Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 343–374
http://www.ssca.org.in/journal

# Bayesian Integration for Small Areas by Supplementing a Probability Sample with a Non-probability Sample

**Balgobin Nandram[1] and J. N. K. Rao[2]**

[1] *Department of Mathematical Sciences*
*Worcester Polytechnic Institute, Worcester, MA 01609, United States of America*
[2] *School of Mathematics and Statistics, Carleton University, Ottawa, K1S 5B6, Canada*

*(This is a special invited paper on request from the Chair Editor.)*

Received: 26 June 2023; Revised: 5 January 2024; Accepted: 8 January 2024

**Abstract**

We consider the problem of data integration in small area estimation, where a non-probability sample (nps) and a relatively small probability sample (ps) are available from each area. By definition, for the nps, there are no survey weights, but for the ps, there are survey weights. A recent method, based on a pseudo-likelihood, is used to estimate the survey weights in the nps, and thereafter assumed known. The key issue we address is that the nps, although much larger than the ps, can lead to a biased estimator of a finite population quantity of each area but with much smaller variance. We assume that there are common covariates and responses for everyone in the two samples, no covariates available for nonsampled units, and no overlaps of the two samples by area. In the data integration, we use the nps to construct a prior for the ps, and partial discounting of the nps is incorporated to avoid a dominance of the prior. Inverse probability weighting is used to assist Bayesian predictive inference via surrogate sampling of the finite population means and percentiles. The Gibbs sampler, with some collapsing to speed up convergence and to provide strong mixing, is carefully executed to fit the joint posterior density. In our illustrative example on body mass index, our data-integrated model is preferred over the ps only model and other competitors. The data-integrated model provides small area estimates, roughly similar to those of the ps only model, with larger precision.

**Key words:** Bayesian diagnostics; Finite population quantities; Gibbs sampler; Inverse probability weighting; Power prior; Surrogate samples.

## 1. Introduction

We assume that there are data from a number of small areas, and from each area we have a non-probability sample (nps,1) and a probability sample (ps, 2), the probability sample being much smaller than the non-probability sample. The problem is how to improve

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

inference for each area based on the ps, but supplemented by the nps, and we do not want the nps to dominate the analysis. While the nps may be biased, the ps is considered unbiased when the survey weights are incorporated. In a similar manner, because of its size the nps will provide improved precision but it will provide biased estimates, which we do not want to happen. Probability sampling is the gold standard among all data collection procedures, but this is still problematic because nonresponse has become a serious concern. How can we provide small area estimates with relatively small bias, possibly closed to the ps, with better precision than only the ps can provide?

There are efforts to combine both probability and nonprobability samples to produce a single inference that compensates for the limitations of each process. Typically the nonprobability sample is relatively large, as in big data, but one needs to be careful with the bias it introduces into the final estimates. Meng (2018) argued that a small bias in big data can be catastrophic; see also Nandram and Rao (2021, 2023) for a review and an interpretation of Meng (2018) relevant to nonprobability sample. Perhaps if one uses only the covariates in the big data, there may not be significant bias, but it is a different issue if one wants to use the study variable from the big data as well.

Most of the work on nonprobability sampling has been in the non-Bayesian setting, mostly randomization-based analysis. For example, Elliott and Haviland (2007) evaluated a composite estimator to supplement a standard probability sample with a nonprobability sample. They showed that the estimator, based on a linear combination of both sample processes and a bias function, can produce estimates with a smaller mean squared error (MSE) relative to a probability-only sample. See Elliott and Valliant (2017) for an informative review of the design-based approach, where they discussed quasi-randomization.

Sakshaug *et al.* Blom (2019) and Wisnioski *et al.* (2020) introduced a Bayesian approach in which survey weights are incorporated as a covariate and there is no need to estimate the probabilities of the nps. The underlying idea is that probability and nonprobability samples can be integrated in a way that exploits their advantages to compensate for their weaknesses and improve estimation of model parameters. Salvatore *et al.* (2023) used a similar idea for binary data via logistic regression. Nandram and Rao (2021, 2023) showed how to combine a nps and a ps using a Bayesian model. They argued that the nps should be used to construct a prior, together with a discounting factor, and to obtain a prior for the hyper-parameters in the model, which begins with a weighted likelihood. As pointed out by both Sakshaug *et al.* (2019) and Wisnioski *et al.* (2020), it will be better to use a nonprobability sample to supplement a probability sample; see also Nandram and Rao (2021, 2023). Salvatore *et al.* (2023) also supported our idea.

Chen, Li and Wu (2020) used a ps and a nps to obtain survey weights in the nps in the design-based approach, where they made strong use of the Horvitz-Thompson and the Hajek estimators. There was no study variable in the ps and so this is really a very limited data integration problem. Actually their method cannot be extended to accommodate a study variable in the ps. Also, their method uses logistic regression to construct the propensity scores and then the survey (design) weights are obtained by taking reciprocals. This is ignorable selection. A summary of the approach of Chen, Li and Wu (2020) for propensity scores is given in Appendix A. For small area estimation, the computational procedure of Chen, Li and Wu (2020) is unstable, so we had to do this procedure for the entire ensemble at once, not each area at a time.

As pointed out by a reviewer, we should present the reasons why we do not use non-ignorable selection. We used ignorable selection because within the framework of Chen, Li and Wu (2020), it is not possible to obtain the propensity scores unless you want to use a mass imputation to 'manufacture' the values of the study variable for the probability sample. This is not in the spirit of our paper because we have the study variable for both the non-probability sample and the probability sample; yet we use the method of Chen, Li and Wu (2020) to get the propensity scores. We realized that we could have used both samples to get the propensity scores in the non-probability samples, but it does not fit directly into the framework of Chen, Li and Wu (2020). Non-ignorable selection is defined as

$$f(I = k, y \mid \underline{x}) = P(I = k \mid y, \underline{x}) f(y \mid \underline{x}), k = 0, 1,$$

where $I$ is the participation variable, $y$ the study variable and $\underline{x}$ the covariates. We have ignorable selection if $P(I \mid y, \underline{x}) = P(I \mid \underline{x})$, which is simpler than non-ignorable selection. Clearly, non-ignorable selection is preferred but it leads to computational instability. See Nandram and Choi (2010) and Nandram (2022) for more discussions on non-ignorability. One difficulty is that one needs $y$ to be strongly related to $\underline{x}$ and at the same time, both $\underline{x}$ and $y$ are used as covariates in the participation model. New research is needed at least within the Bayesian paradigm; see Marella (2023) for recent work on nonignorability, not within the Bayesian paradigm though. Data integration can be discussed without mentioning how the data are selected; see Salvatore *et al.* (2023) for binary data and others.

It is worth noting that all the above mentioned work do not consider data integration for small areas. Beaumont (2020) argued that it is sensible to use a non-probability sample to supplement a probability sample in small area estimation; see also Beaumont and Rao (2021). For one thing, small sample sizes within small areas do not lead to adequate precision. The small area model will include random effects as an attempt to discriminate the areas. These works use the area-level Fay-Herriot model. However, there is virtually no work using the unit level model like that of Battese, Harter and Fuller (1988) for integration of a non-probability sample and a probability sample partly because it is a less practical to get unit-level data in both the nps and the ps, but it is possible. Again see Nandram and Rao (2021, 2023).

Rao (2020) stated that a non-probability sample can be used to construct covariates for probability samples in small area estimation. The use of area level big data as additional predictors in the area level model has the potential of providing good predictors for modeling. He mentioned four applications that have used big data covariates in an area level model; see Marchetti *et al.* (2015), Porter *et al.* (2014), Schmid *et al.* (2017) and Muhyi *et al.* (2019) for the four applications. Rao (2020) also cited applications where unit level models are used; see Chambers *et al.* (2019). Again, if one wants to use both the study variable and the covariates from the big data, one might need the unknown selection probabilities. However, one does not really need to estimate the selection probabilities because one can use structural (measurement error) models; see discussions in the concluding remarks and Berg *et al.* (2021). One drawback of structural models is that there will be non-identifiable parameters which will create difficulties in model fitting, especially if Markov chain Monte Carlo methods must be used.

In our paper, we actually used a power prior to discount the non-probability sample, which we treat as historical data to construct a prior distribution for the parameters of the probability sample. The parameters in the two models are basically the same, and their priors come from the non-probability sample. In general, if you start with a density, $g(y \mid \underline{\theta})$, we can penalize it by using

$$f(y \mid \underline{\theta}, a) = \frac{\{g(y \mid \underline{\theta})\}^a}{\int \{g(y \mid \underline{\theta})\}^a dy}, 0 \leq a \leq 1.$$

So we actually use $f(y \mid \underline{\theta}, a)$ for the non-probability sample and $g(y \mid \underline{\theta})$ for the probability sample. For example, if $a = 1$, there will be no discounting, and if $a = 0$, the non-probability sample will not be used. Details of the power prior in data integration is reviewed in Nandram and Rao (2021, 2022); see Ibrahim and Chen (2000) and Ibrahim *et al.* (2015) for a review and many applications of the power prior in more general settings.

Small area estimation (SAE) is an important problem facing many government agencies. They want to do estimation for each area, but for most small areas the direct estimates are unreliable. Then, pooling of the data over the entire ensemble is required to get reliable estimates for each area. While the SAE problem is difficult in its own right, there is additional complexity to integrate the non-probability sample and the probability sample.

To focus our development, we study body mass index (BMI) as the variable of interest with covariates, age, race and sex, from eight counties in California, based on a probability sample. The covariates, responses (BMI) and survey weights are all known. We construct a small-area example out of these data with two samples from each of the eight counties (about 80% for nps and 20% for the ps). To form a practical example, we discarded the weights from the nps and they are assumed unknown. The population size of each county is roughly the sum of the survey weights in the ps. Here, the covariates, responses and survey weights in the nps are respectively $(\underline{x}_{1ij}, y_{1ij}, w_{1ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{1i}$, and the covariates, responses and survey weights of the ps are $(\underline{x}_{2ij}, y_{2ij}, w_{2ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{2i}$; the survey weights $w_{1ij}$ are unknown in the nps.

The small area model has the following features.

a. The two sets of covariates are commensurate (*i.e.*, the same covariates are measured in the non-probability sample and the probability sample; or at least only a common set of covariates will be used).

b. Pooling will take place using a common set of regression coefficients and variance components over all areas in the two samples. The nps is essentially used to construct a prior for the hyperparameters and this prior is discounted using a power prior.

c. Within an area, the random effects are the same in the model that links the non-probability sample and the probability sample.

d. It is possible to have some areas with only a probability sample, and some areas with only a non-probability sample, but there must be a common set. This can be done within our approach, but we will not pursue this issue further in this paper.

Finally, a reviewer asked why there is a need for super-population models. Clearly, it will be better to do data integration without specifying the super-population model. Most

Bayesian methods used the super-population model; Wang *et al.* (2018) is an exception and it uses an approximate Bayesian method. In fact, they used the sampling distribution of a summary statistic to derive the posterior distribution of the parameters of interest, but this is not quite within the Bayesian paradigm. However, there is a need to robustify both models for the study variable and the participation variable. We have indicated how to do so in the concluding remarks, and this is an on-going activity. In our on-going research work, by using a double mass imputation (Kim, *et al.* 2021, Chen, *et al.* 2022), we can avoid a participation model but we do need a robust model for the study variable if we use a Bayesian method. One of the authors gave a couple of talks on this topic already.

This paper has five sections, including this one, and it is an extension of Nandram and Rao (2021, 2023) to cover Bayesian data integration for small areas. In Section 2, we review the single area model of Nandram and Rao (2021, 2023). In Section 3, we discuss small area estimation using a unit-level model, show how to operationalize the proposed model to provide fast computation for a large number of areas, and describe how to estimate finite population percentiles. In Section 4, we provide the analysis of the numerical example as we described above. Section 5 provides some concluding remarks and extensions. The appendices provide technical details on propensity scores, computation for the small area model, Bayesian model diagnostics, and the ps only model.

## 2. Review of the single area model

In this paper, we extend the single area model of Nandram and Rao (2021, 2023) to accommodate data integration for small areas. Therefore, it is pertinent for us to describe the single area model to motivate the small area model.

We have two samples from a single area, which are the nps (1) and the ps (2). We have $(W_{ti}, \underline{x}_{ti}, y_{ti}), i = 1, \ldots, n_t, t = 1, 2$, where $W_{1i}$ are unknown, but $W_{2i}$ are assumed known. We plan to construct a prior for the regression coefficients and the variance parameters using a discount factor (power prior) to help mitigate the nps from dominating the ps. (Throughout, as covariates are assumed fixed, conditioning on them will be omitted.)

For the nps, propensity scores, assumed strictly positive, are estimated using logistic regression (Chen, Li and Wu, 2020; see Appendix A of the current paper for a review), so for the nps probability enters through quasi-randomization (*e.g.*, Elliott and Valliant, 2017). The method of CLW is used to estimate the propensity scores, $\pi_{1i}$, and the weights of the nps are $W_{1i} = N\frac{1/\pi_i}{\sum_{j=1}^{n_1} 1/\pi_j}, i = 1, \ldots, n_1$, where $N$ is the population size, and the Horvitz-Thompson estimator of $N$ is $\sum_{i=1}^{n_2} W_{2i}$. This assumes ignorability in which given the covariates, the study variable and the participation variable are independent and it also assumes that the propensity scores depend only on the covariates, which is not unreasonable; see Nandram (2022) for a discussion about nonignorability. These estimated weights, $W_{1i}$, are assumed known throughout our work. In our models, associated with weighted likelihood, we use normalized densities with adjusted weights to get a more appropriate measure of variability. The adjusted weights are

$$w_{ti} = \hat{n}_t W_{ti} / \sum_{j=1}^{n_t} W_{tj}, \hat{n}_t = \left( \sum_{j=1}^{n_t} W_{tj} \right)^2 / \sum_{j=1}^{n_t} W_{tj}^2, i = 1, \ldots, n_t, t = 1, 2,$$

where $\hat{n}_t$ is the effective sample size; see Potthof *et al.* (1992).

The population model, which we assume holds, is

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

A finite population quantity (mean or percentile) can be estimated using surrogate sampling (Nandram 2007). That is, the entire population is sampled given $(\underline{\beta}, \sigma^2)$. However, the question is how to get samples of $(\underline{\beta}, \sigma^2)$, and this is where most of the work is needed. We need to adjust the population model to accommodate the two samples, in which the nps is penalized using a power prior; see Nandram and Rao (2021, 2023) for a quick review of the power prior and how it is used in our work.

The model that combines the two samples, in which the nps is used to supplement the ps is

$$y_{ti} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\underline{x}_i'\underline{\beta}, \frac{\sigma^2}{a_t w_{ti}}\right),$$

$$\pi(\underline{\beta}, \sigma^2, a) \propto 1/\sigma^2, a_2 = 1, 0 < a_1 = a < 1, i = 1, \ldots, n_t, t = 1, 2,$$

where $a$ is the discounting factor with a uniform prior and $w_{ti}$ are adjusted weights. The joint posterior density of $(\underline{\beta}, \sigma^2, a)$ has been shown to be proper and it can be fit using a grid sample (the posterior density of $a$ is non-standard); see Nandram and Rao (2021, 2023) for details.

Nandram and Rao (2021, 2023) obtained Bayesian predictive inference for the finite population mean using

$$\pi(\bar{Y} \mid \underline{y}_1, \underline{y}_2) = \int f(\bar{Y} \mid \underline{\beta}, \sigma^2)\pi(\underline{\beta}, \sigma^2 \mid \underline{y}_1, \underline{y}_2) d\underline{\beta} d\sigma^2,$$

where $\underline{y}_1$ and $\underline{y}_2$ are the two samples. Note that $f(\bar{Y} \mid \underline{\beta}, \sigma^2)$ does not depend on $(\underline{y}_1, \underline{y}_2)$, unlike standard Bayesian predictive inference, a feature of surrogate sampling; see Nandram (2007). Note that

$$\bar{Y} \mid \underline{\beta}, \sigma^2 \sim \text{Normal}\left(\underline{\bar{X}}'\underline{\beta}, \frac{\sigma^2}{N}\right),$$

where we use the Horvitz-Thompson estimator of the finite population mean vector covariate, $\underline{\bar{X}}$, which is $\frac{1}{N}\sum_{i=1}^{n_2} W_{2i}\underline{x}_{2i}$; this is actually the Hajek estimator because $N$ is assumed unknown.

Inference about a finite population percentile is a related, but different, problem. This is discussed in Section 3. Inference about the finite population percentiles is also a problem in our study on body mass index (*e.g.*, the $85^{th}$ percentile is a measure of overweight).

## 3. A small area model for data integration

We show how to extend the model of Nandram and Rao (2021) to accommodate a number of areas. This uses an extended version of the unit-level model of Battese, Harter and Fuller (BHF, 1988). See also Toto and Nandram (2011) and Molina, Nandram and Rao (2014) for the Bayesian version of the BHF model.

We assume there are $\ell$ areas and within the $i^{th}$ area, there are a non-probability sample of size $n_{1i}$ and a probability sample of size $n_{2i}$ where "1" and "2" respectively refer to the non-probability sample and the probability sample, maintaining the notation in the single area example, and the population size is $N_i$. [Note that the nps and the ps of each area come from the same distinct sub-population; so there is single subscript in $N_i$.] For $i = 1, \ldots, \ell$, the covariates are $(\underline{x}_{sij}, j = 1, \ldots, n_{sij}, s = 1, 2)$, but the covariates are unobserved for the nonsampled units, and the responses are $y_{sij}, j = 1, \ldots, n_{si}$. There are also survey weights for the probability sample, denoted by $W_{2i}$ (known). There are no survey weights for the non-probability sample and these are estimated using the method of Chen, Li and Wu (2020); again see Appendix A. The population size for the $i^{th}$ area is estimated by $N_i = \sum_{j=1}^{n_{2i}} W_{2ij}, i = 1, \ldots, \ell$. Bayesian predictive inference is required for the finite population area means,

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}, i = 1, \ldots, \ell,$$

based on the non-probability samples and probability samples, where $y_{ij}$ are the unknown population values. Of course, the model permits the use of the non-probability sample, as we have seen for the single sample model. That is, there is pooling across areas and within areas from both the non-probability sample and the probability sample.

As we have stated, the discounting factors will only be included for the nps, which will be used to construct the prior (the nps is viewed as historical data) and the ps will be used as the actual data. For generality, these discounting factors depend on areas. That is, for $s = 1$ (nps), $a_{si} = a_i, i = 1, \ldots, \ell$ (allowing discounting) and for $s = 2$ (ps), $a_{si} = 1, i = 1, \ldots, \ell$ (no discounting).

### 3.1. Proposed small area model

Our model for the two samples over the areas is

$$y_{sij} \mid \nu_i, \underline{\beta} \overset{ind}{\sim} \text{Normal}(\underline{x}_{sij}\underline{\beta} + \nu_i, \frac{\sigma^2}{a_{si} w_{sij}}), j = 1, \ldots, n_{si}, s = 1, 2,$$

where $w_{sij}$ are the adjusted weights within areas. The weights for the nps are obtained using the method of Chen, Li and Wu (2020) over the entire ensemble (assumed known henceforth), and then the weights for both the nps and ps are used to provide the adjusted weights, as was done in the single area example. The fact that we are assuming the estimated weights are known is an important caveat of our work, and this is on-going research activity. A priori, for the random effects, we assume that

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

and for the hyperparameters, we assume

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, 0 < \rho < 1.$$

Again note that these are two BHF models, one for the non-probability samples and the other for the probability samples. But they are connected because they have the same parameters (except the nps has the discounting factors), and this is how we link the nps, ps and the small areas.

For the discounting factors $0 \le a_i \le 1$, we will assume that for $i = 1, \ldots, \ell$,

$$a_i \mid \theta, \gamma \overset{ind}{\sim} \text{Beta}\left\{\theta \frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right\}, 0 < \theta, \gamma < 1.$$

We need to specify the priors for $\theta$ and $\gamma$. We make a modest assumption that the distribution of each $a_i$ is log-concave, and a sufficient condition for this to happen is that $\theta\frac{1-\gamma}{\gamma} > 1$ and $(1-\theta)\frac{1-\gamma}{\gamma} > 1$. (A log-concave density has very nice properties, specifically its moment generating function exists.) This means that $0 < \gamma < \frac{1}{3}$, $\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}$. Therefore, the prior for $(\underline{a}, \theta, \gamma, \rho)$ is

$$\pi(\underline{a}, \theta, \gamma, \rho) = \left\{\prod_{i=1}^{\ell} \frac{a_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-a_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\}}\right\}, 0 < \gamma < \frac{1}{3}, \frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}, 0 < \rho < 1.$$

Note that this model holds for the entire population with $w_{sij} \equiv 1$.

Using Bayes' theorem, letting $\underline{y}$ (both nps and ps) denote the vector of all observations, the joint posterior density is

$$\pi(\underline{\nu}, \underline{a}, \underline{\beta}, \sigma^2, \rho, \theta, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2}\prod_{i=1}^{\ell}\left\{\left[\prod_{j=1}^{n_{1i}}\sqrt{\frac{a_i w_{1ij}}{2\pi\sigma^2}}e^{-\frac{a_i w_{1ij}}{2\sigma^2}(y_{1ij}-\underline{x}'_{1ij}\underline{\beta}-\nu_i)^2}\prod_{j=1}^{n_{2i}}\sqrt{\frac{w_{2ij}}{2\pi\sigma^2}}e^{-\frac{w_{2ij}}{2\sigma^2}(y_{2ij}-\underline{x}'_{2ij}\underline{\beta}-\nu_i)^2}\right]\right.$$
$$\left.\times\sqrt{\frac{1-\rho}{2\pi\rho\sigma^2}}e^{-\frac{1-\rho}{2\rho\sigma^2}\nu_i^2}\;\frac{a_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-a_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\}}\right\}. \tag{1}$$

Letting $\Omega_1 = (\underline{a}, \theta, \gamma, \rho)$ and $\Omega_2 = (\underline{\nu}, \underline{\beta}, \sigma^2)$, to fit the posterior density in (1), we will first integrate out $\Omega_2$ and sample the posterior density of $\Omega_1 \mid \underline{y}$ using the Gibbs sampler; see Appendix B. Then, we can sample $\Omega_2 \mid \Omega_1, \underline{y}$ using the composition method via

$$\pi(\Omega_2 \mid \Omega_1, \underline{y}) = \pi_1(\sigma^2 \mid \Omega_1, \underline{y})\pi_2(\underline{\beta} \mid \sigma^2, \Omega_1, \underline{y})\pi_3(\underline{\nu} \mid \underline{\beta}, \sigma^2, \Omega_1, \underline{y}),$$

where $\pi_1(\sigma^2 \mid \Omega_1, \underline{y})$, $\pi_2(\underline{\beta} \mid \sigma^2, \Omega_1, \underline{y})$ and $\pi_3(\underline{\nu} \mid \underline{\beta}, \sigma^2, \Omega_1, \underline{y})$ are all in standard forms, inverse gamma, p-variate normal and independent normals respectively; see Appendix B. This strategy provides a more efficient computational algorithm (better convergence and mixing of the Gibbs sampler).

Bayesian predictive inference is required for $\bar{Y}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} y_{ij}$, where $y_{ij}$ are the population values (unknown). As the sample values, $y_{sij}$, are corrupted because of the survey weights, we cannot use them. So we use surrogate sampling; in principle the entire population is drawn, not the values for the individual units though. Therefore,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\underline{\bar{X}}'_i\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}\right), i = 1, \ldots, \ell,$$

where $\underline{\bar{X}}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} \underline{x}_{2ij}$ and $N_i$ are assumed unknown. We use the Horvitz-Thompson estimators $\underline{\bar{x}}_{2i} = \frac{\sum_{j\in S_{2i}} w_{2ij}\underline{x}_{2ij}}{\sum_{j\in S_{2i}} w_{2ij}}$ and $\sum_{j\in S_{2i}} w_{2ij}$ to estimate $\underline{\bar{X}}_{2i}$ and $N_i$ respectively (inverse probability weighted estimators - IPW), where $S_{2i}$ is the set of units in the $i^{th}$ area of the ps. Then,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\underline{\bar{x}}'_{2i}\underline{\beta} + \nu_i, \frac{\sigma^2}{\sum_{j\in S_{2i}} w_{2ij}}\right), i = 1, \dots, \ell. \quad (2)$$

Once we have drawn $(\underline{\nu}, \underline{\beta}, \sigma^2)$ using the Gibbs sampler, we simply draw the $Y_i$ from (2). According to the model, all the sampled data are used in the predictive inference.

Observe that $\text{E}(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2, \rho) = \underline{\bar{x}}'_{2i}\underline{\beta} + \lambda_i(\bar{\bar{y}}_i - \underline{\bar{\bar{x}}}_i{}'\underline{\beta})$, where

$$\lambda_i = \frac{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij} + (1-\rho)}, \phi_{sij} = \frac{a_{si}w_{sij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}},$$

$$\bar{\bar{y}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}y_{sij}, \quad \underline{\bar{\bar{x}}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}\underline{x}_{sij};$$

see Appendix B for definitions. Then,

$$\text{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) = \lambda_i\bar{\bar{y}}_i + (1-\lambda_i)\underline{\bar{\bar{x}}}_i{}'\underline{\beta} + (\underline{\bar{x}}_{2i} - \underline{\bar{\bar{x}}}_i)'\underline{\beta}$$

and

$$\text{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) = \left\{\frac{1}{\sum_{j=1}^{n_{2i}} w_{2ij}} + \frac{\rho}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij} + (1-\rho)}\right\}\sigma^2.$$

These can be used to form Rao-Blackwellized density estimators for $\bar{Y}_i$.

More importantly, we can study the behavior of $\text{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$ and $\text{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$ to see the importance of $\rho$. As $\rho \to 0$, $\lambda_i \to 0$,

$$\text{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \underline{x}'_{2i}\underline{\beta}$$

and

$$\text{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \frac{\sigma^2}{\sum_{j=1}^{n_{2i}} w_{2ij}}.$$

That is, the non-probability sample does not play a major role. As $\rho \to 1$, $\lambda_i \to 1$,

$$\text{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \underline{x}'_{2i}\underline{\beta} + (\bar{\bar{y}}_i - \underline{\bar{\bar{x}}}'\underline{\beta})$$

and

$$\text{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \left\{\frac{1}{a\sum_{j=1}^{n_{1i}} w_{1ij} + \sum_{j=1}^{n_{2i}} w_{2ij}} + \frac{1}{\sum_{j=1}^{n_{2i}} w_{2ij}}\right\}\sigma^2.$$

Both samples are important.

### 3.2. Operationalizing the small area model

Apart from the exchangeable assumption on the $a_i$, the current small area model is essentially robust with respect to the $a_i$. But with a large number of areas, it will be too slow to sample all the $a_i$ using the grid method. One possibility is to smooth out the $a_i$ in an attempt to operationalize the algorithm.

We can assume that the $a_i$ are "proportional" to the sample sizes or better yet to their logarithms. This will also eliminate the exchangeability assumption. Therefore, one possibility is to take

$$a_i = \frac{e^{\gamma_0 + \gamma_1 \log(n_i)}}{1 + e^{\gamma_0 + \gamma_1 \log(n_i)}}, i = 1, \ldots, \ell,$$

where for the $i^{th}$ area, $n_i$ is the sample size of the nonprobability sample or the total sample size. We are assuming here that $-\infty < \gamma_0 < \infty, 0 < \gamma_1 < \infty$.

Then, clearly

$$a_i = \frac{\alpha_0 n_i^{\gamma_1}}{1 + \alpha_0 n_i^{\gamma_1}}, \alpha_0 = e^{\gamma_0}, i = 1, \ldots, \ell.$$

Now, letting $\alpha_0 = \frac{\phi_0}{1-\phi_0}$ and $\alpha_1 = \frac{\phi_1}{1-\phi_1}$, we have

$$a_i = \frac{\phi_0 n_i^{\frac{\phi_1}{1-\phi_1}}}{1 - \phi_0 + \phi_0 n_i^{\frac{\phi_1}{1-\phi_1}}}, i = 1, \ldots, \ell, \tag{3}$$

where $0 < \phi_0, \phi_1 < 1$. Note that if $\phi_1 = 0$, then $a_i = \phi_0$ and there will be no dependence on the $n_i$. Now, simply substitute the $a_i$ in (3) into the SAE model and use the prior

$$\phi_0, \phi_1 \overset{ind}{\sim} \text{Uniform}(0, 1).$$

This reduces the number of parameters for this part of the model from $\ell + 2$ to just two and actually the two parameters, $\theta$ and $\gamma$, are now eliminated or replaced by $\phi_0$ and $\phi_1$. So if $\ell$ is large, not just 8, there will be large gains in computational time. This is how the procedure is operationalized.

### 3.3. Percentiles

As we consider each area individually, we can drop the subscript, $i$, to get the population model

$$y_j \mid \underline{\beta}, \nu, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_j' \underline{\beta} + \nu, \sigma^2), j = 1, \ldots, N.$$

We recall that the nonsampled covariates are unknown. In principle, if we can get the nonsampled covariates, then, given $\underline{\beta}, \nu, \sigma^2$, we can sample $y_j, j = 1, \ldots, N$. Then, for $0 < \gamma < 1$, the $[\gamma N]$ percentile is $Y_{[\gamma N]}$, an order statistic. But this procedure is prohibitively expensive because the nonsampled covariates are unknown and $N$ is large.

However, it is possible to obtain finite population percentiles (needed for BMI data) without the nonsampled covariates. For BMI, the $85^{th}$ and $95^{th}$ percentiles respectively measure overweight and obsesity. First, note that

$$P(Y_j < t_j \mid \nu, \underline{\beta}, \sigma^2) = \Phi \left\{ \frac{t_j - \underline{x}_j' \underline{\beta} - \nu}{\sigma} \right\},$$

where $\Phi(\cdot)$ is the standard normal cdf. Therefore, with $0 < \gamma < 1$, the $100(1-\gamma)^{th}$ percentile of $Y_j$ is $t_j = \underline{x}_j'\underline{\beta} + \nu + \sigma\Phi^{-1}(\gamma)$, Then, for the $h^{th}$ iterate from the Gibbs sampler, the $100(1-\gamma)^{th}$ percentile of $Y_j$ is

$$t_j^{(h)} = \underline{x}_j'\underline{\beta}^{(h)} + \nu^{(h)} + \sigma^{(h)}\Phi^{-1}(\gamma),$$

and the $100(1-\gamma)^{th}$ finite population percentile is $\frac{\sum_{j=1}^{n_2} W_{2j} t_j^{(h)}}{\sum_{j=1}^{n_2} W_{2j}}$. Some improvements can be made; actually such improvements are not necessary because $N$ is very large, and like the finite population mean, the variance is approximately zero.

Walker (1968) showed that the sample $\gamma$-quantile, $Y_{([N\gamma])} \sim aN\{\epsilon_\gamma, \frac{\gamma(1-\gamma)}{Nf^2(\epsilon_\gamma)}\}$, where $\epsilon_\gamma$ is the $\gamma^{th}$ quantile of the population, $f(\cdot)$, which is assumed to be continuous with $f(\epsilon_\gamma) > 0$ and $F(\epsilon_\gamma) = \gamma$ uniquely. Here, we simply take $\epsilon_\gamma = \frac{\sum_{j=1}^{n_{2i}} W_{2ij} t_{ij}^{(h)}}{\sum_{j=1}^{n_{2i}} W_{2ij}}$ and because the variance is $o(\frac{1}{N})$ and $N$ is very large, essentially $Y_{([N\gamma])}$ is a point mass at $\epsilon_\gamma$. A similar result holds for $\bar{Y}_i$.

One question is how to define $f(y)$. We write $y_j \mid \underline{\beta}, \nu, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_j'\underline{\beta} + \nu, \sigma^2), j = 1, \ldots, N$. Then, we replace $\underline{x}_j, j = 1, \ldots, N$, by the weighted average, $\underline{d} = \frac{\sum_{j=1}^{n_2} W_{2j}\underline{x}_j}{\sum_{j=1}^{n_2} W_{2j}}$, to get $y_j \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{d}'\underline{\beta} + \nu, \sigma^2), j = 1, \ldots, N$. Finally, $f(\epsilon_\gamma) = \frac{1}{\sigma}\phi(\frac{\epsilon_\gamma - \underline{d}'\underline{\beta} - \nu}{\sigma})$, where $\phi(\cdot)$ is the standard normal density.

## 4. Numerical example on small area estimation

We use the BMI data from the eight counties of California to construct a practical example; see Nandram and Choi (2010) for design issues in the National Health and Nutrition Examination Survey (NHANES III). We use Bayesian model diagnostics to compare all the models. Then, we compare our selected model with data integration and the ps only model via Bayesian predictive inference of the finite population mean and the $85^{th}$ finite population percentile.

But, first we discuss the performance of the Gibbs sampler for the model with discounting (the other models are similar). The entire computation consists of three parts (a) constructing the unknown survey weights for the nps, (b) fitting the individual area model, and (c) fitting of the small area model. The entire computation took nearly 40 minutes with (c) taking almost all the time. We started the Gibbs sampler arbitrarily by taking the $a_i$ to be the corresponding posterior means from the individual area model, set $\rho = .5$, its mid range, and as the mid point of the interval $(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$ is .5, set $\theta = .5$ and and $\gamma = 1/6$, its mid range. We ran $21,000$ iterates, used 1000 as a "burn in" and systematically selected every twentieth to get a 'random' sample of $M = 1,000$. We also performed the diagnostic procedures for the Gibbs sampler. The auto-correlations are not significant, the trace plots show no trend, Geweke tests of stationarity are all passed and the effective sample size are all satisfactory, mostly near to 1000. Table 1 has the p-values and the effective sample sizes. The fact that the effective sample size (ESS) is about 550, not 1000, for $\theta$ and $\gamma$ is not a problem because $\theta$ and $\gamma$ are hyperparameters of the $a_i$, which perform well.

In Table 2 we present diagnostic measures to compare the small area models. These are the negative log pseudo marginal likelihood (LPML), the deviance information criterion (DIC), the Bayesian predictive p-value (BPP), a divergence measure (DM) and the posterior root mean square error (PRMSE); see Appendix C for a review of the definitions of these measures. Smaller values of all quantities, except BPP, show better fit; values of BPP in $(.05, .95)$ show good fitting models.

All measures show that the model without discounting is not competitive, and DM and PRMSE show that the PS only model is not competitive, leaving us with two models, discounting and logit. In terms of PRMSE, the model with discounting is approximately 10% better than the logit model, which is not robust because it assumes linearity between the discounting factors and log sample sizes, thereby making the model with discounting the best. Also, the posterior standard deviations of the finite population means of the different areas under the model with discounting are at least as similar to those from the other models, better than the ps only model.

**Table 1: Gibbs sampler diagnostics for the model with discounting using the BMI data of the eight counties**

| Parameter | $n_1$ | $n_2$ | Pval | ESS |
|---|---|---|---|---|
| $a_1$ | 140 | 24 | 0.804 | 1000 |
| $a_2$ | 138 | 38 | 0.750 | 1000 |
| $a_3$ | 667 | 128 | 0.395 | 1000 |
| $a_4$ | 133 | 29 | 0.709 | 1000 |
| $a_5$ | 96 | 29 | 0.813 | 1000 |
| $a_6$ | 119 | 22 | 0.144 | 1000 |
| $a_7$ | 100 | 28 | 0.332 | 884 |
| $a_8$ | 137 | 39 | 0.447 | 1000 |
| $\rho$ | - | - | 0.465 | 1000 |
| $\theta$ | - | - | 0.886 | 541 |
| $\gamma$ | - | - | 0.473 | 545 |

NOTE: Pval is the $p$-value of the Geweke test and ESS is the effective sample size of the Gibbs sampler

**Table 2: Comparison of five models using BMI data of eight counties**

| Model | LPML | DIC | BPP | DM | PRMSE |
|---|---|---|---|---|---|
| Discounting | 977.491<br>(0.8) | 1946.369<br>(1.3) | 0.553<br>(-) | 2.626<br>(-2.0) | 1.606<br>(-52.4) |
| Logit | 975.866<br>(0.7) | 1943.152<br>(1.1) | 0.528<br>(-) | 2.623<br>(-2.2) | 1.783<br>(-47.1) |
| No discounting | 1235.930<br>(27.5) | 2472.066<br>(28.6) | 1.000<br>(-) | 2.616<br>(-2.5) | 1.718<br>(-49.1) |
| No nps weights | 978.573<br>(0.9) | 1948.031<br>(1.3) | .541<br>(-) | 2.597<br>(-3.1) | 1.521<br>(-54.9) |
| PS only | 969.371 | 1922.219 | 0.493 | 2.682 | 3.373 |

NOTE: For PRMSE, the true value is taken to be the weighted average of all BMI values. The model with discounting is the one described, the logit model regresses the $a_i$ on the logarithm of sample sizes, and the model without discounting has all $a_i$ set to unity. The measures are calculated for PS data only. Gibbs sampling is needed for the models with discounting. Wang *et al.* (2011) has the divergence measure (DM). The parenthesis $(\cdot)$ shows the percent each measure is larger than the one for the ps. The model with discounting has PRMSE 9.9% smaller than the logit model.

Table 3 has posterior inference about the discounting factors. There are some discrimination among the small areas as the $a_i$ range from .066 to .141. The posterior standard deviations are small making the CVs standing between .102 and .160 and so the inference is very precise and reliable. Consequently, the 95% HPDIs are reasonably tight. Therefore, as there is much discounting, the $a_i$ are playing a consequential role in this application. Nandram and Rao (2021, 2023) gave interpretations of the discounting factor for a single area.

For comparisons, we use the following idea in Tables 4 & 5. For two standard deviations, $a, b$, assuming independence, $\max(a, b) \leq \sqrt{(a^2 + b^2)} \leq a + b$. That is, assuming independence of two sources, the standard deviation of the difference is at least the larger one.

In Table 4, we compare inference about the finite population means using integrated data and the probability sample only (ps only model). Note that the data from the nps are not used in the ps model only; see Appendix D for a discussion of the ps only model. As expected, there are large gains in precision over the ps only model when the model with discounting is used. Three of the PMs under the model with discounting are smaller than the corresponding ones under ps only model. Therefore, there is possibly some selection bias in the model with discounting. The 95% HPDIs for the nps have considerable overlaps on

**Table 3: Posterior summaries of the discounting factors for BMI data of eight counties**

| County | $n_1$ | $n_2$ | PM | PSD | NSE | CV | 95% HPDI |
|---|---|---|---|---|---|---|---|
| 1 | 140 | 24 | 0.130 | 0.019 | 0.001 | 0.147 | (0.097, 0.171) |
| 2 | 138 | 38 | 0.066 | 0.010 | 0.000 | 0.160 | (0.043, 0.085) |
| 3 | 667 | 128 | 0.111 | 0.011 | 0.000 | 0.102 | (0.091, 0.132) |
| 4 | 133 | 29 | 0.112 | 0.017 | 0.001 | 0.149 | (0.081, 0.146) |
| 5 | 96 | 29 | 0.095 | 0.015 | 0.001 | 0.158 | (0.069, 0.130) |
| 6 | 119 | 22 | 0.141 | 0.022 | 0.001 | 0.155 | (0.101, 0.184) |
| 7 | 100 | 28 | 0.101 | 0.016 | 0.001 | 0.160 | (0.071, 0.131) |
| 8 | 137 | 39 | 0.099 | 0.015 | 0.000 | 0.148 | (0.071, 0.126) |

NOTE: The discounting factors, $a_i$, are small.

the right to those of the ps. Therefore, there is not much difference between the two models in terms of PMs.

In Table 5, we compare inference about the $85^{th}$ percentile of the finite population using the model with discounting and the probability sample only. Again, as expected, there are large gains in precision when the model with discounting is used. Three of the PMs under the model with discounting are smaller than the corresponding ones under ps model only. For each area, the intervals under the nps overlap considerably on the right of those for the ps. Therefore, again there is possibly some selection bias in the model with discounting. There are similar results for the $90^{th}$ and $95^{th}$ percentiles (not shown) with much larger variability, of course.

In Figures 1 & 2 we show plots of the posterior densities of the finite population means by county. For all counties, the model with discounting gives more precise estimates than the ps only model, and the plots overlap with various degrees, with the plot of the nps to the right of the ps, indicating some degree of selection bias remaining; five counties (2, 3, 4, 5, 7), appear similar. There appears to be no differences in sample size except for county 3 with very large county size (667, 128).

**Table 4: Comparison of the nps model (with discounting) and the ps only model via posterior summaries of the finite population mean of the eight counties using the BMI data**

| County | $n_1$ | $n_2$ | Model | PM | PSD | NSE | CV | 95% HPDI |
|---|---|---|---|---|---|---|---|---|
| 1 | 140 | 24 | nps | 26.193* | 0.293 | 0.009 | 0.011 | (25.559, 26.704) |
| 1 | 140 | 24 | ps | 25.229 | 0.450 | 0.014 | 0.018 | (24.437 26.199) |
| 2 | 138 | 38 | nps | 27.483* | 0.299 | 0.010 | 0.011 | (26.908, 28.052) |
| 2 | 138 | 38 | ps | 27.100 | 0.363 | 0.010 | 0.013 | (26.393 27.740) |
| 3 | 667 | 128 | nps | 26.931* | 0.149 | 0.006 | 0.005 | (26.642, 27.219) |
| 3 | 667 | 128 | ps | 26.769 | 0.222 | 0.006 | 0.008 | (26.346 27.204) |
| 4 | 133 | 29 | nps | 26.299 | 0.364 | 0.010 | 0.014 | (25.593, 26.951) |
| 4 | 133 | 29 | ps | 26.481 | 0.878 | 0.026 | 0.033 | (24.535 28.090) |
| 5 | 96 | 29 | nps | 27.017 | 0.355 | 0.011 | 0.013 | (26.290, 27.652) |
| 5 | 96 | 29 | ps | 27.416 | 0.521 | 0.017 | 0.019 | (26.356 28.339) |
| 6 | 119 | 22 | nps | 26.352* | 0.299 | 0.008 | 0.011 | (25.841, 26.954) |
| 6 | 119 | 22 | ps | 25.102 | 0.469 | 0.013 | 0.019 | (24.100 25.939) |
| 7 | 100 | 28 | nps | 26.845* | 0.305 | 0.010 | 0.011 | (26.253, 27.389) |
| 7 | 100 | 28 | ps | 26.467 | 0.416 | 0.014 | 0.016 | (25.720 27.297) |
| 8 | 137 | 39 | nps | 27.350 | 0.295 | 0.012 | 0.011 | (26.789, 27.930) |
| 8 | 137 | 39 | ps | 28.406 | 0.457 | 0.013 | 0.016 | (27.530 29.276) |

NOTE: Posterior inference is based on 1000 iterates that provide posterior mean, PM, posterior standard deviation, PSD, numerical standard error, NSE, coefficient of variation, CV, and 95% highest posterior density interval, HPDI. The PMs of the model with data discounting are larger than those under the PS only model by $3.8, 1.4, .6, -.7, -1.5, 5.0, 1.4, -3.7$ percent. PMs are larger for counties marked $(*)$.

**Table 5: Comparison of the nps model (with discounting) and the ps only model via posterior summaries of the finite population $85^{th}$ percentile of the eight counties using the BMI data**

| County | $n_1$ | $n_2$ | Model | PM | PSD | NSE | CV | 95% HPDI |
|---|---|---|---|---|---|---|---|---|
| 1 | 140 | 24 | nps | 27.762* | 0.311 | 0.011 | 0.011 | (27.126, 28.309) |
| 1 | 140 | 24 | ps | 26.724 | 0.455 | 0.015 | 0.017 | (25.856, 27.649) |
| 2 | 138 | 38 | nps | 29.036* | 0.318 | 0.010 | 0.011 | (28.392, 29.625) |
| 2 | 138 | 38 | ps | 28.574 | 0.376 | 0.012 | 0.013 | (27.846, 29.290) |
| 3 | 667 | 128 | nps | 28.490* | 0.169 | 0.006 | 0.006 | (28.128, 28.790) |
| 3 | 667 | 128 | ps | 28.255 | 0.235 | 0.006 | 0.008 | (27.836, 28.774) |
| 4 | 133 | 29 | nps | 27.859 | 0.378 | 0.011 | 0.014 | (27.105, 28.553) |
| 4 | 133 | 29 | ps | 27.955 | 0.827 | 0.022 | 0.030 | (26.149, 29.415) |
| 5 | 96 | 29 | nps | 28.580 | 0.351 | 0.012 | 0.012 | (27.924, 29.302) |
| 5 | 96 | 29 | ps | 28.908 | 0.505 | 0.018 | 0.017 | (27.867, 29.300) |
| 6 | 119 | 22 | nps | 27.932* | 0.332 | 0.011 | 0.011 | (27.268, 28.553) |
| 6 | 119 | 22 | ps | 26.600 | 0.475 | 0.014 | 0.018 | (25.700, 27.540) |
| 7 | 100 | 28 | nps | 28.409* | 0.323 | 0.012 | 0.011 | (27.786, 29.013) |
| 7 | 100 | 28 | ps | 27.934 | 0.429 | 0.011 | 0.015 | (27.089, 28.756) |
| 8 | 137 | 39 | nps | 28.905 | 0.297 | 0.009 | 0.010 | (28.352, 29.494) |
| 8 | 137 | 39 | ps | 29.913 | 0.422 | 0.015 | 0.014 | (29.091, 30.726) |

NOTE: Posterior inference is based on 1000 iterates that provide PM, posterior mean, PSD, posterior standard deviation, $W$, width of the 95% HPD interval and CV, coefficient of variation. PMs are larger for counties marked ($*$).

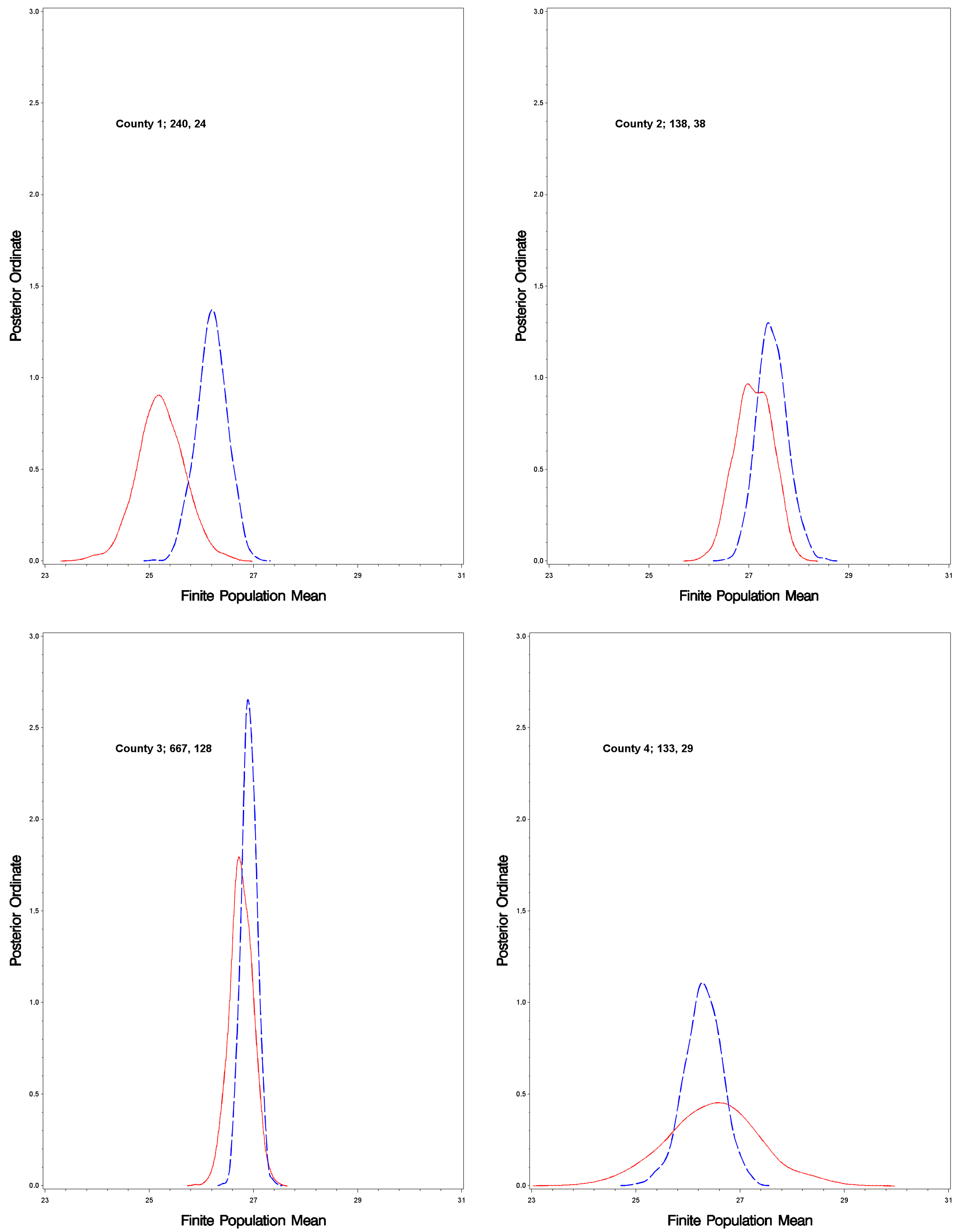


**Figure 1: Comparison for the posterior distributions of the finite population mean for nps and ps models by county (dashed: discounting model; solid: ps only model)**

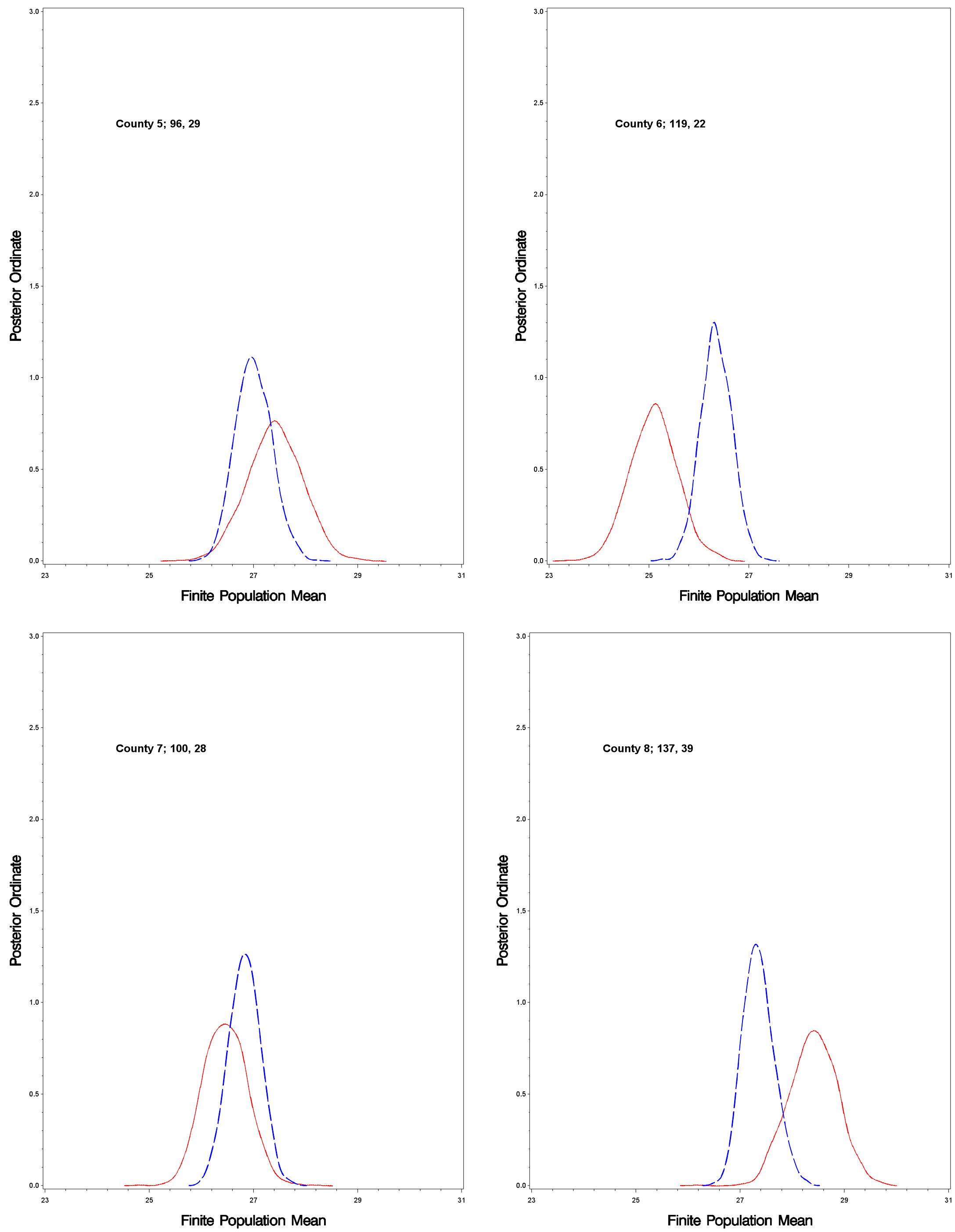


**Figure 2: Comparison for the posterior distributions of the finite population mean for nps and ps models by county (dashed: discounting model; solid: ps only model)**

## 5. Concluding remarks

This section has two subsections. The first subsection is a summary of the paper with general comments and the second subsection is on robustification of the models for the study variable and the participation variable.

### 5.1. Summary and comments

In our illustrative example on body mass index, our data-integrated model with discounting is preferred over the ps only model and other competitors. The logit data-integrated model is a strong competitor. The data-integrated model provides small area estimates, roughly similar to those of the ps only model, with larger precision. It is difficult to remove all biases completely. We outline some important problems we are currently working on, particularly how the assumptions on the participation variable and the study variable can be relaxed.

We have shown how to extend our approach to cover small area estimation. We have done so for the unit-level small area model (a bit less practical); this is an extension of Nandram and Rao (2021, 2023) to cover small areas. We have extended Toto and Nandram (2010) or Molina, Nandram and Rao (2014), who provided a Bayesian approach, to solve the problem without combining a nps and a ps. However, our work here was motivated by Beaumont (2020), Rao (2020) and Beaumont and Rao (2020) but these authors provided limited discussion of unit-level models; Beaumont and Rao (2020) showed how to use the area-level Fay-Herriot model to improve inference for the small areas in the ps, covariates being drawn from the nps (Big Data).

The assumption of normality on the BMI data is perhaps not a very good one because the BMI data are skewed (true for most continuous survey data) and discrete; see Yin and Nandram (2020 a,b) on how the Dirichlet process is used for BMI data without data integration. Also, more robust methods on propensity scores are needed. Stick-breaking priors can be used to provide more robust models, but these models are difficult to fit when all uncertainty is taken into account and this is on-going work; see Ishwaran and James (2001). It is also possible to use BART in data integration (*e.g.*, Rafei, *et al.* 2021). But BART is not a fully Bayesian procedure because it double-uses the data, it suffers from overshrinkage, and there is no underlying theory of BART (just a machine learning algorithm like random forest); see Hill, Linero and Murray (2020) for more detailed discussions and criticisms about BART. Yet, one does not need to express a relation between study variable and covariates; see Lockwood (2023, PhD Dissertation) for an important advance.

It is possible to avoid estimation of the survey weights of the non-probability sample by using a structural (measurement error) model; see Berg *et al.* (2021) for a start. We have been doing similar work at National Agriculural Statistics Service, USDA. For the nps (1), we consider

$$y_{1ij} \overset{ind}{\sim} \text{Normal}\left\{\gamma_0 + \gamma_1(\underline{x}'_{1ij}\underline{\beta} + \nu_i), \frac{\sigma^2}{a_i}\right\}, j = 1, \ldots, n_{1i}, i = 1, \ldots, \ell,$$

and for the ps (2),

$$y_{2ij} \overset{ind}{\sim} \text{Normal}\left\{\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{2ij}}\right\}, j = 1, \ldots, n_{2i}, i = 1, \ldots, \ell.$$

Here, $\gamma_0$ and $\gamma_1$ are weakly identified and can lead to poor performance of a Gibbs sampler. One can define the true values of $y_{2ij}$ as $\theta_{ij} = \underline{x}'_{2ij}\underline{\beta} + \nu_i$. We do not need to estimate nps weights. Note again that $n_{1i}$ is much larger than $n_{2i}$, and a discount factor is used to increase variability and help avoiding the nps to dominate the ps. Note that the parameters, $\underline{\beta}$, $\sigma^2$ and $\nu_i$ are the same in both the nps and the ps. Finally, a standard assumption on the area random effects is

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

Of course, this can be overcome using the Pitman-Yor stick breaking procedure. Because of non-identifiability issues, we will assume that $\gamma_0$ and $\gamma_1$ are independent with

$$\gamma_0 \sim \text{Uniform}(c_1, c_2), \gamma_1 \sim \text{Uniform}(d_1, d_2),$$

where $(c_1, c_2)$ and $(d_1, d_2)$ are to be specified using exploratory data analysis. This can be done by fitting $\bar{y}_{1i} = \gamma_0 + \gamma_1\bar{y}_{2i} + e_i, i = 1, \ldots, \ell$, and using the bootstrap distributions of the least squares estimators of $\gamma_0$ and $\gamma_1$ to get their ranges. For the $a_i$, we will assume the same prior as before, and we also assume that

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

Also, as before prediction is done by using

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

and the prediction procedure is similar to the one done earlier. For

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{\bar{X}}'_i\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}), i = 1, \ldots, \ell,$$

where $\underline{\bar{X}}_i = \frac{\sum_{j=1}^{N_i} \underline{x}_{ij}}{N_i}$ is unknown and $N_i$ may also be unknown. Design-based estimators of $N_i$ and $\underline{\bar{X}}_i$ are respectively $N_i = \sum_{i=1}^{n_{2i}} W_{2ij}$ and $\underline{\bar{X}}_i = \frac{\sum_{j=1}^{n_{2i}} W_{2ij}\underline{x}_{2ij}}{N_i}$ (Hajek or Horvitz-Thompson). Inference for finite population percentiles is also possible.

## 5.2. Robustification

Looking towards double robustness as in non-Bayesian methods, we can use a mixture model for the study variable and a t-link for the participation variable of any number of areas within the Bayesian paradigm.

### 5.2.1. Robustification of the model of the study variable

For the study variable, we use a three-component mixture model. For the non-probability sample,

$$f(y_{1ij} \mid \nu_i, \underline{\beta}, p, q, \rho, \gamma) = (1-p-q)\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\rho\gamma\sigma^2}{aw_{1ij}})$$

$$+p\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\gamma\sigma^2}{aw_{1ij}}) + q\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{aw_{1ij}}), i = 1, \ldots, n_{1i}.$$

and, for the probability sample, we have

$$f(y_{2ij} \mid \nu_i, \underline{\beta}, p, q, \rho, \gamma) = (1-p-q)\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\rho\gamma\sigma^2}{w_{2i}})$$

$$+p\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\gamma\sigma^2}{w_{2ij}}) + q\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{2ij}}), i = 1, \ldots, n_{2i}, i = 1, \ldots, \ell.$$

Finally,

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

It is also sensible to use the constraint $p > q$ and $0 < p, q, p+q, \rho, \gamma < 1$. In each case, the first component corresponds to ordinary observations, the second component corresponds to mild outliers and the third component to severe outliers. See Chakraborty, Datta, and Mandal (2019) for the much simpler two-component mixture model. There is on-going work on this topic.

### 5.2.2. Robustification of the model of the participation variable

We consider the following mixture model for the selection indicators, $r_i, i = 1, \ldots, N$, and we consider one large area (all areas combined). We make the robust assumption,

$$r_i \mid T = g, \underline{\theta} \overset{ind}{\sim} \text{Bernoulli}\{\mathcal{T}_{a_g}(\underline{z}'_i\underline{\theta})\}, i = 1, \ldots, N,$$

$$P(T = g \mid \lambda_g) = \lambda_g, g = 1, \ldots, G,$$

where $(a_g, \lambda_g), g = 1, \ldots, G$, and $G$ are to be specified. We define the propensity scores as

$$\pi_i = \sum_{g=1}^{G} \lambda_g \mathcal{T}_{a_g}(\underline{z}'_i\underline{\theta}), i = 1, \ldots, N.$$

We can now develop a pseudo-density for each $g$ and average all the pseudo-densities over $g$. Specifically, we have the mixture pseudo-density,

$$P(\underline{r} \mid \underline{z}, \underline{\theta}) = \sum_{g=1}^{G} \lambda_g \prod_{i=1}^{n_1} \left\{ \frac{\mathcal{T}_{a_g}(\underline{z}'_{1i}\underline{\theta})}{1 - \mathcal{T}_{a_g}(\underline{z}'_{1i}\underline{\theta})} \right\} \prod_{i=1}^{n_2} \left\{ 1 - \mathcal{T}_{a_g}(\underline{z}'_{2i}\underline{\theta}) \right\}^{W_{2i}}, \tag{4}$$

where $\mathcal{T}_{a_g}, g = 1, \ldots, G$, is the Student's $t$ cdf on $a_g$ degrees of freedom. The estimated propensity scores we need are then

$$\hat{\pi}_i = \sum_{g=1}^{G} \lambda_g \mathcal{T}_{a_g}(\underset{\sim}{z}'_{1i}\hat{\underset{\sim}{\theta}}), i = 1, \ldots, n_1,$$

where $\hat{\underset{\sim}{\theta}} = E(\underset{\sim}{\theta} \mid \underset{\sim}{r})$; it is possible to use other summaries as well (*e.g.*, the posterior median or the posterior mode).

This is a generalization of the logistic regression model, and it covers many cases (Cauchy, logistic and normal). It is well-known that when the Student's $t$ density and/or the logistic distribution are appropriately rescaled, a plot of the quantiles of the Student's t density on roughly 8 degrees of freedom versus the quantiles of the logistic distribution is almost a $45^o$ straight line through the origin. Here $\lambda_g, g = 1, \ldots, G$, are specified weights at degrees of freedom $a_g, g = 1, \ldots, G$, and to look at variation around the logistic distribution, we can place more probability at $a_g = 8$. For example, we have used $a_g = 1, 4, 8, 13, 20, 30, 40, 50$ for $G = 8$, $a_g = 40, 50$ will be close to a standard normal density, and $\lambda_g = .125, .125, .25, .125, .125, .125, .080, .045$. There is on-going work on this topic.

**Acknowledgments**

Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram) and J. N. K. Rao was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. Balgobin Nandram presented invited talks on this topic at the SAE2022 conference, University of Maryland, and at Banaras Hindu University, India, February 2023. The reviewers helped to improve the presentation.

## APPENDIX A: Propensity scores

Let $\underset{\sim}{x}_i, i = 1, \ldots, N$, denote the covariates; these are observed in the ps and the nps, but they are not observed for the rest of the population. Again, for the nps, we have $\underset{\sim}{x}_{1i}, i = 1, \ldots, n_1$, and for the ps, we have $\underset{\sim}{x}_{2i}, i = 1, \ldots, n_2$. Chen, Li and Wu (2020) has a method to get the propensity scores for the nps, and therefore the survey weights, which they defined as the reciprocals of the propensity scores. They assume that the propensity scores can be modeled parametrically using

$$\pi_i = P(r_i = 1 \mid \underset{\sim}{x}_i) = \pi(\underset{\sim}{x}_i; \underset{\sim}{\theta}),$$

with independence over $i$, where $\underset{\sim}{\theta}$ are to be estimated. Here $r_i = 1$ for the ps or nps; $r_i = 0$ for the nonsamples. Then, the likelihood function is

$$\ell(\underset{\sim}{\theta}) = \prod_{i=1}^{N} \{\pi(\underset{\sim}{x}_i; \underset{\sim}{\theta})\}^{r_i}\{1 - \pi(\underset{\sim}{x}_i; \underset{\sim}{\theta})\}^{1-r_i}.$$

The propensity scores are obtained in two steps.

First, they wrote the log-likelihood as

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

Second, they used the pseudo-log-likelihood by replacing the second term by the Horvitz-Thompson estimator since the nonsample $\underline{x}_i$ are unknown, as

$$\ell^*(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{x}_{1i};\underline{\theta})}{1-\pi(\underline{x}_{1i};\underline{\theta})}\right\} + \sum_{i=1}^{n_2} W_{2i}\log\{1-\pi(\underline{x}_{2i};\underline{\theta})\},$$

which can now be maximized for $\hat{\underline{\theta}}$. The propensity scores for the nps are then $\pi(\underline{x}_{1i};\hat{\underline{\theta}}), i = 1,\dots,n_1$. Henceforth, they specialize to logistic regression.

One caveat is that the propensity scores are not really selection probabilities (*i.e.*, quasi-randomization). This is true because the propensity scores must be obtained for the entire population (*i.e.*, all $N$ units) and then calibrated to the nps sample size. Only in this case, quasi-randomization makes any sense at all. This is still an open problem. Also, they assumed ignorability (given the covariates, the participation variable is independent of the study variable), but see Nandram (2022) for nonignorability. Chen, Li and Wu (2020) did not assume non-ignorability because they assumed that the study variable is missing in the probability sample; they need to mass impute the the missing values, but this is not in the spirit of their work.

## APPENDIX B: Computation for the small area model

We discuss how to fit the proposed model. Recall $\Omega_1 = (\underline{a},\theta,\gamma,\rho)$ and $\Omega_2 = (\underline{\nu},\underline{\beta},\sigma^2)$. Our strategy is to integrate out $\Omega_2$ from $\pi(\Omega_1,\Omega_2 \mid \underline{y})$ to get $\pi(\Omega_1 \mid \underline{y})$ and then sample $\pi(\Omega_1 \mid \underline{y})$ using the Griddy-Gibbs sampler (Ritter and Tanner, 1992).

For convenience, we will keep $a_{si}, s = 1,2, i = 1,\dots,\ell$, free in $(0,1)$ and sometimes $a_{1i} = a_i$ and $a_{2i} = 1$, $i = 1,\dots,\ell$. Then, letting $n = \sum_{s=1}^{2}\sum_{i=1}^{\ell} n_{si}$, the total number of observations,

$$\pi(\Omega_1,\Omega_2 \mid \underline{y}) \propto \pi(\Omega_1)\left(\prod_{i=1}^{\ell}\sqrt{a_i}\right)\times$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n+\ell}{2}+1}\left(\frac{1-\rho}{\rho}\right)^{\ell/2}\prod_{i=1}^{\ell}\left[e^{-\frac{1}{2\rho\sigma^2}\left\{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(y_{sij}-\underline{x}'_{sij}\underline{\beta}-\nu_i)^2+(1-\rho)\nu_i^2\right\}}\right]. \quad \text{(B.1)}$$

We will integrate out $\Omega_2$. Momentarily, we will drop $\pi(\Omega_1)$, but we will retain $\prod_{i=1}^{\ell}\sqrt{a_i}$.

Define the following quantities,

$$\lambda_i = \frac{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}+(1-\rho)}, \quad \phi_{sij} = \frac{a_{si}w_{sij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}},$$

$$\bar{\bar{y}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}}\phi_{sij}y_{sij}, \quad \bar{\bar{\underline{x}}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}}\phi_{sij}\underline{x}_{sij},$$

$$\tilde{y}_{sij} = y_{sij} - \bar{\bar{y}}_i, \quad \tilde{\underline{x}}_{sij} = \underline{x}_{sij} - \bar{\bar{\underline{x}}}_i.$$

Note that while the $\lambda_i$ are functions of $\rho$, but the $\phi_{sij}$, $\bar{\bar{y}}_i$ and $\bar{\bar{\underline{x}}}_i$ are not functions of $\rho$.

We can now rewrite the exponent in (B.1),

$$\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(y_{sij} - \underline{x}'_{sij}\underline{\beta} - \nu_i)^2 + \frac{1-\rho}{\rho}\nu_i^2\right\}\right\},$$

as

$$\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(\tilde{y}_{sij} - \tilde{\underline{x}}'_{sij}\underline{\beta})^2 + \frac{1-\rho}{\rho}(\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij})(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'\underline{\beta} - \nu_i)^2\right\}\right\}.$$

Then, it is easy to show that

$$\nu_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y} \overset{ind}{\sim} \text{Normal}\{\hat{\nu}_i, \frac{\rho}{1-\rho}\sigma^2(1-\lambda_i)\}, i = 1, \ldots, \ell,$$

where $\hat{\nu}_i = \lambda_i(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'_i\underline{\beta})$. This is a standard form in small area estimation and it combines both the probability sample and the non-probability sample over all areas; note the common $\underline{\beta}$ and $\sigma^2$.

Then, integrating out the $\nu_i$ from (B.1), we have

$$\pi(\underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \prod_{i=1}^{\ell} \sqrt{a_i(1-\lambda_i)}$$

$$\times \prod_{i=1}^{\ell}\left[\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(\tilde{y}_{sij} - \tilde{\underline{x}}'_{sij}\underline{\beta})^2 + P_i(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'_i\underline{\beta})^2\right\}\right\}\right], \tag{B.2}$$

where

$$P_i = \left(\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\right)(1-\lambda_i)^2 + \frac{1-\rho}{\rho}\lambda_i^2, i = 1, \ldots, \ell.$$

Then,

$$\underline{\beta} \mid \sigma^2, \rho, \underline{y} \sim \text{Normal}\{\hat{\underline{\beta}}, \sigma^2\Delta\},$$

where

$$\Delta = \left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{x}'_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{\underline{x}}}'_i\right\}^{-1}$$

and

$$\hat{\underline{\beta}} = \left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{x}'_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{\underline{x}}}'_i\right\}^{-1}\left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{y}_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{y}}_i\right\}.$$

Then integrating $\underline{\beta}$ from (B.2), we have

$$\pi(\sigma^2, \rho \mid \underline{y}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}+1} \mid \Delta \mid^{1/2} \prod_{i=1}^{\ell} \sqrt{a_i(1-\lambda_i)}$$

$$\times e^{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\{\tilde{y}_{sij}-\underline{\tilde{x}}'_{sij}\underline{\hat{\beta}}\}^2+\sum_{i=1}^{\ell} P_i(\bar{\bar{y}}_i-\underline{\bar{\bar{x}}}'_i\underline{\hat{\beta}})^2\right\}}. \tag{B.3}$$

Therefore,

$$\sigma^2 \mid \rho, \underline{y} \sim \text{InvGam}\left\{\frac{n-p}{2}, \frac{Q}{2}\right\}, \tag{B.4}$$

where

$$Q = \sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\{\tilde{y}_{sij} - \underline{\tilde{x}}'_{sij}\underline{\hat{\beta}}\}^2 + \sum_{i=1}^{\ell} P_i(\bar{\bar{y}}_i - \underline{\bar{\bar{x}}}'_i\underline{\hat{\beta}})^2.$$

Integrating out $\sigma^2$ from (B.3), we have

$$\pi(\rho \mid \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell} \sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}}, 0 \le \rho \le 1. \tag{B.5}$$

Actually $\pi(\rho \mid \Omega_1, \underline{y})$ is defined for all values of $\rho$ in $[0,1]$ because the $P_i$ and $\lambda_i$ are well defined for all values of $\rho$ in $[0,1]$. Note that the $a_i$ are constants (given) above, specifically they are constants in (B.5).

Bringing back $\pi(\Omega_1)$ into the picture, we have

$$\pi(\Omega_1 \mid \underline{y}) \propto \pi(\Omega_1)\pi(\rho \mid \underline{y}),$$

and therefore,

$$\pi(\Omega_1 \mid \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell} \sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}} \left\{\prod_{i=1}^{\ell} \frac{a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}}{B\{\theta(\frac{1-\gamma}{\gamma}), (1-\theta)(\frac{1-\gamma}{\gamma})\}}\right\}, \tag{B.6}$$

$\frac{\gamma}{1-\gamma} \le \theta \le \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3, 0 \le \rho \le 1$. It is worth noting that the $a_i$ are not independent; $\Delta$ and $Q$ contain all the $a_i$, which is contained by $\lambda_i$ also.

In (B.6), $\pi(\Omega_1 \mid \underline{y})$ is well defined for all values of $\underline{a}, \theta, \gamma, \rho$ because $0 < a_i < 1, i = 1, \ldots, \ell$, $0 < \rho < 1$, $0 < \gamma < \frac{1}{3}$, $\frac{\gamma}{1-\gamma} \le \theta \le \frac{1-2\gamma}{1-\gamma}$. Therefore, it follows that the joint posterior density $\pi(\Omega_1, \Omega_2 \mid \underline{y})$ is proper. Next, we present the rather obvious conditional posterior densities (CPDs) necessary to run the Gibbs sampler.

First, we consider the CPD of the $a_i, i = 1, \ldots, \ell$. Letting $\underline{a}_{(i)} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_\ell)', i = 1, \ldots, \ell$ ($a_i$ is eliminated), then for $0 < a_i < 1$,

$$\pi(a_i \mid \underline{a}_{(i)}, \rho, \theta, \gamma, \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell} \sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}} \left\{\prod_{i=1}^{\ell} a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}\right\}. \tag{B.7}$$

Second, the CPD of $\rho$ is

$$\pi(\rho \mid \underline{a}, \theta, \gamma, \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell} \sqrt{(1-\lambda_i)}}{Q^{(n-p)/2}}, 0 < \rho < 1. \tag{B.8}$$

Third, the joint CPD of $(\theta, \gamma)$ is

$$\pi(\theta, \gamma \mid \underline{a}, \rho, \underline{y}) \propto \left\{ \prod_{i=1}^{\ell} \frac{a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}}{B\{\theta(\frac{1-\gamma}{\gamma}), (1-\theta)(\frac{1-\gamma}{\gamma})\}} \right\}, \frac{\gamma}{1-\gamma} \leq \theta \leq \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3. \tag{B.9}$$

$\frac{\gamma}{1-\gamma} \leq \theta \leq \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3$. The CPD of $\theta$ or $\gamma$ is easy to write down.

We note that all the CPDs are nonstandard, but all the parameters lie in $(0, 1)$, so we have used a grid method, with 100 grid points, to sample each of the CPDs. The number grid points can be reduced for the $a_i$ perhaps to 50 or so, but we need the number grid points to be around 100 for $(\rho, \theta, \gamma)$; hyperparameters are more difficult to sample. We have done this, and we have reduced the entire computation time from roughly 40 minutes to 20 minutes with little change in the results.

## APPENDIX C: Bayesian model diagnostics and measures

We test concordance of the ps (2) part of the model,

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

with the observed data of the ps (2). It is not sensible to study concordance with the observed data of the nps (1) because they are biased. The posterior density of $(\underline{\nu}, \underline{\beta}, \sigma^2)$ comes from their respective models. We describe five Bayesian measures, which are the negative log-pseudo marginal likelihood (LPML), the deviance information criterion (DIC), the Bayesian predictive p-value (BPP), the divergence measure (DM) and the posterior root mean squared error (PRMSE); LPML and DM are based on Bayesian cross-validation.

First, the conditional posterior ordinate is $CPO_{ij} = f(y_{2ij} \mid \underline{y}_{(2ij)})$, where $\underline{y}_{(2ij)}$ is the vector of all values excluding $y_{(2ij)}$. Let $\underline{\Omega} = (\underline{\nu}', \underline{\beta}', \sigma^2)'$ and $\underline{\Omega}^{(h)}$ denote the $h^{th}$ iterate from the Gibbs sampler of the appropriate parameters. Then, $CPO_{ij}$ is usually estimated by

$$\widehat{CPO}_{ij} = \left[ \frac{1}{M} \sum_{h=1}^{M} \frac{1}{f(y_{2ij} \mid \underline{\Omega}^{(h)})} \right]^{-1},$$

the harmonic mean, and $LPML = \sum_{i=1}^{\ell} \sum_{j=1}^{n_{2i}} \log(\widehat{CPO}_{ij})$.

Second, letting $\hat{\underline{\Omega}}$ denote the posterior mean of $\underline{\Omega}$, the DIC is

$$DIC = 2\hat{D}(\underline{y}) - D(\underline{y} \mid \hat{\underline{\Omega}}),$$

where $D(\underline{y} \mid \underline{\Omega}) = -2 \log\{f(\underline{y} \mid \underline{\Omega})\}$ and $\hat{D}(\underline{y}) = \frac{1}{M} \sum_{h=1}^{M} D(\underline{y} \mid \underline{\Omega}^{(h)})$.

Third, letting $T_2$ denote a test (discrepancy) function, the BPP is

$$P(T_2^{rep} > T_2^{obs} \mid \underline{y}^{obs}),$$

where we have used

$$T_2 = \sum_{i=1}^{\ell} \sum_{j=1}^{n_{2i}} W_{2ij} \frac{(y_{2ij} - \underline{x}'_{2ij}\underline{\beta} - \nu_i)^2}{\sigma^2}.$$

Fourth, the divergence measure is

$$DM = \frac{1}{\sum_{i=1}^{\ell} n_{2i}} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} |y_{2ij} - E(y_{2ij}|\underline{y}_{(2ij)})|;$$

see Wang *et al.* (2012).

Fifth, letting $T = \sum_{s=1}^{2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_{si}} W_{sij} y_{sij} / \sum_{s=1}^{2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_{si}} W_{sij}$,

$$PRMSE = \sqrt{\sum_{i=1}^{\ell} \{(PM_{2i} - T)^2 + PSD_{2i}^2\}},$$

where $PM_{2i} = \mathrm{E}(\bar{Y}_{2i}|\underline{y}_1, \underline{y}_2)$ and $PSD_{2i}^2 = \mathrm{Var}(\bar{Y}_{2i}|\underline{y}_1, \underline{y}_2)$.

## APPENDIX D: Adding survey weights into the bayesian BHF model

We describe how to fit the ps only model. This is essentially adding survey weights to the BHF model.

The population model is

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

where $\underline{x}_{ij}$ has $p$ components, including an intercept, and

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

The reparameterization with respect to $\rho$ is similar, but slightly different, to the one we have used before. The correlation of the values within an area is $\rho$, and the model is defined for all values of $\rho$ in $[0, 1]$. Let $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$, the finite population mean of the $i^{th}$ area, and let $\underline{\bar{X}}_i$ denote the finite population mean covariate vector.

Therefore,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \rho \overset{ind}{\sim} \mathrm{Normal}\{\underline{\bar{X}}'_i\underline{\beta} + \nu_i, \frac{(1-\rho)\sigma^2}{N_i}\}.$$

Then, integrating out the $\nu_i$, we have

$$\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho \overset{ind}{\sim} \mathrm{Normal}\{\underline{\bar{X}}'_i\underline{\beta}, \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{N_i}\}.$$

So that $\sigma^2$ has a direct impact in prediction even when the $N_i$ are very large, and $\rho$ plays an important role here. This is different from the case when there is just a single sample (*i.e.*, no random effects), where for large $N_i$, the variance is approximately 0.

The sample model is

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

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

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

Letting $n = \sum_{i=1}^{\ell} n_i$, the total number of observations over the $\ell$ small areas, the joint posterior density is

$$\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{(n+\ell)/2+1}} \frac{1}{(\rho)^{\ell/2}} \frac{1}{(1-\rho)^{n/2}}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho \sum_{j=1}^{n_i} w_{ij}(y_{ij} - \underline{x}'_{ij}\underline{\beta} - \nu_i)^2 + (1-\rho)\nu_i^2\right\}\right]. \tag{D.1}$$

We will decompose $\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y})$ as

$$\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y}) = \pi_1(\underline{\nu} \mid \underline{\beta}, \sigma^2, \rho, \underline{y})\pi_2(\underline{\beta} \mid \sigma^2, \rho, \underline{y})\pi_3(\sigma^2 \mid \rho, \underline{y})\pi_4(\rho \mid \underline{y}),$$

where $\pi_1(\underline{\nu} \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$, $\pi_2(\underline{\beta} \mid \sigma^2, \rho, \underline{y})$, $\pi_3(\sigma^2 \mid \rho, \underline{y})$, except $\pi_4(\rho \mid \underline{y})$, are all in standard forms. Next, we will demonstrate this decomposition, and at the same time, we will prove propriety of the joint posterior density.

For $i = 1, \ldots, \ell$, let $\bar{\underline{x}}_i = \frac{\sum_{j=1}^{n_i} w_{ij}\underline{x}_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$, $\bar{y}_i = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$, and $\lambda_i = \frac{\rho\sum_{j=1}^{n_i} w_{ij}}{\rho\sum_{j=1}^{n_i} w_{ij}+1-\rho}$. Note that the $\lambda_i$ are not functions of $\sigma^2$. Then, it is easy to show that

$$\nu_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y} \overset{ind}{\sim} \text{Normal}\{\hat{\nu}_i, (1-\lambda_i)\rho\sigma^2\}, i = 1, \ldots, \ell,$$

where $\hat{\nu}_i = \lambda_i(\bar{y}_i - \bar{\underline{x}}'_i\underline{\beta})$.

Let $t_{ij} = y_{ij} - \lambda_i\bar{y}_i$ and $\underline{d}_{ij} = \underline{x}_{ij} - \lambda_i\bar{\underline{x}}_i, i = 1, \ldots, \ell$. Then, integrating $\nu_i$ from (D.1), we have

$$\pi(\underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{n/2+1}} \frac{1}{(1-\rho)^{n/2}} \prod_{i=1}^{\ell} \sqrt{1-\lambda_i}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\underline{\beta})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \bar{\underline{x}}'_i\underline{\beta})^2\right\}\right]. \tag{D.2}$$

Now, let

$$\hat{\underline{\beta}} = \Delta\left\{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}\underline{d}_{ij}t_{ij} + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2\bar{\underline{x}}_i y_i\right\},$$

where

$$\Delta^{-1} = \rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}\underline{d}_{ij}\underline{d}'_{ij} + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2\bar{\underline{x}}_i\bar{\underline{x}}'_i.$$

Note that $\hat{\beta}$ does not depend on $\sigma^2$. Then, it is easy to show that

$$\underline{\beta} \mid \sigma^2, \rho, \underline{y} \sim \text{Normal}(\hat{\underline{\beta}}, \rho(1-\rho)\sigma^2\Delta).$$

Now, integrating $\underline{\beta}$ from (D.2), we have

$$\pi(\sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{(n-p)/2+1}} \frac{|\Delta|}{(1-\rho)^{(n-p)/2}} \rho^{p/2} \prod_{i=1}^{\ell} \sqrt{1-\lambda_i}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\hat{\underline{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\hat{\underline{\beta}})^2\right\}\right]. \qquad \text{(D.3)}$$

Finally, it follows easily that

$$\sigma^2 \mid \rho, \underline{y} \sim \text{InvGam}\left\{\frac{n-p}{2}, \frac{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\hat{\underline{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\hat{\underline{\beta}})^2}{2\rho(1-\rho)}\right\}$$

and integrating $\sigma^2$ from (D.3), we have

$$\pi(\rho \mid \underline{y}) \propto \prod_{i=1}^{\ell} \frac{1}{(\rho\sum_{j=1}^{n_i} w_{ij} + 1 - \rho)^{1/2}}$$

$$\times \frac{\rho^{n/2}(1-\rho)^{\ell/2}|\Delta|^{1/2}}{\{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\hat{\underline{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\hat{\underline{\beta}})^2\}^{(n-p)/2}}, 0 < \rho < 1. \qquad \text{(D.4)}$$

Note that $\pi(\rho \mid \underline{y})$ is defined for all values of $\rho \in [0,1]$; we only need $\Delta$ to be well defined, and this is true because $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} \underline{d}_{ij}\underline{d}'_{ij}$ is full rank for all values of $\rho$ (*i.e.*, the matrix $(\underline{x}'_{ij})$ is full rank provided that it has at least $p$ linearly independent rows). Of course, $n > p$ as in standard regression problems. Therefore, the joint posterior density is proper.

## References

Berg, E., Im, J., Zhu, Z., Lewis-Beck, C., and Li, J. (2021). Integration of statistical and administrative agricultural data from Namibia. *Statistical Journal of the IAOS*, **37**, 557–578, DOI: 10.3233/SJI-200634.

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.

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.

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.

Chambers, R. L., Fabrizi, E., and Salvati, N. (2019). Small area estimation with linked data. *arXiv:1904.0036v1 [Stat.ME], 31 March 2019*, pg. 1-29.

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.

Chen, S., Yang, S., and Kim, J. K. (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology.* **10**, 1-24, DOI: 10.1093/jssam/smaa036.

Elliott, M. N. and A. Haviland (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, **33**, 211–215.

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

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Applications*, **7**, 251-278, DOI: 10.1146/annurev-statistics-031219-041110.

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

Ibrahim, J. G., Chen, M-H., Gwon, Y., and Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, **34**, 3724-3749, DOI: 10.1002/sim.6728.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-173.

Kim, J. K., 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.

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.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rjnzivillo, S., Pappalardo, L., and Gabrjelli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, **31**, 263-281.

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.

Meng, X-L (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, **12**, 685–726, DOI: 10.1214/18-AOAS1161SF.

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.

Muhyi, F. A., Sartono, B., Sulvianti, I. D., and Kurnia, A. (2019). Twitter utilization in application of small area estimation to estimate electability of candidate central java governor. *IOP Conference Series in Earth and Environmental Science*, **299** 012033, 1-10.

Nandram, B. (2022). A Bayesian assessment of non-ignorable selection of a non-probability Sample. *Indian Bayesians' News Letter, Invited Paper*, 1-16.

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.

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 small 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. (2023). Overcoming challenges associated with early bayesian state estimation of planted acres in the United States. Special Proceedings: Society of Statistics, Computing and Applications, ISBN #: 978-81-950383-2-9, 25th Annual Conference, 15-17 February 2023; pp 51-78.

Porter, A. T., Holan, S. H., Wikle, C. K., and Cressie, N. (2014). Spatial Fay-Herriot model for small area estimation with functional covariates. *Spatial Statistics*, **10**, 27-42.

Potthoff, R. F., Woodbury, M. A., and Manton, K. G. (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.

Rafei, A., Flannagan,C. A. C., West, B. T., and Elliott, M. R. (2022). Robust bayesian inference for big data: Combining sensor-based records with traditional survey. *arxiv:2101.07456V2 [Stat.ME], 26 March 2022*, pp. 1-58.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya*, Series B, 3-33, DOI: 10.1007/s13571-020-00227-w.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, **87**, 861-868, DOI: 10.2307/2290225.

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: 10.1093/jssam/smad041.

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*, 00,1-35, DOI: 10.1093/jssam/smad041.

Schmid, T., Bruckschen, F., Salvati, N., and Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society*, Series A, **180**, 1163-1190, DOI: 10.1111/rssa.12305.

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.

Yin, J. and Nandram, B. (2020a). A bayesian small area model with dirichlet processes on responses. *Statistics in Transition, New Series*, **21**, 1-19, DOI: 10.21307/stattrans-2020-041.

Yin, J. and Nandram, B. (2020b). A nonparametric bayesian analysis of response data with gaps, outliers and ties. *Statistics and Applications, New Series*, **18**, 121-141, ISSN 2452-7395(online).

Walker, A. M. (1968). A note on the asymptotic distribution of the sample quantiles. *Journal of the Royal Statistical Society*, Series B, **30**, 570-575.

Wang, J. C., Scott, H. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 84-106, D0I: 10.107/513253-011-0067-5.

Wang, Z., Kim, J. K., and Yang, S. (2018). Approximate bayesian inference under informative sampling. *Biometrika*, **105**, 91-10, DOI: 10.1093/biomet/asx073.

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.