamma)/\sqrt{\rho\sigma^2}\} \right]^{1-z_i} \left[\frac{p}{\sqrt{\sigma^2}} \phi\{(y_i - \mu)/\sqrt{\sigma^2}\} \right]^{z_i} \right\}, \tag{B.2}$$

where $0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0$. It is now easy to run a Gibbs sampler to fit the joint posterior density in (B.2). It is advisable to sample the joint conditional posterior density of (μ, σ^2) (i.e., blocking).

Note that if we have only the first (or the second) component in the model (i.e., $p = 0$ or $p = 1$), then γ and ρ will not be identifiable. Therefore, it is necessary to assume there are at least two observations from each component of the mixture to avoid impropriety of the joint posterior density (i.e., $2 \leq \sum_{i=1}^{n} z_i \leq n - 2$). One way to do this is to arrange $y_1, \ldots, y_n$ in increasing order $y_{(i)}, i = 1, \ldots, n$, and take the corresponding $z_1 = z_2 = 0$ and $z_{n-1} = z_n = 1$, where $z_i, i = 3, \ldots, n - 2$, are determined from the joint posterior density. Doing so will avoid specifying this difficult constraint in the joint posterior density in (B.2).

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Beaumont, J-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, **46**, 1-28.

Beaumont, J-F. and Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, **83**, 11-22, DOI: 10.17226/24893.

Battese, G. E., Harter, R., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. **83**, 28-36, DOI: 10.2307/2288915.

Chakraborty, A., Datta, G. S., and Mandal, A. (2019). A robust hierarchical Bayes small area estimation for nested error linear regression model. *International Statistical Reviews*, **87**, S1, S158-S156, DOI: 10.1111/insr.12283.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011-2021, DOI: 10.1080/01621459.2019.1677241.

Chung, H. C. and Datta, G. S. (2022). Bayesian spatial models for estimating means of sampled and nonsampled small areas. *Survey Methodology*, **48**, 463-489.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, **32**, 249–264, DOI: 10.1214/16-STS598.

Goyal, S., Datta, G. S., and Mandal, A. (2020). A hierarchical Bayes unit-level small area estimation model for normal mixture populations. *Sankhya*, Series B, S1-S27, DOI: 10.1007/s13571-019-00216-8.

He, Z. and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, **56**, 360-367, DOI: 10.1111/j.0006-341X.2000.00360.x

Ibrahim, J. G. and Chen, M-H. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46-60, DOI: 10.1214/ss/1009212673.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-172.

Kim, J., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society*, Series A, **184**, 941-963.

Leon, de A. R. and Carriere, K. C. (2005). A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, **92**, 174-185, DOI: 10.1016/j.jmva.2003.08.006.

Lockwood, A. (2023). *Bayesian Predictive Inference for a Study Variable Without Specifying a Link to the Covariates.* PhD Dissertation, Department of Mathematical Sciences, Worcester Polytechnic Institute, pg. 1-110.

Marella, D. (2023). Adjusting for selection bias in non-probability samples by empirical likelihood approach. *Journal of Official Statistics*, **39**, 2023, 151-172, DOI: 10.2478/JOS-2023-0008.

Molina, I., Nandram, B., and Rao, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, **8**, 852-.885, DOI: 10.1214/13-AOAS702.

Nandram, B., Toto, M. C. S., and Choi, J. W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, **81**, 1593-1608.

Nandram, B. (2007). Bayesian predictive inference under informative sampling Via surrogate samples. In *Bayesian Statistics and Its Applications, Eds. S.K. Upadhyay, Umesh Singh and Dipak K. Dey*, Anamaya, New Delhi, Chapter 25, 356-374.

Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, **105**, 120-135, DOI: 10.1198/jasa.2009.ap08443.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 4-17, DOI: 10.5539/ijsp.v10n6p5.

Nandram, B. and Rao, J. N. K (2021). A Bayesian approach for integrating a probability sample with a nonprobability sample. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1568-1603.

Nandram, B. and Rao, J. N. K (2023). Bayesian predictive inference when integrating a nonprobability sample and a probability sample. *arXiv:2305.08997v1 [Stat.ME], 15 May 2023*, pg. 1-35.

Nandram, B. and Rao, J. N. K. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications*, **22** 345-376, ISSN 2454-7395.

Nandram, B. (2023). Overcoming challenges associated with early Bayesian state estimation of planted acres in the United States. *Special Proceedings of the Twenty-fifth Conference of the Society of Statistics, Computer and Applications*, ISBN #: 978-81-

950383-2-9, 25th Annual Conference, 15-17 February 2023; pp 51-78.

Nandram, B. (2022), A Bayesian assessment of non-ignorable selection of a non-probability Sample. *Indian Bayesians' News Letter, Invited Paper*, **14**, November 2022, 7-20.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya*, Series B, **83**, 242-272, DOI: 10.1007/s13571-020-00227-w.

Rafei, A., Elliott, M. R., and Flannagan, C. A. C. (2022). Robust and efficient Bayesian inference for non-probability samples. *arXiv:2203.14355Vi*, pp. 1-46.

Sakshaug, J. W., Wisniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, **35**, 653-681, DOI: 10.2478/jos-2019-0027.

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wisniowski, A., and Struminskaya, B. (2023). Bayesian integration of probability and non-probability samples for logistic regression. *Journal of Survey Statistics and Methodology*, 00,c1-35, DOI: 10.1093/jssam/smad041.

Scott, A. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, **64**, 830-840, DOI: 10.1080/01621459.1969.10501015.

Toto, M. C. S. and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference*, **140**, 2963-2979, DOI: 10.1016/j.jspi.2010.03.043.

Wisniowski, A., Sakshaug, J. W., Ruiz, D. A. P., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey Inference, *Journal of Survey Statistics and Methodology*, **8**, 120-147, DOI: 10.1093/jssam/smz051.

Wu, C. (2022). Statistical Inference with non-probability survey samples. *Survey Methodology (With Discussions)*, **48**, 283-311.