A Bayesian Approach for Integrating a Small Probability Sample with a Non-probability Sample

Balgobin Nandram* J. N. K. Rao†

Abstract

We consider the problem of integrating a small probability sample (ps) and a non-probability sample (nps). By definition, for the nps, there are no survey weights, but for the ps, there are survey weights. The key issue is that the nps, although much larger than the ps, can lead to a biased estimator of a finite population quantity but with much smaller variance. We begin with a relatively simple problem in which the population is assumed to be homogeneous and there are no common units in the ps and the nps. We assume that there are covariates and responses for everyone in the two samples, and there are no covariates available for nonsampled units. We use the nps to construct a prior for the ps, particularly in small area estimation. We also introduce partial discounting to avoid a dominance of the prior. Inverse probability weighted estimators are used to do Bayesian predictive inference of the finite population mean. We show how to extend our procedure to cover estimation for small areas, where auxiliary information is much needed, and the nps is used as the prior with partial discounting. In our illustrative example on body mass index and our simulation study, we compare our procedures with inference from the ps only estimate. Our procedure provides improved estimates over the ps only estimate.

Key Words: Big data, Covariates, Finite population mean, Inverse probability weighting, Power prior, Selection bias, Small area, Surrogate samples

1. Introduction

Undoubtedly, probability sampling is the gold standard among all data collection procedures. It is based on randomization, and leads to unbiased and consistent estimates when a proper estimation procedure is implemented. Yet probability sampling schemes pose difficulties because of high nonresponse rates causing them to lose their much-needed probabilistic structure. On the other hand, nonprobability samples lack this probabilistic structure and hence are grossly inaccurate. For one thing, they are not representative of the population from which they are drawn. However, nonprobability samples are easy to collect and therefore very cheap to run. Faced with expensive surveys, government agencies are left with no choice but to enormously reduce efforts and costs in planning and fielding probability samples. Citro (2014) mentioned seven challenges for official statistics and stated, "In my view, to respond adequately to one or more, let alone all seven, of these challenges, official statistical offices need to move from the probability sample survey paradigm of the past 75 years to a mixed data source paradigm for the future."

*Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609
†School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada

It is not clear whether nonprobability sampling can replace probability sampling but the situation is threatening because of time, cost and nonresponse constraints (e.g., Beaumont, 2020; Rao 2020). Therefore, it is important to join scientists in trying to find a resolution to this problem. Beaumont (2020) reviewed some approaches that can reduce, or even eliminate, the use of probability surveys while preserving valid statistical inference. All his approaches use data from nonprobability samples, but in most approaches probability samples are also used. He particularly discussed the design-based approach based on probability sample, which is nonparametric, and therefore is not subject to the risk of bias due to a mis-specified model, but it can be inefficient. Rao (2020) reviewed various probability survey methods that are used to make valid inferences about finite population parameters. This allowed him to show how these models can be extended to nonprobability samples that can lead to valid inferences by themselves or when combined with probability samples.

It is possible to make inference about a finite population quantity using a single nonprobability sample only; see Rao (2020) for a discussion. One approach uses poststratification via multi-level regression and poststratification (MRP); see Wang et al. (2015). There are at least three problems with MRP and they are sparseness of the poststrata, weak covariates and nonexchangeability of poststrata effects. Therefore, it is unlikely for MRP to eliminate the selection bias inherent in nonprobability samples; see Valliant (2020) who demonstrated that MRP does not work well for nonprobability samples when there are weak covariates. Another approach is to use propensity scores to produce nps survey weights and then proceed as in a regular probability sample; see Elliott and Valliant (2017) for an informative review of quasi-randomization and the super-population approach. Chen, Li and Wu (2020) supplemented a nonprobability sample with a probability sample to estimate propensity scores. A full Bayesian approach in this direction is given by Nandram, Cao, Xu and Bhadra (2019). A third approach is to use a nonignorable selection model to remove the selection bias; see Smith (1983) for pioneering work in this direction. Xu and Nandram (2019) and Xu (2020) used this approach to obtain full Bayesian analyses. The references in these papers provide a historical development of this area. It is difficult to make efficient inference from a nonprobability sample with considerable selection bias. From our investigation, this appears to be correct. After all, a probability sample is the gold standard (high quality), but a nonprobability sample has low quality (large bias, large mean squared error but unrealistically small variance). In this paper, within the Bayesian framework, we use both the quasi-randomization approach and the super-population approach.

The key problem of a non-probability sample is that it is very likely to lead to seriously biased estimates of finite population quantities. Therefore, the large well-documented literature on selection bias is pertinent in the study of non-probability samples; these articles are too numerous to mention here. But see Xu, Nandram and Manandhar (2020) and Choi, Nandram and Kim (2021) for recent applications, and the references therein.

It becomes necessary to combine the two sampling processes. There are efforts to combine both probability and nonprobability samples to produce a single inference that compensates for the limitations of each process. 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. Elliott (2009) proposed a pseudo-design-based estimation procedure that uses a probability sample to estimate pseudo-inclusion probabilities for elements of a nonprobability sample. Both samples are then combined to derive estimates that are shown to have improved accuracy and smaller MSE compared with those derived from a probability-only sample. See Elliott and Valliant 2017) for an informative review. A limitation of these studies is the necessity of a large probability sample to produce robust calibration weights or pseudo-inclusion probabilities.

Sakshaug, Wisniowski, Ruiz and Blom (2019), hencefort SWRB, and Wisnioski, Sakshaug, Ruiz and Blom (2020), henceforth WSRB, gave very clear comparisons of probability sample and nonprobability sample. SWRB stated, "Given the advantages of both sampling schemes, it makes sense to devise a strategy to combine them in a way that is beneficial from both a cost and error perspective." In their conclusion, SWRB stated, "In conclusion, it is interesting to know 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." We are in concordance with these authors.

There are many approaches to make inference about a population using a nonprobability sample only or a nonprobability sample integrated with a probability sample. In the latter case, it is not clear which should be used to supplement the other. SWRB and WSRB, using the Bayesian approach, supported the situation where the nonprobability sample should be used to supplement the probability sample. We concur with these authors. However, it is possible, albeit with less quality, a nonprobability sample only can be used with some calibration (benchmarking) of covariates to make inference about a finite population from the nonprobability sample only; evidently this is risky. Meng (2018) argued that a small bias in big data can be catastrophic; see also Rao (2020) for a review and an interpretation of Meng (2018) relevant to nonprobability samples. As pointed out by SWRB and WSRB, it will be better to use a nonprobability sample to supplement a probability sample.

While both SWRB and WSRB use the Bayesian framework to combine a nonprobability sample and a probability sample, we believe that this is the right way to go. Indeed, as they correctly pointed out, data integration is a big strength of the Bayesian framework. Unlike other methods, we can account for variability with virtually no additional effort. In a non-Bayesian framework, one would need to appeal to other constructs (e.g., variance functions, bootstrapping or asymptotics that understate variability), thereby adding a degree of subjectivity or inconsistency. WSRB provided an improved Bayesian analysis over SWRB. Both SWRB and WSRB are mainly interested in inference about regression coefficients. On the other hand, we are primarily interested in inference about a finite population quantity especially in the presence of a nonprobability sample. We combine the nonprobability sample and the probability sample, where one of them is used to construct the prior. The question we ask is which one of these two approaches is better. We also show how to discount the information provided by the prior in either case. While SWRB did not use survey weights, WSRB did include the calibrated survey weights as a covariate. A major part of our work is to incorporate survey weights into the likelihood.

Like SWRB and WSRB, we use a multiple regression model for the finite population of N units,

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

with appropriate priors on model paramets, $\underset{\sim}{\beta}$ and σ^2, to obtain a fully Bayesian approach. We assume priors on both parameters that are completely improper and noninformative. That is,

$$\pi(\underset{\sim}{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}, \underset{\sim}{\beta} \in R^p, \sigma^2 > 0.$$

Models of this form provide proper posterior distributions provided that the matrix of covariates is full rank. This model holds for all units in the population and so it is called a population model. Within the Bayesian approach, once a full model is written down, it is not subject to change; nothing else should be considered when it is fit using MCMC or other sampling-based methods, otherwise it becomes incoherent. However, for both sampling processes, probability sample and nonprobability sample, we need to make the necessary adjustments to the population model to accommodate the two processes with survey weights and discounting; see Pfeffermann (1993) for an adjustment to a population model to account for departures (e.g., selection bias) from the sampling model.

We observe a probability sample that has survey weights due to the selection of the sample; survey weights carry the selection bias that we want to remove. When a non-probability sample is taken, there are no survey weights and therefore the selection mechanism is not available. Our main problem is to construct a prior based on the ps (nps) when the actual data come from the nps (ps). In the nps, the sample size is n_1 and in the ps the sample size is much less, say n_2. Our ps is small and n_2 is about 10-20% of n_1.

There are three problems; see Rao (2020).

1. Responses are available in both the ps and the nps. In this case, one might use either the ps or the nps as the actual sample. We will need to calibrate the nps to the ps.

2. Responses are not observed in the ps. In this case, we might prefer to use the ps to construct the weights for the nps and use the nps as the actual sample. Again, we will need to calibrate the nps to the ps.

3. Only the nps is available. We assume that population totals are available from a census or administrative records. This can be handled as in Nandram, Cao, Xu and Bhadra (2019) and Xu and Nandram (2019, 2020) and Zhiqing (2020).

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 two examples out of these data, one is a single sample example and the other is a small area example.

First, we use six (6) counties as the nps and the remaining two (2) as the ps. The weights of the nps are discarded, and assumed to be unknown. The population is assumed to be set where the ps is taken, so that the population size is roughly the sum of the survey weights for the ps. The

population size is assumed to be roughly the sum of the survey weights for the ps, and we also assume that both the ps and nps are sampled from this population. The covariates and responses in the nps are $(x_{1i}, y_{1i}), i = 1, \ldots, n_1$. The covariates, responses and survey weights for the ps are $(x_{2i}, y_{2i}, w_{2i}), i = 1, \ldots, n_2$. The weights are missing from the nps, but we denote them by $w_{1i}, i = 1, \ldots, n_1$. We will call this the single area example.

Second, we construct two samples from each of the eight (8) counties. We take about 80% for nps and 20% for the ps. Again, 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. We will call this the small area example. Here, the covariates, responses and survey weights in the nps are respectively $(x_{1ij}, y_{1ij}, w_{1ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{1i}$ and covariates, responses and survey weights of the ps are $(x_{2ij}, y_{2ij}, w_{2ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{2i}$; the survey weights w_{1ij} are unknown in the nps.

Chen, Li and Wu (2020) used a ps and a nps to obtain survey weights in the nps. 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. A summary of this method is given in appendix A. Chen, Li and Wu (2020) did not calibrate the nps to ps. We have winsorized the estimated survey weights in both directions: if they are smaller than 1 or they are extremely large. We have described the calibration and winsorizing (trimming) procedure of the nps in Appendix B; see Haziza and Beaumont (2017) for a review. 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.

Nandram, Choi and Liu (2021) considered a problem similar to the one discussed here, and they have used the BMI data as an illustrative example. However, they did not use the method of Chen, Li and Wu (2020) to estimate the propensity scores (nps weights). Rather they used record linkage via the covariates to fuse the survey weights of the ps to the nps; the survey weights are calibrated to the population size, estimated via the ps. This method avoids the logistic regression (not robust) on the selection indicators, but it borrows the survey weights, albeit not the best idea. As there is no model for the participation variable, the issue of double robustness is null and void. Both Bayesian methods and non-Bayesian methods (least squares together with the bootstrap) are used.

In the single area example, we consider five scenarios (models) to make inference about the finite population mean when we have a probability sample and/or a nonprobability sample. It is possible to ignore the nps altogether, and use only the ps. However, if we have only the nps, we have to make do with what we have; fielding a small parallel ps may be costly for some agencies. It is possible to use only the nps, but this has the risk of large bias and the misleading feeling of small variance just because of its large size. Using only the ps can provide unbiased estimates, but the mean squared error will be large if the sample size of the ps is small. However, the ps gives a good sense of what the point estimate should be. Therefore, when combining the ps and the nps, although with a relatively small sample size, the ps can at least help to guide the estimation procedure of the nps.

As a summary, the novelty of our approach is three-fold. First, we provide a full Bayesian method to combine two likelihoods, one based on the probability sample and the other based on the nonprobability sample, incorporating the survey weights. One of these is used as the prior. Second, the "prior" data are partially discounted using a "power prior", thereby preventing the prior data, ps (nps), to dominate the actual data, nps (ps). Third, we adjust the survey weights to get an effective sample size. This sample size is smaller than the original sample size, thereby accounting for reduced variability induced by the design features in drawing the ps. These traits make our approach novel, more coherent than SWRB and more competitive than WSRB.

This paper has six sections, including this one. In Section 2, we review preparatory materials in the context of data integration on a more informative review of Sakshaug et al. (2019), the use of the power prior, and a review of Meng (2018) for Big Data. In Section 3, we describe our Bayesian methodology that uses discounting and adjusted survey weights for a single area. In Section 4, we discuss an application on body mass index and we specifically describe the five scenarios (models). We also describe a simulation study to make further comparisons of the five scenarios. In Section 5, we discuss small area estimation using a unit-level model together with a numerical example as we described above. Section 6 provides some concluding remarks. The appendices provide technical details on propensity scores, calibration and computational details associated with the single area example and small area example.

2. Preparatory Materials

In this section, we provide a brief review of of data integration as presented by Sakshaug et al. (2019), the power prior and Big Data as in Meng (2018).

2.1 A Quick Review of Sakshaug et al. (2019)

SWRB used a large non-probability sample to supplement a relatively much smaller probability sample (a simple random sample). Essentially, under the Bayesian paradigm, they have used the non-probability sample to provide a prior for the probability sample. They are interested in super-population parameters, but not really finite populations. Also, survey weights are not studied in their work. Therefore, they have studied how well the regression coefficients in the model are estimated. They have also looked at prediction in the sense that they used part of the ps to fit the models and predict the part that is left off. Our work is motivated by SWRB, but we are interested in prediction for a finite population, a much more difficult problem than their prediction problem. Moreover, we want to integrate both datasets into our likelihood function with appropriate penalties and adjustments for survey weights.

SWRB assumed that the nps and the ps are fielded by the same questionnaire. Let $\underset{\sim}{y}_2$ denote the vector of n_2 responses and X_2, a $n_2 \times p$ matrix of covariates associated with the ps, including an intercept. For their baseline model, called the reference model or Model 1, they assume

$$\underset{\sim}{y}_2 \sim \text{Normal}(X_2\underset{\sim}{\beta}, \sigma^2 I).$$

Apriori, they assume that independent priors for the p components of $\underset{\sim}{\beta}$,

$$\beta_j \sim \text{Normal}(\beta_{j0}, \sigma^2_{\beta_{j0}}), j = 1, \ldots, p$$

with $\beta_{j0} = 0$ and $\sigma^2_{\beta_{j0}} = 10^6$. For σ^2, they assumed , $\sigma^2 \sim \text{InvGam}(.001, .001)$, a proper diffuse prior. Denote the posterior mean of $\underset{\sim}{\beta}$ as $\hat{\underset{\sim}{\beta}}_2$. Under this prior, the posterior mean makes no use of the nps data.

In Models 2 and 3, they used the non-probability sample to construct a prior for the probability sample. Let $\underset{\sim}{y}_1$ denote the vector of n_1 responses and X_1, a $n_1 \times p$ matrix of covariates associated with the nps, including an intercept. For Model 2, they have double-used the data, i.e., $\hat{\underset{\sim}{\beta}}_2$, an incoherent procedure in Bayesian statistics. They fit a model, similar to Model 1, to the nps to get $\hat{\underset{\sim}{\beta}}_1$. Then, they assume

$$\beta_j \sim \text{Normal}(\hat{\underset{\sim}{\beta}}_{1j}, (\hat{\underset{\sim}{\beta}}_{1j} - \hat{\underset{\sim}{\beta}}_{2j})^2), j = 1, \ldots, p$$

with the same prior for σ^2. Put aside that this prior double-uses the data, the rationale for the chosen prior variance of β_j is left unclear.

In Model 3, they have bootstrapped the nps data to obtain a prior for the $\underset{\sim}{\beta}$ in the ps. They obtained

$$\beta_j \sim \text{Normal}(\hat{\underset{\sim}{\beta}}_{2j}, \hat{\sigma}^2_{B,j}), j = 1, \ldots, p,$$

where $\hat{\sigma}^2_{B,j}$ is a bootstrap estimate.

One can believe that Model 2 is not sensible. Model 3 indicates a very strong prior on β_j because the bootstrap prior variance above will be very small due to the large size of the nps. Evidently, this is unrealistic and suggests one could make inference from the prior only. Model 1 is fine, but one can simply use the prior $\pi(\sigma^2) \propto 1/\sigma^2$, an objective prior, and there is no conflict. It is just as simple to fit a model to the nps similar to the ps. Moreover, clearly independent priors on the regression coefficients is not sensible. If one needs to use the nonprobability sample to construct a prior for the probability sample, one has to be careful because this prior can dominate the probability sample. In fact, it actually happens in the final estimates. One needs to penalize the prior constructed using the nonprobability sample.

It is possible to convert the nps to a ps. This can be done by first estimating the survey weights using propensity scores (e.g., Chen, Li and Wu 2020), and then adjusting the nps to a simple random sample with a smaller effective sample size using surrogate sampling (Nandram 2007). Assuming no overlaps, one can fit the same model to both datasets (now both are simple random samples). This model is

$$\underset{\sim}{y}_t \mid \underset{\sim}{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(X_t \underset{\sim}{\beta}, \sigma^2), t = 1, 2, \;\; \pi(\underset{\sim}{\beta}, \sigma^2) \propto \frac{1}{\sigma^2},$$

where the length of $\underset{\sim}{y}_1$ is $n_1^* \leq n_1$, with a corresponding adjustment on X_1. [An adjustment is not needed for the simple random sample.] The joint posterior density of $\underset{\sim}{\beta}, \sigma^2$ is proper once at least one of X_1 or X_2 is full rank.

Based on the simulation study they did for both estimation and prediction, the results appear very good; prediction results being less convincing. The results indicate that an increase in bias due to using the nps-based prior, is offset by a reduction in variance, but a degree bias still remains. They also performed an analysis on a real dataset and the results appear satisfactory. In the discussion, they stated, "In conclusion, we find that augmenting a probability sample with a nonprobability sample under the Bayesian framework can produce survey estimates with smaller mean squared error and potentially large cost savings relative to probability-only samples." We concur with SWRB.

2.2 Power Prior

The power prior is an informative prior that combines historical data with current data. It has been extensively used in many different applications over the past thirty years. However, we must be careful that the historical data do not dominate the current data. In an interesting paper, Ibrahim and Chen (2000) reviewed the properties of the power prior for arbitrary regression models (e.g., linear models and generalized linear models). They discussed many applications in diverse disciplines. However, they expressed doubts about the computational aspect of the power prior. Later Ibrahim, Chen, Gwon and Chen (2015) reviewed many theoretical properties for a much larger class of models with many different formulations. There are many versions of the power prior.

In our application, we can treat the probability sample or the non-probability sample as historical data with the appropriate power prior. The power prior is a penalty to the historical data partially discounting it; here the ps or the nps is the historical data. We show how to use the power prior to partially discount the non-probability sample and vice versa.

Suppose $y \mid \underset{\sim}{\theta} \sim f(y \mid \underset{\sim}{\theta})$. Then, one version of the the power prior is

$$f(y \mid \underset{\sim}{\theta}, a) = \frac{\{f(y \mid \underset{\sim}{\theta})\}^a}{\int_y \{f(y \mid \underset{\sim}{\theta})\}^a dy}, 0 \leq a \leq 1.$$

It is important to note that a may not be identifiable in this power prior alone.

If the historical data, $y_{11}, \ldots, y_{1n_1}$, and current data, $y_{21}, \ldots, y_{2n_2}$, are available, then assuming that the historical data and current data are independent, the joint probability 'density' function of the data is

$$f(\underset{\sim}{y}_1, \underset{\sim}{y}_2 \mid \underset{\sim}{\theta}, a) = \prod_{i=1}^{n_1} \frac{\{f(y_{1i} \mid \underset{\sim}{\theta})\}^a}{\int_{y_{1i}} \{f(y_{1i} \mid \underset{\sim}{\theta})\}^a dy_{1i}} \prod_{i=1}^{n_2} f(y_{2i} \mid \underset{\sim}{\theta}), 0 \leq a \leq 1.$$

Here a may be identifiable, but some control may be needed over a in general. Besides for generalized linear models, there may be difficulties in computation, particularly the normalization constant. This problem is particularly severe when Markov chain Monte Carlo methods are used for computations. This prompted Ibrahim, Chen, Gwon and Chen (2015) and others to specify a, and they suggested many methods for doing so. This leads to extensive sensitivity analysis. We keep a random, and if it is needed, a can be specified to be in a sub-interval of $(0,1)$ (e.g., $(.5,1)$

for at most 50% discounting) rather than to actually specify a. Besides it is really a bad idea to specify a in a Bayesian setting because a serves a very important function in the model; the data should 'speak' for it.

Let us consider a simple example for the power prior,

$$y_{11},\ldots,y_{1n_1} \mid \theta,\sigma^2,a \overset{iid}{\sim} \text{Normal}(\theta,\frac{\sigma^2}{a})$$

and for the current data,

$$y_{21},\ldots,y_{2n_2} \mid \theta,\sigma^2 \overset{iid}{\sim} \text{Normal}(\theta,\sigma^2).$$

Clearly, a is not identifiable in the power prior if σ^2 is also unknown. However, if σ^2 is known, a serves the role as a penalty to increase variance (i.e., σ^2 to $\sigma^2/a, 0<a<1$).

Assume the prior $\pi(\theta,\sigma^2,a) \propto 1/\sigma^2$. Letting $D=(\underset{\sim}{y}_1,\underset{\sim}{y}_2)$ and using Bayes' theorem, the joint posterior density is

$$\pi(\theta,\sigma^2,a \mid D) \propto a^{n_1/2}\left(\frac{1}{\sigma^2}\right)^{(n_1+n_2)/2+1} \times$$

$$\exp\left\{-\frac{1}{2\sigma^2}\{a(n_1-1)s_1^2+(n_2-1)s_2^2+an_1(\bar{y}_1-\theta)^2+n_2(\bar{y}_2-\theta)^2\}\right\}, 0\le a\le 1,$$

where $\bar{y}_t$ $s_t^2, t=1,2$, are the sample means and the sample variances.

Letting $\lambda=\frac{an_1}{an_1+n_2}$, it follows that

$$\theta \mid \sigma^2,a,D \sim \text{Normal}\{\lambda\bar{y}_1+(1-\lambda)\bar{y}_2,(1-\lambda)\sigma^2/n_2\},$$

$$\sigma^2 \mid a,D \sim \text{InvGam}\left\{\frac{n_1+n_2-1}{2},\frac{n_2\lambda(\bar{y}_1-\bar{y}_2)^2+a(n_1-1)s_1^2+(n_2-1)s_2^2}{2}\right\}$$

and

$$\pi(a \mid D) \propto \frac{a^{n_1/2}\sqrt{(1-\lambda)/n_2}}{\{(n_2\lambda(\bar{y}_1-\bar{y}_2)^2+a(n_1-1)s_1^2+(n_2-1)s_2^2\}^{(n_1+n_2-1)/2}}, 0\le a\le 1.$$

Therefore, $\pi(a \mid D)$ is well defined for all $0\le a\le 1$. Moreover, there will be no difficulties in computation because a can be sampled using a grid method.

In the single area example, the difference between this illustration and our problem is that we have survey weights and covariates. But the implementation is similar in that we do not need to use a Gibbs sampler; we can use the multiplication rule to draw the samples as in this illustrative example. As pointed out in the literature on power priors, it is possible to have poor mixing when Markov chain Monte Carlo methods (e.g., the Gibbs sampler) are used; this is caused by the stochastic feature in a and this is why researchers have turned away from a full specification of a together with extensive sensitivity analysis, a nuisance. But this slow mixing can be avoided by using a carefully planned block Gibbs sampler instead, as in our implementation of the small area model for the non-probability sample later.

2.3 A Quick Review of Meng (2018)

When the sample is unbalanced with respect to the target population composition, larger data volume increases the relative contribution of selection bias to absolute or squared error. Meng (2018) called this phenomenon a "Big Data Paradox", and he showed both theoretically and empirically that the impact of selection bias on the effective sample size can be extremely large. Primarily, he introduced data defect index, drop out odds and the degree of uncertainty.

He considered a finite population as in survey sampling. So let $R_i, i = 1, \ldots, N$, denote the sample (response) indicators, where $R_i = 1$ if the i^{th} unit is selected and $R_i = 0$ otherwise. He discussed Big Data, self-reported or administrative data, that have no consideration of the R-mechanism. On the contrary, in probability sampling the R-mechanism plays a major role. He further assumed that there is no response error (measurement error). The population values are denoted by $y_1, \ldots, y_N$, with R_i pairing up with y_i. Note that for any R-mechanism, $\sum_{i=1}^{N} R_i = n$, and we are interested in the finite population mean, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, although interest can be on any finite population quantity. Let $\bar{y}$ denote the sample mean and $f = \frac{n}{N}$, the sampling fraction. The discussion applies to both continuous and discrete y. We also let $\text{Var}(y_i) = \sigma_Y^2$.

Meng (2018) brought to light a very important formula for the study of Big Data. He stated that the bias in using $\bar{y}$ to estimate $\bar{Y}$ is given by

$$\bar{y} - \bar{Y} = \rho_{RY} \times \sqrt{\frac{1-f}{f}} \times \sigma_Y,$$

where ρ_{RY}, the correlation between R and Y, is the data quality, $\sqrt{\frac{1-f}{f}}$ is the data quantity and σ_Y is the problem difficulty, where ρ_{RY} is the most important measure in his work. The perfect situation occurs when $\rho_{RY} = 0$. However, ρ_{RY} is not identifiable, creating enormous difficulties in the analysis of Big Data that are mostly very large non-probability samples (without an R-mechanism), where N could be of the order of several millions of records. Data quantity is easy to understand. When $f = 1$, the population values of the variable of interest, y, are fully recorded, and the error is zero, and when $f = 0$, the error is infinite. It is possible to decrease σ_Y by adding more information (i.e., increasing the sample size); this is not necessarily true though. However, the strength of the bias is captured by the magnitude of ρ_{RY} and its sign tells us about the direction of the bias.

Meng (2018) showed that the mean squared error of $\bar{y}$ is given by

$$MSE(\bar{y}) = E_R(\rho_{RY}^2) \times \frac{1-f}{f} \times \sigma_Y^2 = D_I \cdot D_o \cdot D_U,$$

where E_R denotes expectation with respect to any chosen distribution of R, conditional on $\sum_{i=1}^{N} R_i = n$, typical in finite population calculations. He called D_I the data defect index, D_o the drop out odds, and D_U the degree of uncertainty. He pointed out that the most relevant quantity to decrease the $MSE(\bar{y})$ is D_I, but this is also the most difficult quantity to estimate because it is not identifiable; clearly D_I cannot be estimated from the data because the nonsample y's are not available. Meng

(2018) pointed out that the D_I is also useful in the context of missing data because it can used as a measure of non-ignorability. More importantly, he showed that $D_I \propto N^{-1}$ for probability sampling but for non-probability Big Data, D_I does not vanish with large N.

For binary data (i.e., $y_i = 0, 1$), Meng (2018) used the Hoeffding-Frechet bounds to obtain

$$-\min\left\{\sqrt{\frac{D_o}{O_Y}}, \sqrt{\frac{O_Y}{D_o}}\right\} \leq \rho_{RY} \leq \min\left\{\sqrt{O_Y D_o}, \frac{1}{\sqrt{O_Y D_o}}\right\},$$

where $O_Y = \frac{p_Y}{1-p_Y}$ and $p_Y = P(y_i = 1)$. The upper bound is achieved when $R_i = y_i$ and the lower bound is achieved when $R_i = 1 - y_i$. It appears that ρ_{RY} will almost always be very small (close to the lower bound because of its non-identifiability); yet it can be very disastrous.

Finally, he stated the Big Data Paradox, "The bigger the data, the surer you fool yourself." He showed that the effective sample size,

$$n_{eff} = \frac{1}{D_o D_I} \leq n^*_{eff} = \frac{n}{1-f}\frac{1}{ND_I},$$

and because $D_I = O(1)$, this causes ND_I to increase very rapidly with N and the effective sample size (ESS) is $O(1/N)$. That is, the ESS is dramatically reduced; it does not matter how small D_I is, thereby creating a "butterfly" effect. This appears to be a bit controversial. He applied his theory to the 2016 US Presidential election. He attempted to use many datasets to get some control over ρ_{RY}; see Appendix E for estimation of ρ_{RY}.

3. Bayesian Methodology for the Single Area Example

We consider the homogeneous case, consisting of a nonprobability sample and a probability sample. In the construction here, the survey weights are very important and they are to be used to avoid bias due to the sample design and other post-design adjustments such as adjustment for non-response and certain demographics. However, it is worth mentioning again that we are considering the case of a homogeneous sample (population). That is, there are no sub-groups (e,g., small areas) or clustering. We also discuss how to partially discount the prior data (ps or nps).

We note here that the nps and the ps are assumed to be independent samples from the same population (finite population or super-population). The ps is a probability sample, and therefore, it is a representative sample. The nps is a nonprobability sample, and therefore, it is not a representative sample from the same population. We estimate propensity scores (survey weights) tto make the nps compatible with the ps; see Appendix A. If nps (ps) is used as a prior, there is discounting of the nps (ps) because the prior data are historical data.

First, in our approach we need the effective sample size and the adjusted survey weights. Let $y_i, i = 1, \ldots, n$, be independent with $\mathrm{E}(y_i) = \mu_i$, $\mathrm{var}(y_i) = v_i^2$ and $W_i, i = 1, \ldots, n$ are the original survey weights. This is a standard assumption in a super-population model, but it is questionable as the units may not be independent. Then, with just this assumption, Potthoff, Woodbury and

Manton (1992) showed that the equivalent (effective) sample size is n_o, where

$$n_o = \frac{(\sum_{i=1}^n W_i)^2}{\sum_{i=1}^n W_i^2}.$$

The effective sample size, n_o, indicates the extent to which the variance is increased by the unequal weighting. Then, the adjusted survey weights required to eliminate bias introduced by the original survey weights are

$$w_i = n_o \frac{W_i}{\sum_{j=1}^n W_j}, i = 1, \ldots, n. \tag{1}$$

[Note the use of small w for adjusted survey weights.] Here $\sum_{i=1}^n W_i = N$, the population size, and $\sum_{i=1}^n w_i = n_o = \sum_{i=1}^n w_i^2$. We note that n_o has some interesting properties. First, if the W_i are nearly equal, $n_o = n$. Second, if $n > 1$, $n_o > 1$. Third, n_o is invariant to scale and therefore the w_i are invariant to scale. The adjusted weights in (1) will play an important role in the Bayesian methodology.

We can now write the joint density of $y_1, \ldots, y_n$ as a weighted product,

$$g(\underset{\sim}{y} \mid \underset{\sim}{\theta}, \underset{\sim}{w}) = \prod_{i=1}^n \frac{\{f(y_i \mid \underset{\sim}{\theta})\}^{w_i}}{\int \{f(y_i \mid \underset{\sim}{\theta})\}^{w_i} dy_i}, \tag{2}$$

where we have conditioned on $\underset{\sim}{w}$ to cover the case when the survey weights are random. [Henceforth, we will drop the conditioning on $\underset{\sim}{w}$.] Apart from the normalization constant, this is similar to what we do in survey sampling, and the normalization constant will make no difference under normality.

In (2), we are assuming that the population is homogeneous. If the population is heterogeneous, and we know the sub-groups (e.g., small areas or clusters), (2) must be applied to each of them separately. Also, the effective sample size formula given here is applicable under independence.

Second, we describe how to use the power prior to partially discount for the prior data. If we put a prior on $\underset{\sim}{\theta}$, say $\pi(\underset{\sim}{\theta})$, using Bayes' theorem,

$$\pi(\underset{\sim}{\theta} \mid \underset{\sim}{y}) \propto \pi(\underset{\sim}{\theta}) g(\underset{\sim}{y} \mid \theta),$$

which we assume is proper. Actually, we want to construct an informative prior for $\underset{\sim}{\theta}$ using the nps (ps) when the actual data are the ps (nps). We want to do so to avoid the prior from dominating the actual sample information, and therefore, some discounting is necessary. That is, the nps (ps) is used to construct a data-based prior with some discounting; the ps and nps are assumed independent. The power prior can be used for this purpose; see Ibrahim and Chen (2000) and Ibrahim, Chen, Gwon and Chen (2015).

In the case of the regression problem stated here, if the nps is used to construct the prior,

$$\pi(\underset{\sim}{\beta}, \sigma^2 \mid \underset{\sim}{y}_1, a) \propto \frac{1}{\sigma^2} \prod_{i=1}^{n_1} \{\frac{a w_{1i}}{\sigma^2}\}^{1/2} e^{-\frac{a}{2\sigma^2} \sum_{i=1}^{n_1} w_{1i} (y_{1i} - \underset{\sim}{x}_{1i}' \underset{\sim}{\beta})^2},$$

where we have used the prior $\pi(\underset{\sim}{\beta},\sigma^2) \propto 1/\sigma^2$. It is easy to draw samples of $(\underset{\sim}{\beta},\sigma^2)$ because

$$\underset{\sim}{\beta} \mid \sigma^2, \underset{\sim}{y_1} \sim \text{Normal}\left\{\underset{\sim}{\hat{\beta}}, \frac{\sigma^2}{a}\left(\sum_{i=1}^{n_1} w_{1i}\underset{\sim}{x}_{1i}\underset{\sim}{x}_{1i}{}'\right)^{-1}\right\},$$

where $\underset{\sim}{\hat{\beta}} = \left(\sum_{i=1}^{n_1} w_{1i}\underset{\sim}{x}_{1i}\underset{\sim}{x}_{1i}{}'\right)^{-1}\sum_{i=1}^{n_1} w_{1i}\underset{\sim}{x}_{1i}y_{1i}$ and

$$\sigma^2 \mid \underset{\sim}{y_1} \sim \text{IGam}\{\frac{n_1-p}{2}, \ \frac{a\sum_{i=1}^{n_1} w_{1i}(y_{1i}-\underset{\sim}{x}_{1i}'\underset{\sim}{\hat{\beta}})^2}{2}\}.$$

Note that we can specify a because $\frac{\sigma^2}{a}$ is identifiable but not a or σ^2 separately. In fact, this is the prior for $\underset{\sim}{\beta},\sigma^2,a$ when $a \sim \text{Uniform}(0,1)$.

For our problem, if the nps is used to construct the prior and the ps is the actual sample, the joint posterior density is

$$\pi(\underset{\sim}{\theta},a \mid \underset{\sim}{y_1},\underset{\sim}{y_2}) \propto \pi(\underset{\sim}{\theta})\pi(a)\prod_{i=1}^{n_1}\frac{\{f(y_{1i} \mid \underset{\sim}{\theta})\}^{aw_{1i}}}{\int\{f(y_{1i} \mid \underset{\sim}{\theta})\}^{aw_{1i}}dy_{1i}}\prod_{i=1}^{n_2}\frac{\{f(y_{2i} \mid \underset{\sim}{\theta})\}^{w_{2i}}}{\int\{f(y_{2i} \mid \underset{\sim}{\theta})\}^{w_{2i}}dy_{2i}}, 0 \le a \le 1,$$

where a is a discounting factor for the power prior. Similarly, if the ps is used to construct the prior and nps is the actual sample, the joint posterior density is

$$\pi(\underset{\sim}{\theta},a \mid \underset{\sim}{y_1},\underset{\sim}{y_2}) \propto \pi(\underset{\sim}{\theta})\pi(a)\prod_{i=1}^{n_1}\frac{\{f(y_{1i} \mid \underset{\sim}{\theta})\}^{w_{1i}}}{\int\{f(y_{1i} \mid \underset{\sim}{\theta})\}^{w_{1i}}dy_{1i}}\prod_{i=1}^{n_2}\frac{\{f(y_{2i} \mid \underset{\sim}{\theta})\}^{aw_{2i}}}{\int\{f(y_{2i} \mid \underset{\sim}{\theta})\}^{aw_{2i}}dy_{2i}}, 0 \le a \le 1.$$

We note that although a is not identifiable in the prior alone, as the two samples are combined, a becomes identifiable, at least weakly.

We assume that there are no overlaps between the nps and ps. That is, a unit is not captured in both the ps and the nps. We note that to stay within the Bayesian paradigm, we cannot use the responses from the actual sample to construct the prior; all other variables can be used.

With the nps as prior, the appropriate posterior density is

$$\pi(\underset{\sim}{\beta},\sigma^2,a \mid \underset{\sim}{y_1},\underset{\sim}{y_2}) \propto$$

$$\frac{1}{\sigma^2}\prod_{i=1}^{n_1}\{\frac{aw_{1i}}{\sigma^2}\}^{1/2}e^{-\frac{a}{2\sigma^2}\sum_{i=1}^{n_1} w_{1i}(y_{1i}-\underset{\sim}{x}_{1i}'\underset{\sim}{\beta})^2}\prod_{i=1}^{n_2}\{\frac{w_{2i}}{\sigma^2}\}^{1/2}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_2} w_{2i}(y_{2i}-\underset{\sim}{x}_{2i}'\underset{\sim}{\beta})^2}, 0 \le a \le 1,$$

where we assume a noniformative uniform prior on a. Here, the parameters can be drawn by simply using a random sampler (not a Markov chain). It is possible to integrate out $\underset{\sim}{\beta}$ and σ^2 to get the posterior density of a; and samples from the posterior density of a can be drawn easily using a grid method. Also, posterior propriety can be established.

Letting $\underset{\sim}{y} = (\underset{\sim}{y_1},\underset{\sim}{y_2})$, the joint posterior density can be written as

$$\pi(\underset{\sim}{\beta},\sigma^2,a \mid \underset{\sim}{y}) \propto a^{n_1/2}(\frac{1}{\sigma^2})^{\frac{n_1+n_2}{2}+1}e^{-\frac{1}{2\sigma^2}Q}, 0 \le a \le 1,$$

where $Q = a\sum_{i=1}^{n_1} w_{1i}(y_{1i} - \underset{\sim}{x}_{1i}\underset{\sim}{\beta})^2 + \sum_{i=1}^{n_2} w_{i2}(y_{2i} - \underset{\sim}{x}_{2i}\underset{\sim}{\beta})^2$. In Appendix C, we obtain a random sampler to draw from the joint posterior density of $\underset{\sim}{\beta}, \sigma^2, a \mid \underset{\sim}{y}$ and we show that the joint posterior density is proper. We state the main distributions to get a random sampler.

For convenience, letting $a_1 = a, a_2 = 1$ (i.e., nps is used as a prior), we define

$$A = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si} \underset{\sim}{x}_{si}\underset{\sim}{x}'_{si},\ \underset{\sim}{b} = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si}\underset{\sim}{x}_{si}y_{si} \text{ and } d = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si}(y_{si} - \underset{\sim}{x}'_{si}\hat{\underset{\sim}{\beta}})^2.$$

Then, letting $\hat{\underset{\sim}{\beta}} = A^{-1}\underset{\sim}{b}$, we have

$$\underset{\sim}{\beta} \mid \sigma^2, a, \underset{\sim}{y} \sim \text{Normal}(\hat{\underset{\sim}{\beta}}, \sigma^2 A^{-1}),$$

$$\sigma^2 \mid a, \underset{\sim}{y} \sim \text{IGam}\left(\frac{n_1 + n_2 - p}{2}, \frac{d}{2}\right),$$

and

$$\pi(a \mid \underset{\sim}{y}) \propto \frac{a^{n_1/2} \mid A \mid^{-1/2}}{d^{(n_1+n_2-p)/2}}, 0 \le a \le 1.$$

The joint posterior density is proper because $0 \le a \le 1$ and all quantities are well defined, provided the design matrix, X_2, of the ps is full rank. We can draw a sample from $\pi(a \mid \underset{\sim}{y})$ using the grid method. The other parameters, $\underset{\sim}{\beta}$ and σ^2 are drawn in a standard manner. However, the main difficulty is to find the inverse and the determinant of A at each value of a because a is jittered at each draw from its discrete distribution (i.e., the draws of a are different almost surely). It is possible to get around this difficulty using a sub-sampling procedure, albeit with less quality; yet we might want to do so.

Here, we assume that the actual sample is the ps, and the nps plays the primary role in the construction of the prior. By interchanging the roles of a_1 and a_2, we will get the posterior density for the case where nps is the actual sample and the ps plays the primary role in the construction of the prior.

It is difficult to do prediction fully within the Bayesian paradigm because the nonsampled covariates are unknown and the population is fairly large. One can fill in the nonsampled covariates subject to some constraint like the total for each covariate is known. In this case, one would need to sample the entire population using surrogate sampling. That is, once the parameters are estimated from the sample model (population model adjusted with survey weights and the discounting factor), one can then use the appropriate parameters in the population model to sample all the N values of the study variable (i.e., projective inference). This is surrogate sampling (see Nandram 2007, Nandram and Choi 2010 and many others). However, for a large population this is a time-consuming procedure that depends on external sources of information (e.g., webscraping or administrative records for N and the total for each covariate). Based on quasi randomization, we have used an alternative procedure.

It is true that given $\underset{\sim}{\beta}, \sigma^2$ and all the sample data, $\underset{\sim}{y}_s$,

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

where $\bar{\underset{\sim}{X}} = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{x}_i$, population average covariate, and N, the population size, are both unknown, and in the Bayesian paradigm, these are parameters. Here, using the ps and inverse probability weighted estimators,

$$\bar{Y} \mid \left\{ \underset{\sim}{\beta}, \sigma^2, \bar{\underset{\sim}{X}} = \frac{\sum_{i=1}^{n_2} W_{2i} \underset{\sim}{x}_{2i}}{\sum_{i=1}^{n_2} W_{2i}}, N = \sum_{i=1}^{n_2} W_{2i}, \underset{\sim}{y}_s \right\} \sim \text{Normal} \left\{ \frac{\sum_{i=1}^{n_2} W_{2i} \underset{\sim}{x}'_{2i}}{\sum_{i=1}^{n_2} W_{2i}} \underset{\sim}{\beta}, \frac{\sigma^2}{\sum_{i=1}^{n_2} W_{2i}} \right\},$$

where $W_{21}, \ldots, W_{2n_2}$ are the original survey weights in the ps. We have used this same technique for all five scenarios (models); in Scenario G below, we assume simple random sampling (i.e., survey weights are equal). Unfortunately, it is difficult to take the variabilities (not quasi randomization) of the mean and variance into consideration, and further study is needed in a super-population model-based analysis like what we attempt here. One way to do so is to bootstrap the probability sample.

4. Numerical Analysis for the Single Sample Example

In this section, we consider numerical analysis for the single sample example. Specifically, we use the example on BMI data to compare different scenarios (models) and we present a simulation study.

As a preliminary analysis, we have looked at the probability sample only, $(x_{2i}, W_{2i}, y_{2i}), i = 1, \ldots, n_2$. By simply bootstrapping these data, we can provide distributions for $N, \bar{\underset{\sim}{X}}, \bar{Y}$. We have obtained $B = 10{,}000$ Bayesian bootstrap samples, each represented by $(x^*_{2i}, W^*_{2i}, y^*_{2i}), i = 1, \ldots, n_2$. Then, we computed $N_b = \sum_{i=1}^{n_2} W^*_{2i}$, $\bar{\underset{\sim}{X}}_d = \frac{\sum_{i=1}^{n_2} W^*_{2i} x^*_{2i}}{N_b}$ and $\bar{Y}_b = \frac{\sum_{i=1}^{n_2} W^*_{2i} y^*_{2i}}{N_b}, b = 1, \ldots, B$. A 95% credible interval for $\bar{Y}$ is $(25.008, 27.115)$ with posterior mean, $PM = 26.002$ and posterior standard deviation, $PSD = .534$. The key question is, "Can we keep PM the same and considerably reduce the PSD?" We have also a 95% credible interval for N is $(1{,}946{,}029,\ 2{,}837{,}823)$ and for $\bar{\underset{\sim}{X}}$, they are respectively $(42.791,\ 50.380)$, $(.012,\ .060)$ and $(.437,\ 27.115)$, corresponding to age, race and sex. It is possible to use the Bayesian bootstrap distributions to express uncertainty in the prediction about the inverse probability weighted estimators. This can be coupled with the Bayesian method.

We describe five scenarios, which we denote by B, C, D, E, G. We want to see how the weights and the discount factor change the Bayesian predictive inference. The five scenarios (models) are given next.

i. B uses the nps to make inference; the weights are obtained via propensity scores assisted by the ps.

ii. C uses both the ps and the nps; the nps is used as the prior and there is partial discounting.

iii. D uses both the ps and the nps; the ps is used as the prior and there is partial discounting.

iv. E uses only the ps with the survey weights, of course.

v. G uses only the ps; weights are omitted, and therefore G is different from E. However, N is still unknown, and we have used N from the ps.

A comparison of these scenarios is informative and it provides important clues on survey weights, discounting and data integration.

4.1 BMI Data

We analyze the BMI data on the 8 counties in California in the single sample, a nonprobaility sample and a probability sample. We note that the correlation between the BMI values and the survey weights for the probability sample is almost zero ($\approx -.145$), thereby there is little effect from the survey weights; it is a bit disappointing that the correlation is so low, but there are still important differences among the five scenarios.

In Table 1, we present posterior summaries of the finite population mean. We use the posterior mean (PM), posterior standard deviation (PSD), numerical standard error (NSE, not really necessary here), posterior coefficient of variation (PCV) and 95% credible interval. We see that the PM of E is different from the other four models which are very similar. For PSDs, B, C, D are similar with those for E and G much larger. NSEs and PCVs are very good for all models. The 95% HPD intervals for B, C, D are similar and to the right of E; G has wide intervals. As expected, E should be unbiased with large PSD.

In Figure 1, we have compared the posterior densities of the finite population mean for the five methods. All posterior densities are unimodal with B, C and D concentrated around 27; E and G are different from these. In the plot, E is to the left; E and G have very large spread.

The discount factor a is significant when the nps is used as the prior [95% HPD interval is $(.670, .945)$] and it is nearly 0 when the ps is used as the prior [95% HPD interval is $(.989, .999)$]. This is sensible because the ps is small and of high quality and the nps is large but of low quality.

We have looked at similar posterior summaries for the regression coefficients, variance, σ^2, and the discounting factor, a. We note that SWRB and WSRB were mainly concerned about regression coefficients. Inference about β_1, the intercept, is similar over the five models. However, there are some differences of the other regression coefficients. For models C and D, all parameters are important. In models A, E and G, β_3, β_4 are important; for models B and G, β_2 is also important. The PSDs for models B, C, D and E, which are similar (D has the smallest PSD), are much smaller than those for model G.

We use the Bayesian bootstrap to assist taking care of the variability in the estimated survey weights for the non-probability sample. Specifically, we incorporate the variability of the estimated weights, $W_{2j}, j = 1, \ldots, n_2$, in the models. We also assess the variability in estimating the population size by $\sum_{j=1}^{n_2} W_{2j} = N$ and $\bar{x}_{\tilde{2}} = \frac{\sum_{j=1}^{n_2} W_{2j} x_{\tilde{2}j}}{\sum_{j=1}^{n_2} W_{2j}}$. The procedure has the following steps.

a. Use the Bayesian bootstrap to draw a random sample from the nps and ps respectively.

b. Compute $\sum_{j=1}^{n_2} W_{2j} = N$ and $\bar{x}_{\tilde{2}} = \frac{\sum_{j=1}^{n_2} W_{2j} x_{\tilde{2}j}}{\sum_{j=1}^{n_2} W_{2j}}$.

Table 1: Comparison of five models using BMI data

Model	PM	PSD	PCV	95% CI
B	27.321	0.153	0.006	(27.029, 27.630)
C	27.045	0.134	0.005	(26.787, 27.301)
D	27.097	0.135	0.005	(26.825, 27.350)
E	25.979	0.299	0.012	(25.395, 26.562)
G	26.856	0.304	0.011	(26.285, 27.470)

NOTE: The models are B: nps only; C: data integration with nps as prior; D: data integration with ps as prior; E: ps only; G: ps without survey weights. When the nps is used as the prior, the 95% HPD interval for the discount factor (a) is $(0.670, 0.945)$ and when ps is used as the prior, it is $(0.989, 0.999)$.

c. Fit the models and do the prediction.

Repeat (a), (b) and (c) 1000 times to get the posterior distribution of the finite population mean. To get the posterior density of $\bar{Y}$, letting $\underset{\sim}{\Omega}$ denote the vector of super-population parameters, we use the following decomposition,

$$[\bar{Y}, \bar{X}, N, \underset{\sim}{\Omega}, \underset{\sim}{W}, \underset{\sim}{y_s}] = [\bar{Y} \mid \bar{X}, N, \underset{\sim}{\Omega}, \underset{\sim}{W}, \underset{\sim}{y_s}] \times [\underset{\sim}{y_s} \mid \underset{\sim}{W}, \underset{\sim}{\Omega}] \times [\underset{\sim}{\Omega}] \times [\bar{X}, N, \underset{\sim}{W}],$$

where $[\bar{X}, N, \underset{\sim}{W}]$ is the bootstrap 'posterior' distribution. Throughout, we condition on $\underset{\sim}{x}$ and we assume independence of the sample indicators and the study variables.

In Table 2 we compare Models B, C and D to assess the underestimation in assuming the estimated weights are known. The biggest increase is in B while C and D show very little increase in standard deviations. For E, the sample weights are known (not estimated) and there is still an increase in standard deviation. Therefore, the increase in standard deviation does not come only from assuming the estimated weights are known. This phenomenon should be looked at more carefully. Note *G* does not have survey weights, except for *N* from the ps.

4.2 Simulation Study

We follow Chen, Li and Wu (2020), henceforth CLW, design-based method for generating the finite population and the sample. However, we associate the data simulation within the framework of the BMI data. The BMI data have the following structure,

$$E(y_i) = 23.8449 + .0559x_{1i} + 2.2656x_{2i} + .0452x_{3i}, i = 1, \ldots, N$$

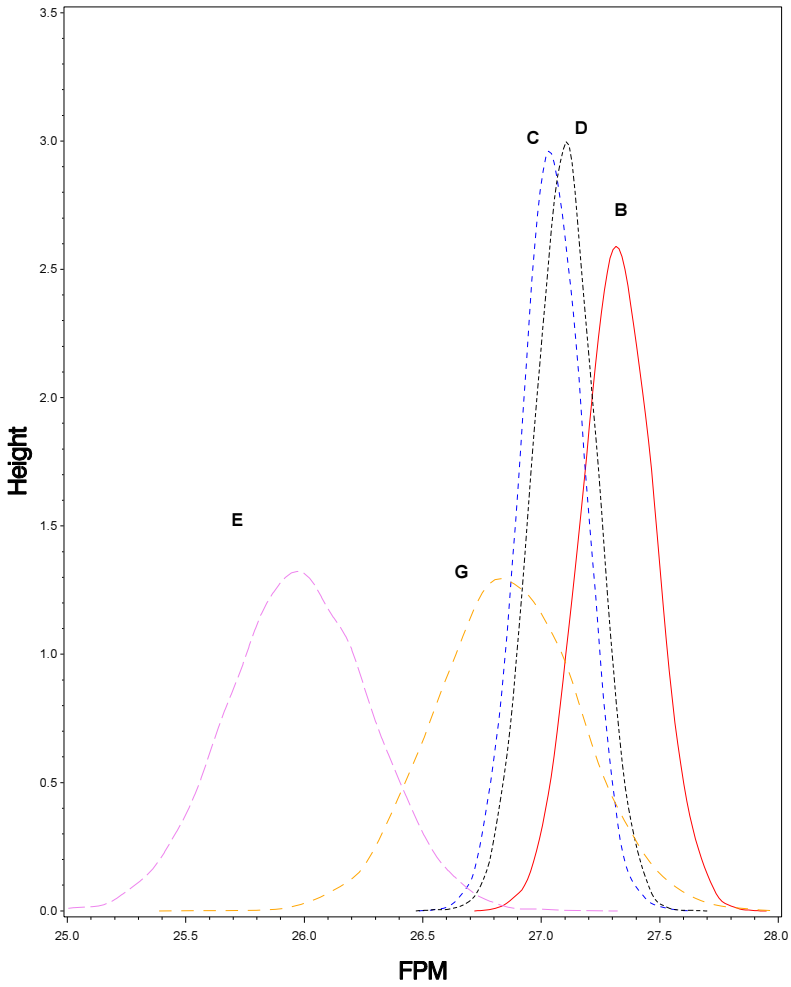

Figure 1: Comparison for the posterior distributions of the finite population mean by the five models (B, C, D, E, G)

at least for the samples (nps and ps). We adjust this a bit by replacing .0452 by .2525 to avoid computational instability. Following CLW, we have taken $N = 20000$, $n_1 = 1500$ and $n_2 = 300$.

Modifying sampling process, we perform the following steps, for $i = 1, \ldots, N$,

i. Draw

$$x_{1i} \overset{ind}{\sim} \text{Uniform}(20, 90)$$

and set $b_i = \{23.8449 + .0559 x_{1i}\}^{\frac{1}{10}}$;

Table 2: Bootstrap study: Underestimation of variability for four selected models

Model	PM	PSD	CV	NSE	95% CI
No Bayesian bootstrap					
B	27.321	0.153	0.002	0.006	(27.029, 27.630)
C	27.045	0.134	0.002	0.005	(26.787, 27.310)
D	27.098	0.135	0.001	0.005	(26.824, 27.350)
E	25.979	0.299	0.003	0.012	(25.395, 26.562)
G	26.856	0.304	0.011	0.009	(26.285, 27.470)
Bayesian bootstrap					
B	27.424	0.470	0.018	0.017	(26.895, 27.845)
C	26.951	0.218	0.006	0.008	(26.510, 27.334)
D	27.088	0.208	0.006	0.008	(26.651, 27.436)
E	25.984	0.371	0.011	0.018	(25.288, 26.772)
G	26.840	0.303	0.009	0.011	(26.211, 27.374)

NOTE: The bootstrap posterior distribution is based on 1000 samples that provide PM, posterior mean, PSD, posterior standard deviation, W, width of the 95% HPD interval and CV, coefficient of variation. In C the prior is the ps and in D the prior is the nps.

ii. Draw

$$x_{2i} \mid x_{1i} \overset{ind}{\sim} \text{Bernoulli}\{\frac{e^{b_i}}{1+e^{b_i}}\}$$

and set $b_i = \{23.8449 + .0559x_{1i} + 2.2656x_{2i}\}^{\frac{1}{10}}$;

iii. Draw

$$x_{3i} \mid x_{1i}x_{2i} \overset{ind}{\sim} \text{Bernoulli}\{\frac{e^{b_i}}{1+e^{b_i}}\};$$

iv. Finally, construct

$$y_i = 23.8449 + .0559x_{1i} + 2.2656x_{2i} + .2525x_{3i} + e_i, e_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2).$$

This gives us the finite population of values $(\underset{\sim}{x}_i, y_i), i = 1, \ldots, N$. So we can compute the true value of $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Following CLW, we have selected σ^2 such that the correlation, Cor$(23.8449 +$

$.0559x_{1i} + 2.2656x_{2i} + .2525x_{3i}, y_i) = \rho$, and we have selected ρ as $\rho = ..20, .30, 50, .80$. This is done by by trial and error.

Table 3: Simulation Study I: No misspecification

		Model				
Measure	ρ	B	C	D	E	G
ARB	0.20	0.006	0.006	0.005	0.017	0.015
	0.30	0.003	0.003	0.003	0.010	0.008
	0.50	0.002	0.002	0.002	0.006	0.005
	0.80	0.001	0.001	0.001	0.002	0.002
PRMSE	0.20	0.282	0.270	0.264	0.705	0.730
	0.30	0.164	0.160	0.155	0.421	0.424
	0.50	0.091	0.089	0.086	0.234	0.236
	0.80	0.039	0.038	0.037	0.101	0.101
Cov	0.20	0.997	0.944	0.978	0.823	0.939
	0.30	0.999	0.949	0.980	0.834	0.960
	0.50	0.999	0.948	0.980	0.833	0.960
	0.80	0.999	0.949	0.980	0.834	0.960
Wid	0.20	0.863	0.757	0.771	1.698	2.064
	0.30	0.526	0.461	0.471	1.034	1.254
	0.50	0.292	0.257	0.262	0.575	0.698
	0.80	0.126	0.110	0.113	0.247	0.300

NOTE: Here, x_3 is used in both the response model and the participation model.

The selection probabilities are from

$$\pi_{1i} = \exp(b_i)/(1+\exp(b_i), b_i = \theta_0 + .1x_{1i} + 0.2x_{2i} + 0.1x_{3i}, i = 1, \ldots, N$$

and we vary θ_0 is selected by trial and error such that $\sum_{i=1}^{N} \pi_{1i} = n_1$. We select π_{2i} such that

$$\pi_{2i} = n_2 z_i / \sum_{i=1}^{N} z_i, z_i = \theta_1 + x_{1i} + 0.2x_{2i} + 0.1x_{3i},$$

where θ_1 is selected, again by trial and error, to ensure $\max\{z_i\}/\min\{z_i\} \approx 50$. This deviates a little bit from CLW.

As in CLW, the nps is taken using Poisson sampling with probabilities π_{1i} and target sample size n_1, and the ps taken using randomized systematic PPS sampling with target sample size n_2.

We have run the simulations as follows.

a. For each setting of ρ, we have generated one finite population, and we took 1000 samples (a nps and a ps) from it. We use the following notations. T is true finite population mean, PM is the posterior mean and PSD is posterior standard deviation; (C025, C975) is the 95% highest posterior density interval (HPDI);

b. We computed the absolute relative bias: $ARB = \mid (PM - T)/T \mid$; posterior root mean squared error, $PRMSE = \sqrt{(PM-T)^2 + PSD^2}$ and incidence: $I = 1$ if a 95% HPDIs containing T, 0 otherwise; width, $Wid = C975 - C025$;

c. Finally, we averaged the 1000 runs; coverage, Cov, is the proportion of HPDIs containing T.

In Table 3, we present simulation comparisons. The smallest ARBs come from B, C, D and E, G are slightly larger. The PRMSEs of B, C, D are smaller than E, G with C, D slightly smaller than B. The coverage for E is below the nominal value of 95%; B and D are too conservative but C is just about the nominal value. For Wid, C and D dominate the others, considerably shorter than B, and E and G are too wide. We note that ARB, PRMSE and Wid decrease with increasing ρ; this must be true because larger ρ means smaller σ^2 in the simulation runs (a possible defect in the simulation design). It is clear that C is the winner and B and D are competitive. For the discount factor a, the PM (PSD), averaged over the simulation runs, are for C .569 $(.051)$ (i.e., considerable discounting) and for D .981 $(.016)$ (i.e., no discounting) with very little changes over ρ. This indicates that one should use the nps as the prior with a penalty and inference should be made using the ps (poor coverage although very wide) integrated with the nps. We have seen similar results in the example on BMI data.

We have also looked at two cases of mis-specifications. The first case omitted the third covariate from model fitting, and the second case the third covariate is not used for data collection and model fitting; the third covariate is used in getting the propensity scores. We found that the two models with discounting are competitive with the others, but for all models coverage decreases as ρ increases.

5. A Small Area Model for a Non-probability Sample

Beaumont (2020) argued that it is sensible to use a non-probability sample to supplement a probability sample in small area estimation. Rao (2020) stated that a non-probability sample can be used to construct covariates for probability samples in small area estimation; he cited five papers where this has been done. However, we can use the non-probability sample to supplement a probability sample within small areas similar to what we have done in Model 3. The small area model will include random effects as an attempt to discriminate the areas.

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 and the probability sample; or at least only a common set of covariates will be used). Of course, they generally differ by unit.

b. Within an area, the random effects are the same in the non-probability sample and the probability sample.

c. Pooling will take place using a common set of regression coefficients and variance components over all areas in the two samples.

d. It is possible to have some areas with only a probability sample, and some areas with only a non-probability sample.

We use an extended version of the model of Battese, Harter and Fuller (BHF, 1988).

We assume there are ℓ areas and within the i^{th} area, there are a non-probability and a probability sample of sizes n_{1i} and n_{2i} and population sizes N_{1i} and 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. The covariates are $(\underset{\sim}{x}_{si}, i = 1,\ldots,n_{si}, s = 1,2)$, but the covariates are unobserved for the nonsamples. The responses are $y_{si}, i = 1,\ldots,n_{si}$. There are also survey weights $W_{si}, i = 1,\ldots,n_{si}$, such that $\sum_{j=1}^{n_{si}} W_{sij} = N_{si}$; so N_{si} may be known or unknown. Bayesian predictive inference is required for

$$\bar{Y}_i = \frac{1}{N_{2i}} \sum_{j=1}^{N_{2i}} y_{2ij}, i = 1,\ldots,\ell,$$

the finite population means, based on the probability sample. Of course, the model permits the use of the non-probability sample, as we have seen for the single sample model and as we will see for the small area model soon. That is, there is pooling across areas and within areas from both the non-probability sample and the probability sample.

As we stated, the discounting factors will only be used for the nps, which will be used to construct the prior and the ps will be used as the actual data and 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).

5.1 Small Area Model

Our model for the two samples over the areas is

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

where w_{sij} are the adjusted weights within areas,

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

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

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.

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 θ and γ. 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 $(\underset{\sim}{a},\theta,\gamma,\rho)$ is

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

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

Using Bayes' theorem, letting $\underset{\sim}{y}$ denote the vector of all observations, the joint posterior density is

$$\pi(\underset{\sim}{\nu},\underset{\sim}{a},\underset{\sim}{\beta},\sigma^2,\rho,\theta,\gamma \mid \underset{\sim}{y}) \propto$$
$$\frac{1}{\sigma^2}\prod_{i=1}^{\ell}\left\{\left[\prod_{j=1}^{n_{1i}}\sqrt{\frac{a_i w_{1ij}}{2\pi\sigma^2}}e^{-\frac{a_i w_{1ij}}{2\sigma^2}(y_{1ij}-\underset{\sim}{x}'_{1ij}\underset{\sim}{\beta}-\nu_i)^2}\prod_{j=1}^{n_{2i}}\sqrt{\frac{w_{2ij}}{2\pi\sigma^2}}e^{-\frac{w_{2ij}}{2\sigma^2}(y_{2ij}-\underset{\sim}{x}'_{2ij}\underset{\sim}{\beta}-\nu_i)^2}\right]\right.$$
$$\left.\times\sqrt{\frac{1-\rho}{2\pi\rho\sigma^2}}e^{-\frac{1-\rho}{2\rho\sigma^2}\nu_i^2}\ \frac{a_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-a_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\{\theta\frac{1-\gamma}{\gamma},(1-\theta)\frac{1-\gamma}{\gamma}\}}\right\}. \tag{3}$$

Letting $\Omega_1 = (\underset{\sim}{a},\theta,\gamma,\rho)$ and $\Omega_2 = (\underset{\sim}{\nu},\underset{\sim}{\beta},\sigma^2$, to fit the posterior density in (3), we will first integrate out Ω_2 and sample the posterior density of $\Omega_1 \mid \underset{\sim}{y}$ using the Gibbs sampler; see Appendix D. Then, we can sample $\Omega_2 \mid \Omega_1,\underset{\sim}{y}$ using the composition method via

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

where $\pi_1(\sigma^2 \mid \Omega_1,\underset{\sim}{y})$, $\pi_2(\underset{\sim}{\beta} \mid \sigma^2,\Omega_1,\underset{\sim}{y})$ and $\pi_3(\underset{\sim}{\nu} \mid \underset{\sim}{\beta},\sigma^2,\Omega_1,\underset{\sim}{y})$ are all in standard forms, inverse gamma, p-variate normal and independent normals respectively; see Appendix D.

Bayesian predictive inference is required for $\bar{Y}_{2i} = \frac{1}{N_i}\sum_{i=1}^{N_i} y_{2ij}$. As the 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,

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

where $\bar{\underset{\sim}{X}}_{2i} = \frac{1}{N_{2i}} \sum_{i=1}^{N_{2i}} \underset{\sim}{x}_{2ij}$ and N_{2i} are unknown. We use the Horvitz-Thompson estimators $\bar{\underset{\sim}{x}}_{2i} = \frac{\sum_{j \in S_{2i}} w_{2ij} \underset{\sim}{x}_{2ij}}{\sum_{j \in S_{2i}} w_{2ij}}$ and $\sum_{j \in S_{2i}} w_{2ij}$ (IPW, inverse probability weighted estimators). Then,

$$\bar{Y}_{2i} \mid \nu_i, \underset{\sim}{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\bar{\underset{\sim}{x}}_{2i}' \underset{\sim}{\beta} + \nu_i, \frac{\sigma^2}{\sum_{j \in S_{2i}} w_{2ij}} \right), i = 1, \ldots, \ell. \tag{4}$$

Once we have drawn $(\underset{\sim}{\nu}, \underset{\sim}{\beta}, \sigma^2)$ using the Gibbs sampler, we simply draw the Y_{2i} from (4). According to the model, all the sampled data are used in the predictive inference.

Observe that $\mathrm{E}(\bar{Y}_i \mid \nu_i, \underset{\sim}{\beta}, \sigma^2, \rho) = \bar{\underset{\sim}{x}}_{2i}' \underset{\sim}{\beta} + \lambda_i (\bar{\bar{y}}_i - \bar{\underset{\sim}{\bar{x}}}_i' \underset{\sim}{\beta})$; see Appendix D for definitions. Then,

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

and

$$\mathrm{Var}(\bar{Y}_{2i} \mid \underset{\sim}{\beta}, \sigma^2, \rho, \underset{\sim}{y}) = \left\{ \left(\sum_{j \in S_{2i}} w_{2ij} \right)^{-1} + \frac{\rho}{1 - \rho} (1 - \lambda_i) \right\} \sigma^2.$$

These can be used to form Rao-Blackwellized density estimators for $\bar{Y}_{2i}$.

5.2 Operationalizing the SAE 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 not be easy 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

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

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

$$a_i = \frac{\phi_0 n_i^{\frac{\phi_1}{1 - \phi_1}}}{1 - \phi_0 + \phi_0 n_i^{\frac{\phi_1}{1 - \phi_1}}}, i = 1, \ldots, \ell, \tag{5}$$

where $0 < \phi_0, \phi_1 < 1$. Note that if $\phi_1 = 0$, $a_i = \phi_0$ and there will be no dependence on the n_i. Now, simply substitute the a_i in the SAE model by (5) 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, θ and γ, are now eliminated or replaced by ϕ_0 and ϕ_1.

5.3 Numerical Example on Small Area Estimation

For the eight counties of California, which we use as a numerical example, the Gibbs sampler is efficient. 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 $5,500$ iterates, used 500 as a "burn in" and systematically selected every fifth to get a 'random' sample of $M = 1,000$. We performed the diagnostic procedures. 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. 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 15 minutes with (c) taking almost all the time.

Table 4: BMI data from eight counties

Model	LPML	DIC	BPP	DM	PRMSE
Discounting	977.491	1946.369	0.553	2.626	1.606
Logit	975.866	1943.152	0.528	2.623	1.783
No discounting	1235.930	2472.066	1.000	2.616	1.718
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).

In Table 4 we present 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). 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, 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.

6. Concluding Remarks

Two important findings show up. First, if one has the study variable on both the ps and the nps, and the nps is much larger than the ps, as is usually the case, then it is better to use the nps to construct the prior with partial discounting. Second, if the ps is used to construct the prior with partial discounting, there will be virtually no discounting. Apparently this is sensible because the ps is much smaller than the nps and of higher quality. It is erroneous to use the ps as the prior or to use the ps to supplement the nps, rather one should use the nps to supplement the ps, provided the study variable is available in the ps.

We have also 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); we have extended Toto and Nandram (2010) or Molina, Nandram and Rao (2014), who provided a Bayesian approach, to solve the problem. However, our work here was motivated by Beaumont (2020) and Rao (2020) and Beaumont and Rao (2020) but these authors did not discuss unit-level models; Beaumont and Rao (2020) showed how to use the Fay-Herriot model to improve inference for the small areas in the ps, covariates being drawn from the nps.

There is a need to express uncertainty in the estimation of propensity scores in the Bayesian approach. We have shown how to use the bootstrap (Bayesian or non-Bayesian) method to incorporate uncertainty in the estimated survey weights for the single sample example. There is considerable underestimation in Scenario B with C and D showing much less underestimation; there is a significant underestimation in Scenario E. However, the bootstrap is not the best way to do this; one would need a model to contain the unknown survey weights with estimation being done in the same model. This is under study. Also, it is possible to have the discounting factor varies with the observation. For example, using the nps as the prior, we can easily replace a by $a^{1/w_{1i}}$, reflecting less discounting for observations with larger survey weights.

The assumption of normality on the BMI data is perhaps not a very good one because the BMI data are skewed and discrete; see Yin and Nandram (2020 a,b). Also, more robust methods on propensity scores are needed. The generalized Dirichlet process, the Pitman-Yor process, or general 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; see Ishwaran and James (2001). It is also possible to use BART in data integration (e.g., Rafei, et al. 2021); one does not need to express a relation between study variable and covariates. 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).

Acknowledgments

Balgobin Nandram gave an invited presentation on one version of this paper at the 2021 annual meeting of Statistical Society of Canada. This work was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

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 very clever way to get the propensity scores for the nps, and therefore the survey weights, which are 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 first clever idea is to write the log-likelihood as

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

The second clever idea is to use the pseudo-log-likelihood by replacing the second term by the Horvitz-Thompson estimator since the nonsample $\underset{\sim}{x}_i$ are unknown, as

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

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

However, the gradient vector of the log-pseudo-likelihood in the general form is

$$\Delta(\underset{\sim}{\theta}) = \sum_{i=1}^{n_1} \{1 - \pi(\underset{\sim}{x}_{1i}; \underset{\sim}{\theta})\} \frac{\partial \pi(\underset{\sim}{x}_{1i}; \underset{\sim}{\theta})}{\pi(\underset{\sim}{x}_{1i}; \underset{\sim}{\theta})} - \sum_{i=1}^{n_2} w_{2i} \frac{\partial \pi(\underset{\sim}{x}_{2i}; \underset{\sim}{\theta})}{1 - \pi(\underset{\sim}{x}_{2i}; \underset{\sim}{\theta})},$$

where $\partial\pi(\underset{\sim}{x}_{1i};\underset{\sim}{\theta})$ or $\partial\pi(\underset{\sim}{x}_{2i};\underset{\sim}{\theta})$ is the gradient vector with respect to $\underset{\sim}{\theta}$. Then, $\Delta(\underset{\sim}{\hat{\theta}}) = \underset{\sim}{0}$ give the solutions. In the case of logistic regression, they have used the Newton-Ralphson method to solve the equations, starting with $\underset{\sim}{\hat{\theta}} = \underset{\sim}{0}$, but we note that the Newton-Raphson's method is sensitive to these starts.

APPENDIX B: Calibration

We show how to calibrate the nps to the ps. Here, we simply need the population totals from the ps. We are assuming that the ps is very small, so that the totals from the ps are not very reliable. We can obtain the totals from a census, administrative records or web scraping. Let $\underset{\sim}{t}$ denote the vector of the p totals including the intercept; note that it appears Haziza and Beaumont (2017) did not use the intercept but this is necessary. Generally, the basic weighting system ensures consistency of a survey with a census by reducing nonsampling errors (e.g., response errors and coverage errors) and improves precision; see Haziza and Beaumont (2017) for more discussion.

Let the original survey weights in the ps be $w_j, j = 1,\ldots,n$. [Momentarily we drop the subscript on n.] We search for a calibrated weighting system $\tilde{w}_j, j = 1,\ldots,n$, such that $\sum_{j=1}^{n} \tilde{w}_j \underset{\sim}{x}_j = \underset{\sim}{t}$, the calibration equations. We want the $\tilde{w}_j, j = 1,\ldots,n$, to be as close as possible to $w_j, j = 1,\ldots,n$. Haziza and Beaumont (2017) judged closeness by a distance function, $G(u)$, where

a. $G(u) \geq 0$ and $G(1) = 0$;

b. $G(u)$ is differentiable, $g(u) = G'(u), g(1) = 0$, and strictly convex.

They need to minimize $\sum_{j=1}^{n} \frac{\tilde{w}_j}{q_j} G(\frac{\tilde{w}_j}{w_j})$ over $\tilde{w}_j, j = 1,\ldots,n$, where q_j denote the importance of unit j, subject to the constraint $\sum_{j=1}^{n} \tilde{w}_j \underset{\sim}{x}_j = \underset{\sim}{t}$.

Haziza and Beaumont (2017) considered the function,

$$\phi(\tilde{w_1},\ldots,\tilde{w_n},\underset{\sim}{\lambda}) = \sum_{j=1}^{n} \frac{\tilde{w}_j G(\frac{\tilde{w}_j}{w_j})}{q_j} - \underset{\sim}{\lambda}'(\sum_{j=1}^{n} \tilde{w}_j \underset{\sim}{x}_j - \underset{\sim}{t}),$$

where $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_p)'$ are Lagrangian multipliers. Differentiating $\phi(\tilde{w_1},\ldots,\tilde{w_n},\underset{\sim}{\lambda})$ with respect to $\tilde{w}_j$, they got

$$\tilde{w}_j = w_j g^{-1}(q_j \underset{\sim}{\lambda}' \underset{\sim}{x}_j), j = 1,\ldots,n.$$

Therefore,

$$\sum_{j=1}^{n} w_j g^{-1}(q_j \underset{\sim}{\lambda}' \underset{\sim}{x}_j) \underset{\sim}{x}_j = \underset{\sim}{t}.$$

Here, our method differs from Haziza and Beaumont (2020); they used the Newton-Ralphson method. We use the Nelder-Mead to minimize $\sum_{k=1}^{p} \mid \sum_{j=1}^{n} w_j g^{-1}(q_j \underset{\sim}{\lambda}' \underset{\sim}{x}_j) x_{jk} - t_k \mid$ over $\underset{\sim}{\lambda}$, forcing each component down to zero, to get $\underset{\sim}{\hat{\lambda}}$. The weights are then,

$$\tilde{w}_j = w_j g^{-1}(q_j \underset{\sim}{\hat{\lambda}}' \underset{\sim}{x}_j), j = 1,\ldots,n.$$

In our case, we use the simple Euclidean distance. This gives closed-form answers; apparently this was not recognized by Haziza and Beaumont (2017). We choose $q_j = 1, j = 1, \ldots, n$. For $G(u) = (u-1)^2$ is a legitimate distance function because $G(1) = 0$, $g(u) = 2(u-1)$, $g(1) = 0$, $g'(u) = 2 > 0$ and so $G(u)$ is strictly convex. Also $g^{-1}(y) = 1 + \frac{y}{2}$. In this case, we have

$$\sum_{j=1}^{n} w_j (1 + \frac{\lambda' x_j}{2}) x_{jk} = t_k, k = 1, \ldots, p.$$

That is,

$$\sum_{k'=1}^{p} \sum_{j=1}^{n} \lambda_{k'} x_{jk'} x_{jk} = 2(t_k - \sum_{j=1}^{n} w_j x_{jk}), k = 1, \ldots, p,$$

and

$$A\underset{\sim}{\lambda} = \underset{\sim}{b}, \;\; A = \left(\sum_{j=1}^{n} w_j x_{jk'} x_{ik} \right)_{(k',k)}, \;\; \underset{\sim}{b} = 2(\underset{\sim}{t} - \sum_{j=1}^{n} w_j \underset{\sim}{x}_j).$$

Therefore, assuming A is invertible, $\hat{\lambda} = A^{-1}\underset{\sim}{b}$ and the calibrated weights are $\tilde{w}_j = w_j(1 + \frac{\hat{\lambda}' x_j}{2}), j = 1, \ldots, n$. It is worth noting that if we use the ps to get $\underset{\sim}{t}$, $\sum_{j=1}^{n} w_j \underset{\sim}{x}_j = \underset{\sim}{t}$ and $\underset{\sim}{\lambda} = \underset{\sim}{0}$. If $\underset{\sim}{b}$ is closed to $\underset{\sim}{0}$, there will be little difference using calibration.

In our example on eight counties in California, we use web scraping. We got a population size of $N = 4,035,862$, and the other totals are age $= 36.7 \times N$, race $= .719 \times N$, sex $= .497 \times N$. We got $\hat{\lambda}_1 = .000713$, $\hat{\lambda}_2 = -0.000001$, $\hat{\lambda}_3 = 0.002198$, $\hat{\lambda}_4 = -0.000124$. It is not surprising then that calibrated weights are barely different from the original weights.

APPENDIX C: Joint Posterior Density when PS is the Actual Sample

We obtain a random sampler to draw $\underset{\sim}{\beta}, \sigma^2, a \mid \underset{\sim}{y}$ and we show that the joint posterior density is proper. Letting $\underset{\sim}{y} = (\underset{\sim}{y}_1, \underset{\sim}{y}_2)$, the joint posterior density is

$$\pi(\underset{\sim}{\beta}, \sigma^2, a \mid \underset{\sim}{y}) \propto a^{n_1/2} (\frac{1}{\sigma^2})^{\frac{n_1+n_2}{2}+1} e^{-\frac{1}{2\sigma^2}Q}, 0 \le a \le 1,$$

where $Q = a\sum_{i=1}^{n_1} w_{1i}(y_{1i} - \underset{\sim}{x}_{1i}\underset{\sim}{\beta})^2 + \sum_{i=1}^{n_2} w_{i2}(y_{2i} - \underset{\sim}{x}_{2i}\underset{\sim}{\beta})^2$. For convenience, letting $a_1 = a, a_2 = 1$ for the case when the nps is used as the prior.

First, let us look at Q and assume the design matrix is full rank at least for the ps. We find the conditional posterior density of $\underset{\sim}{\beta}$, which clearly has a multivariate normal density. We now decide its mean and variance using a standard trick by differentiation. First, letting $\Delta(\underset{\sim}{\beta}) = Q$, we have

$$\Delta'(\underset{\sim}{\beta}) = 2\sum_{s=1}^{2} \sum_{i=1}^{n_s} a_s w_{si} (y_{si} - \underset{\sim}{x}'_{si}\underset{\sim}{\beta}) \underset{\sim}{x}_{si}$$

and the Hessian matrix is

$$\Delta''(\underset{\sim}{\beta}) = 2\sum_{s=1}^{2} \sum_{i=1}^{n_s} a_s w_{si} \underset{\sim}{x}_{si} \underset{\sim}{x}'_{si}.$$

Now, setting $\Delta'(\underset{\sim}{\beta}) = \underset{\sim}{0}$, we get

$$\hat{\underset{\sim}{\beta}} = A^{-1}\underset{\sim}{b}, A = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si} \underset{\sim}{x}_{si}\underset{\sim}{x}'_{si},\ \underset{\sim}{b} = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si}\underset{\sim}{x}_{si} y_{si}.$$

Therefore,

$$\underset{\sim}{\beta} \mid \sigma^2, a, \underset{\sim}{y} \sim \text{Normal}(\hat{\underset{\sim}{\beta}}, \sigma^2 A^{-1}). \tag{C.1}$$

Second, integrating out $\underset{\sim}{\beta}$ from the joint posterior density, we get

$$\pi(\sigma^2, a \mid \underset{\sim}{y}) \propto a^{n_1/2}\left(\frac{1}{\sigma^2}\right)^{(n_1+n_2-p)/2+1} \mid A \mid^{1/2} e^{-\frac{1}{\sigma^2}\sum_{s=1}^{2}\sum_{i=1}^{n_s}(y_{si}-\underset{\sim}{x}'_{si}\hat{\underset{\sim}{\beta}})^2}.$$

Then, letting $d = \sum_{s=1}^{2}\sum_{i=1}^{n_s} a_s w_{si}(y_{si} - \underset{\sim}{x}'_{si}\hat{\underset{\sim}{\beta}})^2$,

$$\sigma^2 \mid a, \underset{\sim}{y} \sim \text{IGam}\left(\frac{n_1+n_2-p}{2}, \frac{d}{2}\right). \tag{C.2}$$

Finally, integrating out σ^2, we have

$$\pi(a \mid \underset{\sim}{y}) \propto \frac{a^{n_1/2} \mid A \mid^{-1/2}}{d^{(n_1+n_2-p)/2}}, 0 \le a \le 1. \tag{C.3}$$

Because $0 \le a \le 1$, all quantities are well defined, and provided the design matrix of the ps is full rank, the joint posterior density is proper. Draws can be made from the joint posterior density using the multiplication rule of probability, drawing samples from (C.3), (C.2) and (C.1) in that order. Samples can be drawn from $\pi(a \mid \underset{\sim}{y})$ in (C.3) using the grid method.

APPENDIX D: Computation for the Small Area Model

We discuss how to fit the model in (3). Recall $\Omega_1 = (\underset{\sim}{a}, \theta, \gamma, \rho)$ and $\Omega_2 = (\underset{\sim}{\nu}, \underset{\sim}{\beta}, \sigma^2)$. Our strategy is to integrate out Ω_2 from $\pi(\Omega_1, \Omega_2 \mid \underset{\sim}{y})$ to get $\pi(\Omega_1 \mid \underset{\sim}{y})$ and then sample $\tilde{\pi}(\Omega_1 \mid \underset{\sim}{y})$ using the Gibbs sampler.

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

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

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n+\ell}{2}+1}\left(\frac{1-\rho}{\rho}\right)^{\ell/2}\prod_{i=1}^{\ell}\left[e^{-\frac{1}{2\rho\sigma^2}\left\{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(y_{sij}-\underset{\sim}{x}'_{sij}\underset{\sim}{\beta}-v_i)^2+(1-\rho)v_i^2\right\}}\right]. \tag{D.1}$$

We will integrate out Ω_2. Momentarily, we will drop $\pi(\Omega_1)$, but we will retain $\prod_{i=1}^{\ell}\sqrt{a_i}$.

Define the following quantities,

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

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

$$\tilde{y}_{sij} = y_{sij} - \bar{\bar{y}}_i, \;\; \underset{\sim}{\tilde{x}}_{sij} = \underset{\sim}{x}_{sij} - \underset{\sim}{\bar{\bar{x}}}_i.$$

Note that while the λ_i are functions of ρ, the ϕ_{sij}, $\bar{\bar{y}}_i$ and $\underset{\sim}{\bar{\bar{x}}}_i$ are not functions of ρ.

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

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

as

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

Then, it is easy to show that

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

where $\hat{\nu}_i = \lambda_i(\bar{\bar{y}}_i - \underset{\sim}{\bar{\bar{x}}}'_i\underset{\sim}{\beta})$. 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 $\underset{\sim}{\beta}$ and σ^2.

Then, integrating out the ν_i from (D.1), we have

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

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

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,

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

where

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

and

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

Then integrating $\underset{\sim}{\beta}$ from (D.2), we have

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

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

Therefore,

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

where

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

Integrating out σ^2 from (D.3), we have

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

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

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

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

and therefore,

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

$\frac{\gamma}{1-\gamma} \le \theta \le \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3, 0 \le \rho \le 1$. It is worth noting that the a_i are not independent; Δ and Q contain all the a_i, which is contained by λ_i also. 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 $\underset{\sim}{a}_{(i)} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_\ell)', i = 1, \ldots, \ell$ (a_i is eliminated), then for $0 < a_i < 1$,

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

Second, the CPD of ρ is

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

Third, the joint CPD of (θ, γ) is

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

$\frac{\gamma}{1-\gamma} \leq \theta \leq \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3$. The CPD of θ or γ 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 (ρ, θ, γ); hyperparameters are more difficult to sample. We have done this, and we have reduced the entire computation time from 15 minutes to 8 minutes with little change in the results.

APPENDIX E: Estimation of ρ_{RY}

The main difficulty is how to get control over ρ_{RY} or D_I. However, if we can field a small probability survey, we can use the magic of probability sampling to get a reasonable estimate of ρ_{RY}, and therefore we can provide a better study of the quality of the non-probability sample. It appears that there is no other sensible way to get control over ρ_{RY}.

Given a population of R and Y, the finite population correlation coefficient is given by

$$\rho_{RY} = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i}^{N}(Y_1 - \bar{Y})^2}}$$

We have

$$\rho_{RY} = \sqrt{\frac{f}{1-f}} \frac{(\bar{y} - \bar{Y})}{\sigma_Y},$$

where f is the sampling fraction, $\bar{Y}$ is the population mean and σ_Y is the standard deviation of Y. With a small probability sample, we can estimate $\bar{Y}$ and σ_Y using inverse probability weighted estimators. Assuming that there is a probability sample $\{(y_i, W_i), i = 1, \ldots, n\}$. Then a design-unbiased estimator of $\bar{Y}$ is $\bar{y}_W$, σ_Y^2 is estimated by $\hat{\sigma}_Y^2 = \sum_{i=1}^{n} W_i (y_i - \bar{y}_W)^2 / \sum_{i=1}^{n} W_i$,

$$\hat{\rho}_{RY} = \sqrt{\frac{f}{1-f}} \frac{(\bar{y} - \bar{y}_W)}{\hat{\sigma}_Y},$$

and we can study the quality of our non-probability sample. If Y is binary,

$$\hat{\rho}_{RY} = \sqrt{\frac{f}{1-f}}\frac{(\bar{y}-\hat{a})}{\hat{\sigma}_Y},$$

where $\hat{\sigma} = \sqrt{\hat{a}(1-\hat{a})}$ with $\hat{a} = \frac{\hat{N}_1}{\hat{N}}$, $\hat{N}_1 = \sum_{i=1}^{n} W_i y_i$, $\hat{N} = \sum_{i=1}^{n} W_i$.

For the single sample example on BMI data, we found that $\hat{\rho}_{RY} = .006$. Meng (2018) stated that such a value of $\hat{\rho}_{RY}$ is large enough to make the selection bias very destructive, especially in data sets with millions of records, much bigger than ours.

REFERENCES

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 (401), 28-36.

Beaumont, J-F. (2020), Are Probability Surveys Bound to Disappear for the Production of Official Statistics? *Survey Methodology*, 46 (1), 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.

Chen, Y., Li, P. and Wu, C. (2020), Doubly Robust Inference With Nonprobability Survey Samples, *Journal of the American Statistical Association*, 115 (532), 2011-2021.

Choi, S., Nandram, B. and Kim, D. (2021), Bayesian Predictive Inference of Small Area Proportions Under Selection Bias, *Survey Methodology*, 47 (1), 91-122.

Citro, C. (2014), From Multiple Modes for Surveys to Multiple Data Sources for Estimates, *Survey Methodology*, 40, 137-161.

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. (2009), Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights, *Survey Practice*, 2, 1–9.

Elliott, M. R. and R. Valliant (2017), Inference for Nonprobability Samples, *Statistical Science*, 32, 249–264.

Ibrahim, J. G. and Chen, M-H. (2000), Power Prior Distributions for Regression Models, *Statistical Science*, 15 (1), 46-60.

Ibrahim, J. G., Chen, M-H., Gwon, Y. and Chen, F. (2015), The Power Prior: Theory and Applications, *Statistics in Medicine*, 34, 3724-3749.

Ishwaran, H. and James, L. F. (2001), Gibbs Sampling Methods for Stick-breaking Priors, *Journal of the American Statistical Association*, 96, 161-173.

Haziza, D. and Beaumont, J-F. (2017), Construction of Weights in Surveys: A Review, *Statistical Science*, 32 (2), 206-226.

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 (2), 852-885.

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 (2), 685–726.

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 (6) (in press).

Nandram, B., Cao, H., Xu, Z., and Bhadra, D. (2019), Bayesian Predictive Inference for Nonprobability Samples with Spatial Poststratification, *Technical Report*, Mathematical Sciences, Worcester Polytechnic Institute.

Pfeffermann, D. (1993), The role of Sampling Weights When Modeling Survey Data, International Statistical Review/Revue Internationale de Statistique, 61, 317–337.

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 (418), 383-396.

Rao, J. N. K. (2020), On Making Valid Inferences by Integrating Data from Surveys and Other Sources, *Sankhya*, Series B, 3-33.

Rafei, A., Flannagan,C. A. C., West, B. T. and Elliott, M. R. (2021) Robust Bayesian Inference for Big Data: Combining Sensor-based Records with Traditional Survey, arxiv:2101.07456Vi, pp. 1-61.

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 (3), 653-681.

Smith, T. M. F. (1983), On the Validity of Inferences from Non-random Samples, *Journal of the Royal Statistical Society*, Series A, 146, 393-403.

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.

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.

Valliant, R. (2020), Comparing Alternatives for Estimation from Nonprobability Samples, *Journal of Survey Statistics and Methodology*, 8, 231-263.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015), Forecasting Elections with Nonrepresentative Polls, *International Journal of Forecasting*, 31, 980-991.

Xu, Z. (2020). *Bayesian Predictive Inference for a Non-probability Sample with Binary Responses from Small Areas*, PhD Dissertation, Department of Mathematical Sciences, Worcester Polytechnic Institute.

Xu, Z. and Nandram, B. (2019), Bayesian Inference for Non-probability Samples with Binary Responses, *Technical Report*, Mathematical Sciences, Worcester Polytechnic Institute.

Xu, Z. and Nandram, B. (2019), Bayesian Inference of Non-probability Samples, JSM Proceedings, 2585–2593, Foundations in Bayesian Statistics Section. Alexandria, VA: American Statistical Association.

Xu, Z., Nandram, B. and Manandhar, B. (2020), Bayesian Inference of a Finite Population Mean Under Length-Biased Sampling, *Statistical Methods and Applications in Forestry and Environmental Sciences*, Eds. Girish Chandra, Raman Nautiyal and Hukum Chandra, pp. 79-103.

Yin, J. and Nandram, B. (2020a), A Bayesian Small Area Model with Dirichlet Processes on Responses, *Statistics in Transition, New Series*, 21 (3), 1-19.

Yin, J. and Nandram, B. (2020b), A Nonparametric Bayesian Analysis of Response Data with Gaps, Outliers and Ties, *Statistics and Applications, New Series*, 18 (2), 121-141.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015), Forecasting Elections with Non-representative Polls, *International Journal of Forecasting*, 31, 980-991.

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.