Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

An Approximate Bayesian Approach to Improving Probability Sample Estimators Using a Supplementary Non-Probability Sample

by Yong You, Abel DaSylva and Jean-François Beaumont

Release date: October 29, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

An Approximate Bayesian Approach to Improving Probability Sample Estimators Using a Supplementary Non-Probability Sample

Yong You, Abel DaSylva and Jean-François Beaumont[1]

Abstract

Non-probability samples are being increasingly explored by National Statistical Offices as a complement to probability samples. We consider the scenario where the variable of interest and auxiliary variables are observed in both a probability and non-probability sample. Our objective is to use data from the non-probability sample to improve the efficiency of survey-weighted estimates obtained from the probability sample. Recently, Sakshaug, Wisniowski, Ruiz and Blom (2019) and Wisniowski, Sakshaug, Ruiz and Blom (2020) proposed a Bayesian approach to integrating data from both samples for the estimation of model parameters. In their approach, non-probability sample data are used to determine the prior distribution of model parameters, and the posterior distribution is obtained under the assumption that the probability sampling design is ignorable (or not informative). We extend this Bayesian approach to the prediction of finite population parameters under non-ignorable (or informative) sampling by conditioning on appropriate survey-weighted statistics. We illustrate the properties of our predictor through a simulation study.

Key Words: Bayesian prediction; Gibbs sampling; Non-ignorable sampling; Statistical data integration.

1. Introduction

Non-probability samples are being increasingly considered by government agencies as a means of reducing survey costs and obtaining more timely estimates than most probability surveys. In particular, following the COVID-19 pandemic, Statistics Canada has launched a series of online volunteer surveys, called crowdsourcing surveys, to obtain information on different aspects of the life of the Canadian population. However, it has been known for decades that the use of non-probability samples alone, such as crowdsourcing samples, may lead to estimates that suffer from significant selection bias. There has been a growing amount of research in recent years on methods that use auxiliary data from a probability sample to reduce the selection bias of non-probability sample estimators. These methods are applicable in the scenario where the variables of interest are observed only in the non-probability sample, but common auxiliary variables are observed in both samples. Propensity score weighting (e.g., Chen, Li and Wu, 2020) provides one approach for integrating data from both samples. It consists of weighting the non-probability sample participants by the inverse of their estimated participation probability. Statistical matching or sample matching (e.g., Yang, Kim and Hwang 2021; or Rivers, 2007) is an alternative. It consists of imputing the missing variables of interest in the probability sample using non-probability sample data. Recent reviews on statistical data integration under this scenario are given in Beaumont (2020), Elliot and Valliant (2017), Rao (2021) and Valliant (2020).

We consider a different data integration scenario, where the variables of interest and the auxiliary variables are observed in both samples. We do not assume the population totals of the auxiliary variables to be known, or the indicator of participation in the non-probability sample to be observed in the probability sample. The literature on this scenario is scarce. Elliott and Haviland (2007) proposed a composite estimator, which is simply a weighted average of the non-probability and probability sample estimators. They pointed out that their proposed estimator requires a relatively large probability sample in order to estimate with sufficient precision the bias of the non-probability sample estimator. More recently, Bayesian estimators that integrate the two samples have been proposed by Sakshaug,

[1]Yong You, Abel DaSylva and Jean-François Beaumont, ICMIC, Statistics Canada, 100 Tunney's pasture driveway, Ottawa ON, K1A0T6. Contact email: yong.you@statcan.gc.ca, abel.dasylva@statcan.gc.ca and jean-francois.beaumont@statcan.gc.ca. **Disclaimer**: The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those implemented by the Agency.

Wisniowski, Ruiz and Blom (2019), Wisniowski, Sakshaug, Ruiz and Blom (2020) and Nandram and Rao (2021). Assuming an ignorable probability sampling design, Sakshaug et al. (2019) and Wisniowski et al. (2020) proposed and assessed through simulation studies an interesting Bayesian approach to integrating data from both samples for the estimation of model parameters. Their approach involves some intriguing prior distributions for the model parameters: non-probability sample data are used to determine the prior mean, but data from both samples are used to determine the prior variance. Indeed, the estimated bias of the non-probability sample estimator is used to inflate the prior variance. In this paper, we extend their Bayesian approach to the estimation of finite population means under a non-ignorable probability sampling design and evaluate it in a simulation study.

2. Estimation problem

Suppose that we are interested in estimating the population mean $\theta = N^{-1}\sum_{i\in U} y_i$, where y_i is the value of the variable of interest *y* for unit *i* of the finite population *U* of size *N*. A probability sample s of size *n* is drawn from *U* using some probability sampling design, and the variable of interest *y* is observed for all units $i \in s$. The survey-weighted estimator of the population mean is $\hat{\theta}_w = \hat{N}^{-1}\sum_{i\in s} w_i y_i$, where $\hat{N} = \sum_{i\in s} w_i$ and w_i is a survey weight for sample unit *i*. The survey weight can be the basic design weight $w_i = 1/\pi_i$, where π_i is the probability that unit *i* is selected in s, or it can be a calibration weight (e.g., Deville and Särndal, 1992). The variable of interest *y* is also observed for all units of a non-probability sample s_{NP} of size n_{NP}. In addition, we observe a vector of *q* auxiliary variables, $\boldsymbol{x}_i$, for all units in both s and s_{NP}. We suppose that $\boldsymbol{x}_i$ includes an intercept.

We postulate the linear model

$$y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \rightarrow N(\boldsymbol{x'}_i\boldsymbol{\beta}, \sigma^2),\ i \in U, \tag{1}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown model parameters. The idea in Sakshaug et al. (2019) and Wiśniowski et al. (2020) is to use the non-probability sample to obtain the prior mean of $\boldsymbol{\beta}$. Like these authors, we consider the following prior distribution for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \rightarrow N\left(\widehat{\boldsymbol{\beta}}_{NP}, \boldsymbol{\Phi}_0\right), \tag{2}$$

where $\widehat{\boldsymbol{\beta}}_{NP} = \left(\sum_{i\in s_{NP}} \boldsymbol{x}_i\boldsymbol{x'}_i\right)^{-1}\sum_{i\in s_{NP}} \boldsymbol{x}_i y_i$ is the maximum likelihood estimator of $\boldsymbol{\beta}$, assuming non-probability sample selection is ignorable (see Rubin, 1976) with respect to model (1), and $\boldsymbol{\Phi}_0$ is a known specified variance-covariance matrix. Greater detail on the choice of $\boldsymbol{\Phi}_0$ is given in Section 3.3. We assume a non-informative prior for σ^2, with the probability density function proportional to σ^{-2}.

Assuming the probability sampling design is ignorable for model-based inferences (e.g., the probability sample is selected using simple random sampling), Sakshaug et al. (2019) and Wiśniowski et al. (2020) estimate $\boldsymbol{\beta}$ from its posterior mean given the observed data, $\boldsymbol{Y}_s = \{y_i, i \in s\}$. Our objective is different. We are interested in predicting the finite population mean $\theta = N^{-1}\sum_{i\in U} y_i$. Assuming $\boldsymbol{x}_i$ is available for the entire population *U* as well as ignorable sampling, we can compute the posterior mean of θ as

$$\tilde{\theta} = E(\theta|\boldsymbol{Y}_s) = N^{-1}\left(\sum_{i\in s} y_i + \sum_{i\in U-s} \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}\right) = N^{-1}\left(\sum_{i\in U} \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}} + \sum_{i\in s}\left(y_i - \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}\right)\right), \tag{3}$$

where $\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{Y}_s)$ is the posterior mean of $\boldsymbol{\beta}$. The posterior mean of θ, given in (3), is computable provided *N* and $\sum_{i\in U} \boldsymbol{x}_i$ are known. In many applications, this is not the case. The unknown population quantities in (3) can be replaced with survey-weighted estimators to obtain

$$\hat{\theta} = \hat{N}^{-1}\left(\sum_{i\in s} w_i\boldsymbol{x'}_i\widehat{\boldsymbol{\beta}} + \sum_{i\in s}\left(y_i - \boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}\right)\right). \tag{4}$$

If the sampling fraction $f = n/N$ is small, the left-hand side of (4) can be approximated as

$$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}) = \hat{N}^{-1}\sum_{i\in s} w_i\boldsymbol{x'}_i\widehat{\boldsymbol{\beta}}. \tag{5}$$

The pseudo maximum likelihood estimator of $\boldsymbol{\beta}$ based on the probability sample data is

$$\widehat{\boldsymbol{\beta}}_w = (\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1} \sum_{i\in s} w_i \boldsymbol{x}_i y_i.$$

Under regularity conditions, $\widehat{\boldsymbol{\beta}}_w$ is consistent for $\boldsymbol{\beta}$ under model (1) and the sampling design. It is straightforward to show that $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_w)$ reduces to the survey-weighted estimator $\hat{\theta}_w = \widehat{N}^{-1} \sum_{i\in s} w_i y_i$, provided an intercept is included in $\boldsymbol{x}_i$. Therefore, the estimator (5) is expected to achieve non-negligible gains in efficiency over the survey-weighted estimator $\hat{\theta}_w$ only when data from both samples are combined for the estimation of $\boldsymbol{\beta}$. The Bayesian approach that we consider accomplishes this data integration by using data from the non-probability sample to determine the prior mean of $\boldsymbol{\beta}$. From (4) and (5), we also observe that both $\hat{\theta}$ and $\hat{\theta}_a$ are expected to be close to $\hat{\theta}_w$ when the linear model (1) has a strong predictive power (small σ^2) so that $y_i \approx \boldsymbol{x}'_i \widehat{\boldsymbol{\beta}}$. This begs the question: Should any auxiliary variable be used for the estimation of finite population parameters?

3. Bayesian inference

3.1 Ignorable probability sampling design

Under certain conditions (see Rubin, 1976), the probability sampling design can be ignored when making model-based inferences (conditional on *s*). In that case, standard results on Bayesian linear regression can be used to obtain the conditional posterior distributions of $\boldsymbol{\beta}$ and σ^2. The conditional posterior distribution of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\boldsymbol{Y}_s, \sigma^2 \rightarrow N\left(\widehat{\boldsymbol{\beta}}_c, \sigma^2 \boldsymbol{\Phi}_c^{-1}\right), \tag{6}$$

where $\boldsymbol{\Phi}_c = \sum_{i\in s} \boldsymbol{x}_i \boldsymbol{x}'_i + \sigma^2 \boldsymbol{\Phi}_0^{-1}$ and $\widehat{\boldsymbol{\beta}}_c = \boldsymbol{\Phi}_c^{-1}\left(\sum_{i\in s} \boldsymbol{x}_i y_i + \sigma^2 \boldsymbol{\Phi}_0^{-1} \widehat{\boldsymbol{\beta}}_{NP}\right)$.

The estimator $\widehat{\boldsymbol{\beta}}_c$ reduces to the unweighted estimator $\widehat{\boldsymbol{\beta}}_{uw} = (\sum_{i\in s} \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1} \sum_{i\in s} \boldsymbol{x}_i y_i$ when a non-informative prior for $\boldsymbol{\beta}$ is used (e.g., when $\boldsymbol{\Phi}_0^{-1}$ is specified close to $\mathbf{0}$). The estimator $\widehat{\boldsymbol{\beta}}_{uw}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under the assumption that the sampling design is ignorable.

The conditional posterior distribution of σ^2 is given by

$$\sigma^2|\boldsymbol{Y}_s, \boldsymbol{\beta} \rightarrow IG\left(\frac{n}{2}, \frac{n}{2} s_{uw}^2(\boldsymbol{\beta})\right), \tag{7}$$

where $IG(a, b)$ stands for the inverse gamma distribution with shape parameter a and scale parameter b, and

$$s_{uw}^2(\boldsymbol{\beta}) = n^{-1} \sum_{i\in s} (y_i - \boldsymbol{x}'_i \boldsymbol{\beta})^2.$$

Using (6) and (7), the iterative Gibbs sampling procedure for Bayesian inference can be applied to generate a large number of values from the posterior distribution $\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}_s$. These generated values allow for approximating the posterior distribution of $\boldsymbol{\beta}$ and σ^2. In particular, Bayesian estimation of $\boldsymbol{\beta}$ is obtained by approximating the posterior mean $\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{Y}_s)$ by the average of the generated values of $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}^*$. The population mean θ is then estimated by $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$.

Under model (1), it is well known that

$$\widehat{\boldsymbol{\beta}}_{uw}\big|\boldsymbol{\beta}, \sigma^2 \rightarrow N(\boldsymbol{\beta}, \sigma^2 (\sum_{i\in s} \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1}) \tag{8}$$

and

$$s_{uw}^2(\boldsymbol{\beta})|\boldsymbol{\beta}, \sigma^2 \rightarrow G\left(\frac{n}{2}, 2\frac{\sigma^2}{n}\right), \tag{9}$$

where $G(a, b)$ stands for the gamma distribution with shape parameter a and scale parameter b. Using (8) and (9), it is not difficult to show that the distribution $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{uw}, \sigma^2$ is identical to the conditional posterior distribution given in (6), and the distribution $\sigma^2|s_{uw}^2(\boldsymbol{\beta}), \boldsymbol{\beta}$ is identical to the conditional posterior distribution given in (7). This observation is key to understanding the main idea behind the extension of this approach to the case of a non-ignorable sampling design. In the next section, we will develop a Gibbs sampling procedure for non-ignorable sampling designs using survey-weighted analogues of (8) and (9).

3.2 Non-ignorable probability sampling design

In practice, from a frequentist model-design perspective, the unweighted estimator $\widehat{\boldsymbol{\beta}}_{uw}$ is used only when the survey weight w_i is the same for all $i \in s$. Otherwise, it is more common to use the survey-weighted estimator $\widehat{\boldsymbol{\beta}}_w$ as it remains consistent under a non-ignorable probability sampling design. We make the usual assumption

$$\widehat{\boldsymbol{\beta}}_w|\boldsymbol{\beta}, \sigma^2 \to N(\boldsymbol{\beta}, \sigma^2\boldsymbol{\Psi}) \,, \tag{10}$$

where $\boldsymbol{\Psi} = \sigma^{-2}\, var_{mp}(\widehat{\boldsymbol{\beta}}_w)$. The subscript *m* refers to model (1) whereas the subscript *p* refers to the probability sampling design. Regularity conditions for the validity of (10) are given in Fuller (2009). Using standard arguments (e.g., Binder and Roberts, 2003), we have

$$\boldsymbol{\Psi} \approx (\textstyle\sum_{i\in U} \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1} + \sigma^{-2} E_m\big[var_p(\widehat{\boldsymbol{\beta}}_w)\big] \,. \tag{11}$$

Note that the first term on the right-hand side of (11) is negligible when the sampling fraction *f* is negligible. Under an ignorable sampling design, it is straightforward to show that

$$\boldsymbol{\Psi} = \sigma^{-2} E_p\big[var_m(\widehat{\boldsymbol{\beta}}_w)\big] = E_p[(\textstyle\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1}(\sum_{i\in s} w_i^2 \boldsymbol{x}_i \boldsymbol{x}'_i)(\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}'_i)^{-1}] \,,$$

which does not depend on $\boldsymbol{\beta}$ or σ^2.

By analogy with the case of ignorable sampling, we consider the conditional posterior distribution $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2$. This simple and clever idea of handling non-ignorable sampling, by conditioning on a consistent estimator, was proposed by Wang, Kim and Yang (2018). From assumption (10) and prior distribution (2), we obtain the conditional posterior distribution

$$\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2 \to N\big(\widehat{\boldsymbol{\beta}}_{wc}, \sigma^2\boldsymbol{\Phi}_{wc}^{-1}\big) \,, \tag{12}$$

where $\boldsymbol{\Phi}_{wc} = \boldsymbol{\Psi}^{-1} + \sigma^2\boldsymbol{\Phi}_0^{-1}$ and $\widehat{\boldsymbol{\beta}}_{wc} = \boldsymbol{\Phi}_{wc}^{-1}\big(\boldsymbol{\Psi}^{-1}\widehat{\boldsymbol{\beta}}_w + \sigma^2\boldsymbol{\Phi}_0^{-1}\widehat{\boldsymbol{\beta}}_{NP}\big)$.

From a frequentist model-design perspective, $s_{uw}^2(\boldsymbol{\beta})$ is not consistent for σ^2. A survey-weighted version of $s_{uw}^2(\boldsymbol{\beta})$ is $s_w^2(\boldsymbol{\beta}) = \widehat{N}^{-1}\sum_{i\in s} w_i(y_i - \boldsymbol{x}'_i\boldsymbol{\beta})^2$. Under regularity conditions, $s_w^2(\boldsymbol{\beta})$ is asymptotically unbiased and consistent for σ^2 under model (1) and the sampling design. We assume that

$$s_w^2(\boldsymbol{\beta})|\boldsymbol{\beta}, \sigma^2 \to G\left(\frac{\lambda}{2}, 2\,\frac{\sigma^2}{\lambda}\right) \,, \tag{13}$$

where $2\lambda^{-1} = \sigma^{-4}\, var_{mp}[s_w^2(\boldsymbol{\beta})]$. The quantity λ can be written as $\lambda = \frac{n}{D}$, where $D = \left(\frac{2\sigma^4}{n}\right)^{-1} var_{mp}[s_w^2(\boldsymbol{\beta})]$ can be interpreted as a design effect. Under regularity conditions, we obtain

$$2\lambda^{-1} \approx \frac{2}{N} + \sigma^{-4} E_m\big\{var_p[s_w^2(\boldsymbol{\beta})]\big\} \,. \tag{14}$$

Again, the first term on the right-hand side of (14) is negligible when the sampling fraction *f* is negligible. Under an

ignorable sampling design, we obtain

$$2\lambda^{-1} = \sigma^{-4} E_p\{var_m[s_w^2(\boldsymbol{\beta})]\} = 2E_p\left(\frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right),$$

which does not depend on $\boldsymbol{\beta}$ or σ^2.

Similar to the case of ignorable sampling, we consider the conditional posterior distribution $\sigma^2|s_w^2(\boldsymbol{\beta}), \boldsymbol{\beta}$. Using (13) and a prior probability density function for σ^2 proportional to σ^{-2}, we obtain

$$\sigma^2|s_w^2(\boldsymbol{\beta}), \boldsymbol{\beta} \rightarrow IG\left(\frac{\lambda}{2}, \frac{\lambda}{2} s_w^2(\boldsymbol{\beta})\right). \tag{15}$$

Using (12) and (15), the Gibbs sampling procedure can be applied to generate a large number of posterior values for $\boldsymbol{\beta}$ and σ^2. The population mean θ is estimated by $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$, where $\widehat{\boldsymbol{\beta}}_{NI}^*$ is the average of the generated values of $\boldsymbol{\beta}$.

In practice, $\boldsymbol{\Psi}$ and λ are unknown and must be replaced by consistent estimators before generating posterior values for $\boldsymbol{\beta}$ and σ^2 using (12) and (15). From (11), a consistent estimator of $\boldsymbol{\Psi}$ is

$$\widehat{\boldsymbol{\Psi}} = \left(\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}'_i\right)^{-1} + \frac{\boldsymbol{v}_p(\widehat{\boldsymbol{\beta}}_w)}{\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)}, \tag{16}$$

where $\boldsymbol{v}_p(\widehat{\boldsymbol{\beta}}_w)$ is a design-consistent estimator of $var_p(\widehat{\boldsymbol{\beta}}_w)$, obtained using linearization or replication variance estimation methods, and

$$\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w) = \frac{n}{n-q} s_w^2(\widehat{\boldsymbol{\beta}}_w).$$

From (14), a consistent estimator of $2\lambda^{-1}$ is

$$2\hat{\lambda}^{-1} = \frac{2}{\hat{N}} + \frac{v_p[s_w^2(\boldsymbol{\beta})]|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_w}}{\left(\tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)\right)^2}, \tag{17}$$

where $v_p[s_w^2(\boldsymbol{\beta})]$ is a design-consistent estimator of $var_p[s_w^2(\boldsymbol{\beta})]$.

3.3 Prior choice

When a non-informative prior for $\boldsymbol{\beta}$ is used (e.g., when $\boldsymbol{\Phi}_0^{-1}$ is specified close to $\mathbf{0}$), the conditional posterior distribution (12) reduces to $\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w, \sigma^2 \rightarrow N\left(\widehat{\boldsymbol{\beta}}_w, \sigma^2\boldsymbol{\Psi}\right)$. Therefore, the posterior mean $\widehat{\boldsymbol{\beta}} = E\left(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_w\right) = \widehat{\boldsymbol{\beta}}_w$ and the estimator (5) becomes $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_w) = \hat{N}^{-1}\sum_{i\in s} w_i \boldsymbol{x}'_i \widehat{\boldsymbol{\beta}}_w = \hat{N}^{-1}\sum_{i\in s} w_i y_i = \hat{\theta}_w$. The Bayesian approach has thus no advantage over the frequentist design-based approach when a non-informative prior for $\boldsymbol{\beta}$ is used.

In the context of an ignorable sampling design, Sakshaug et al. (2019) suggested setting $\boldsymbol{\Phi}_0$ to a diagonal matrix with the j^{th} element on the diagonal equal to $\Phi_{0j} = \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2$, where $\hat{\beta}_{NP,j}$ and $\hat{\beta}_{w,j}$ are the j^{th} elements of $\widehat{\boldsymbol{\beta}}_{NP}$ and $\widehat{\boldsymbol{\beta}}_w$, respectively. The idea is to account for the bias of $\widehat{\boldsymbol{\beta}}_{NP}$ through the specification of the prior variance-covariance matrix for $\boldsymbol{\beta}$. We refer to this prior specification of $\boldsymbol{\Phi}_0$ as the "distance" specification. Sakshaug et al. (2019) pointed out the following caveat:

"*By using the probability-based estimator to construct the prior distribution, the question of using data twice arises. We address this issue by pointing out that the ML estimator from the probability sample (a measure of central tendency) is used to inform the variance, rather than the mean. Further, we use the information from the probability data only in relative comparison to the nonprobability sample. Hence, any potential shrinkage in posterior variance depends on the combination of both data sets, rather than the probability data alone.*"

In the context of a non-ignorable sampling design, the argument is still valid, but the "probability-based estimator" and the "ML estimator from the probability sample" are replaced with the pseudo maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_w$.

Pursuing this idea further, we suggest setting the prior variance-covariance matrix $\boldsymbol{\Phi}_0$ to a diagonal matrix with the j^{th} element on the diagonal equal to

$$\Phi_{0j} = max\left[s_{NP}^2 \Psi_{NP,j}, \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2 - \tilde{s}_w^2(\widehat{\boldsymbol{\beta}}_w)\widehat{\Psi}_j\right], \tag{18}$$

where $s_{NP}^2 = (n_{NP} - q)^{-1} \sum_{i \in s_{NP}} \left(y_i - \boldsymbol{x}'_i \widehat{\boldsymbol{\beta}}_{NP}\right)^2$, and $\Psi_{NP,j}$ and $\widehat{\Psi}_j$ are the j^{th} diagonal elements of the matrices $\left(\sum_{i \in s_{NP}} \boldsymbol{x}_i \boldsymbol{x}'_i\right)^{-1}$ and $\widehat{\boldsymbol{\Psi}}$, respectively. The first element in the max function in (18) is an estimator of the variance of $\hat{\beta}_{NP,j}$ under model (1) and assuming the non-probability selection mechanism is ignorable. The second element is a design-based estimator of the square bias $\left(\hat{\beta}_{NP,j} - \beta_j\right)^2$, where β_j is the j^{th} element of $\boldsymbol{\beta}$. We refer to this prior specification of $\boldsymbol{\Phi}_0$ as the "bias-variance" specification.

Wiśniowski et al. (2020) suggested a few prior variance-covariance matrices of the form $\boldsymbol{\Phi}_0 = \sigma^2 k_0 \mathbf{V}$, where k_0 is a specified constant and $\mathbf{V}$ is a matrix of hyperparameters. A specification that seemed to perform well in their simulation study is $k_0 = \frac{1}{log(n_{NP})}$ and setting $\mathbf{V}$ to a diagonal matrix with the j^{th} element on the diagonal equal to $V_j = max\left[s_{NP}^2 \Psi_{NP,j}, \left(\hat{\beta}_{NP,j} - \hat{\beta}_{w,j}\right)^2\right]$. Wiśniowski et al. (2020) refer to this prior specification of $\boldsymbol{\Phi}_0$ as the "conjugate-distance" specification. It has a form similar to the bias-variance specification but with a scaling factor $\sigma^2 k_0$.

4. Simulation study

4.1 Simulation setup

We conducted a design-based simulation study, following Hidiroglou and You (2016), to evaluate the proposed Bayesian estimator of the population mean and compared it with the Horvitz-Thompson estimator. We created a population consisting of N = 1000 population units as follows:

i) x_i was generated from a gamma distribution with mean μ_x and variance σ_x^2, where μ_x and σ_x^2 are pre-determined constants.

ii) z_i was generated from a gamma distribution with mean μ_z and variance σ_z^2, where μ_z and σ_z^2 are pre-determined constants.

iii) $y_i = \alpha_1 x_i + \alpha_2 z_i + \delta_i$, where $\delta_i \rightarrow N(0, \sigma_\delta^2)$ and σ_δ^2 is a pre-determined constant. The choice of α_1 and α_2 are discussed below. Note that $E(y_i|x_i) = \alpha_0 + \alpha_1 x_i$, where $\alpha_0 = \alpha_2 \mu_z$.

Then, from the population, the non-probability sample was selected by choosing 300 units that have the smallest values of y_i. Probability samples of size 20 and 50 were drawn with Probability Proportional to Size Without Replacement (PPSWR), with z_i as the size measure, as in You, Rao and Kovacevic (2003) and Hidiroglou and You (2016). For estimation, the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ was considered, where $\varepsilon_i \rightarrow N(0, \sigma_\varepsilon^2)$.

To determine a suitable choice of α_1 and α_2, let us first define $e_i = y_i - E(y_i|x_i) = -\alpha_0 + \alpha_2 z_i + \delta_i$. The degree of sampling informativeness depends on the correlation between e_i and z_i. Let us define the square correlation coefficient $\rho_{ez}^2 = \frac{[cov(e_i, z_i)]^2}{var(e_i)\, var(z_i)}$. It is straightforward to show that $\rho_{ez}^2 = \frac{\alpha_2^2 \sigma_z^2}{\alpha_2^2 \sigma_z^2 + \sigma_\delta^2}$. Solving for α_2 yields :

$$\alpha_2 = \sqrt{\frac{\sigma_\delta^2 \rho_{ez}^2}{\sigma_z^2(1-\rho_{ez}^2)}}. \quad (19)$$

The value of ρ_{ez}^2 is set to a pre-determined constant. Similarly, we can also define the square correlation coefficient between x_i and y_i as $\rho_{xy}^2 = \frac{[cov(x_i,y_i)]^2}{var(x_i)\,var(y_i)}$. Again, we can show that $\rho_{xy}^2 = \frac{\alpha_1^2\sigma_x^2}{\alpha_1^2\sigma_x^2+\alpha_2^2\sigma_z^2+\sigma_\delta^2}$. Noting that $\alpha_2^2\sigma_z^2 + \sigma_\delta^2 = \frac{\sigma_\delta^2}{1-\rho_{ez}^2}$ and solving for α_1 yields :

$$\alpha_1 = \sqrt{\frac{\sigma_\delta^2 \rho_{xy}^2}{\sigma_x^2(1-\rho_{xy}^2)(1-\rho_{ez}^2)}}. \quad (20)$$

The value of ρ_{xy}^2 is also set to a pre-determined constant.

The following quantities are set to pre-determined constants: μ_x, σ_x^2, μ_z, σ_z^2, σ_δ^2, ρ_{ez} and ρ_{xy}. Then, α_1 and α_2 are determined as in (20) and (19). In our simulation study, we set μ_x=μ_z=4, σ_x^2=σ_z^2=8, σ_δ^2=36, ρ_{ez}=0.8, and for ρ_{xy}, we consider two cases, ρ_{xy}=0.8 and ρ_{xy}=0.2. The non-probability sample size is 300, and the probability sample size is n = 20 or n = 50. For the Bayesian estimation of β, the prior mean of β is estimated based on the non-probability sample, while the prior variance is the proposed bias-variance specification given in (18).

4.2 Results

We compare the Bayesian predictors of the population mean $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$ and $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$, described in Section 2, with the Horvitz-Thompson (HT) estimator $\hat{\theta}_w$ (with $w_i = \pi_i^{-1}$) by computing the Monte Carlo Absolute Relative Bias (ARB) and Relative Root Mean Square Error (RRMSE) of these estimators. The ARB is defined as

$$ARB = \left|\frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\theta}^{(r)}-\theta)}{\theta}\right|,$$

where $\hat{\theta}^{(r)}$ is $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$, $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$ or $\hat{\theta}_w$ based on *r*-th simulation, and *R* = 5000. The RRMSE is defined as

$$RRMSE = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)}-\theta)^2}}{\theta}.$$

Recall that $\widehat{\boldsymbol{\beta}}^*$ is obtained under the assumption that the probability sampling design is ignorable, whereas $\widehat{\boldsymbol{\beta}}_{NI}^*$ is obtained assuming a non-ignorable probability sampling design. Table 3.2.1 presents the comparison of the HT estimator and Bayesian predictors, in terms of ARB and RRMSE, when ρ_{xy}=0.8, which indicates a strong association between x_i and y_i.

Table 4.2.1
Comparison of ARB and RRMSE, $\boldsymbol{\rho_{xy} = 0.8}$

Estimator	Sample size n = 20		Sample size n = 50	
	ARB	RRMSE	ARB	RRMSE
$\hat{\theta}_w$	0.88%	21.7%	0.91%	13.9%
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$	5.3%	19.5%	5.7%	12.7%
$\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$	0.95%	16.8%	0.98%	10.5%

From Table 4.2.1, it is clear that the bias of $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$ is negligible as its Monte Carlo ARB is just slightly larger than that of the unbiased HT estimator. The ARBs of $\hat{\theta}_w$ and $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$ are smaller than 1% for both sample sizes, whereas $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}^*)$ has a moderate bias slightly larger than 5%. The HT estimator has the largest RRMSE, and $\hat{\theta}_a(\widehat{\boldsymbol{\beta}}_{NI}^*)$ has the smallest RRMSE. Thus, it is clear that properly accounting for the probability sampling design in the estimation

of $\boldsymbol{\beta}$ leads to a decreased bias and a smaller RRMSE. The results in Table 4.2.1 show that the proposed Bayesian predictor performs very well when the probability sampling design is not ignorable.

Table 4.2.2
Comparison of ARB and RRMSE, $\rho_{yx} = 0.2$

Estimator	Sample size n = 20		Sample size n = 50	
	ARB	RRMSE	ARB	RRMSE
$\hat{\theta}_w$	0.98%	15.1%	0.93%	9.72%
$\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*)$	15.6%	26.5%	19.5%	22.9%
$\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*_{NI})$	1.86%	20.5%	1.48%	12.6%

Table 4.2.2 presents the comparison of the HT estimator and Bayesian predictors when ρ_{xy}=0.2, which indicates a weak association between x_i and y_i. The HT estimator performs the best in that scenario with the smallest ARB and the smallest RRMSE. The predictor $\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*)$ performs the worst with a very large bias and RRMSE, whereas $\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*_{NI})$ has a small ARB, below 2% but slightly larger than the ARB of the HT estimator, and a smaller RRMSE than $\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*)$. Thus, properly accounting for the probability sampling design in the estimation of $\boldsymbol{\beta}$ can substantially reduce the bias, even when the association between x_i and y_i is weak. However, the Bayesian predictor $\hat{\theta}_a(\hat{\boldsymbol{\beta}}^*_{NI})$ is not necessarily more efficient than the HT estimator in that situation.

5. Conclusion

We have proposed a predictor of a population mean using a Bayesian linear model that allows us to combine data from a probability and non-probability sample. Under a suitable informative prior distribution for the model parameters, we have shown the benefit of using a non-probability sample to improve estimates from a probability sample, particularly when the association between auxiliary variables and the variable of interest is not weak. However, if the association is perfect, no efficiency gains can be achieved through the use of a non-probability sample as our predictor reduces to the standard survey-weighted estimator. We have also shown the importance of accounting for the probability sampling design when it is non-ignorable. Our proposed predictor of the population mean performed well in our simulation study.

In future research, we plan to investigate the estimation of the mean square error and compare our Bayesian predictor with the one proposed by Nandram and Rao (2021). Another topic of interest, especially in social statistical programs, is the extension to the estimation of a proportion (i.e., the case of a binary variable of interest) using a Bayesian logistic model.

References

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

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data*, Chambers, R.L., and Skinner, C.J. (Eds.), Wiley, New York.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Deville, J.-C., and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

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

Elliott, M., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.
Fuller, W.A. (2009). *Sampling Statistics*, Wiley, New York.

Hidiroglou, M., and You, Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodology*, 42, 41-46.

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

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83, 242-272.

Rivers, D. (2007). Sampling from web surveys. In *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

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

Valliant (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Wiśniowski, 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.

Wang, Z., Kim, J.K., and Yang, S. (2018). Approximate Bayesian inference under informative sampling. *Biometrika*, 105, 91-102.

Yang, S., Kim, J.K., and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 29-58.

You, Y., Rao, J.N.K., and Kovacevic, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. In *Proceedings of 2003 Statistics Canada's XXth International Methodology Symposium*, Statistics Canada.