The Survey Statistician, 2021, Vol. 83, 11.-22.

Ask the Experts

Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies?

Jean-Francois Beaumont [1] **and J. N. K. Rao** [2]

[1] Statistics Canada, jean-francois.beaumont@canada.ca

[2] Carleton University, jrao@math.carleton.ca

Abstract

There is a growing interest in National Statistical Offices to produce Official Statistics using non-probability sample data, such as big data or data from a volunteer web survey, either alone or in combination with probability sample data. The main motivation for using non-probability samples is their low cost and respondent burden, and quick turnaround since they allow for producing estimates shortly after the information needs have been identified. However, non-probability samples are not a panacea. They are well known to produce estimates that may be fraught with significant selection bias. We first discuss this important limitation, along with an illustration, and then describe some remedies through inverse probability weighting or mass imputation. We also discuss how to integrate data from probability and non-probability samples through the Fay-Herriot model used in Small Area Estimation. We conclude with a few remarks on some challenges that statisticians are facing when implementing data integration methods.

Key words: Big data, Inverse probability weighting, Mass imputation, Selection bias, Small area estimation.

1 Introduction

Large-scale sample surveys, based on properly designed probability samples, have long been used to obtain reliable estimates of population totals, means and other descriptive parameters. Repeated sampling (or design-based approach) is widely employed for this purpose ever since the landmark paper by Neyman (1934) laid the theoretical foundations of the design-based approach to inference. Attractive features of this approach include design-consistent estimation of the parameters and associated mean squared errors, and normal theory confidence intervals on the parameters, that are valid, at least for large enough samples, “whatever the unknown properties of the finite population” (Neyman, 1934).

Efficient sampling designs minimizing cost subject to specified precision of estimators and “optimal” estimation of parameters taking advantage of supplementary information, such as censuses and administrative data, were proposed. Methods for combining two or more independent probability samples were also developed to increase the efficiency of estimators for a given cost. Many

important large-scale surveys conducted by official statistical agencies, such as labor force, business, and agricultural surveys, continue to follow the traditional approach.

In the early days of probability sampling evolution, surveys were generally much simpler than they are nowadays, and data were largely collected through personal interviews or through mail questionnaires followed by personal interviews of nonrespondents. Physical measurements were also used. Data collection issues received much attention in recent years to control costs and maintain response rates using new modes of data collection adapted to technological changes. Despite those efforts to collect designed data under probability sampling, response rates are decreasing, and costs and response burden are increasing. On the other hand, largely due to technological innovations, large amounts of inexpensive data, called big data or organic data, and data from nonprobability samples (especially self-selection web surveys) are now accessible. Big data include transaction data, social media data, scrape data from websites, sensor data and satellite images. Such data have the potential of providing estimates in near real time, unlike traditional designed data collected from probability samples.

Statistical agencies publishing official statistics are now undertaking modernization initiatives by finding new ways to integrate data from a variety of sources and produce "reliable" official statistics quickly. However, naïve use of data from nonprobability samples or big data can lead to serious selection bias problems. Without using suitable adjustments to account for selection bias, it can lead to the big data paradox: the bigger the data, the surer we fool ourselves, as demonstrated in Section 3 (Meng, 2018) and in the illustration in Section 4. We discuss some remedies in Section 5 to reduce the pitfalls arising from making inferences from nonprobability samples or big data. We show that the methods designed for making inferences from probability samples can be adapted for nonprobability samples. In Section 6, we discuss how small area estimation techniques can be used to integrate data from probability and non-probability samples. We provide a few concluding remarks in the last section on some of the practical challenges that require further thinking.

2 Probability sampling

2.1 Design-based approach

A distinctive feature of a probability sample A is that it ensures every unit i in the finite population U has a known nonzero inclusion probability π_i, leading to design weights $d_i = \pi_i^{-1}$ and a basic design-unbiased expansion estimator $\hat{Y} = \sum_{i \in A} d_i y_i$ of the finite population total $Y = \sum_{i \in U} y_i$ of a variable of interest y. Extensive research was conducted to improve the efficiency of the expansion estimator through the use of auxiliary variables $\mathbf{x}$ with known population totals $\mathbf{X}$. This is accomplished at the design stage through probability proportional to size sampling and stratification or at the estimation stage through ratio or regression estimation or both. The resulting improved estimators are not necessarily design-unbiased, but they are design-consistent in large samples. A well-known example is the ratio estimator $\hat{Y}_r = (\hat{Y} / \hat{X})X$ extensively used in practice, where $\hat{X}$ is the expansion estimator of the known total X. The ratio estimator may be expressed as a calibration estimator $\hat{Y}_r = \sum_{i \in A} w_i y_i$ with calibration weights $w_i = (X / \hat{X})d_i$ that ensure the calibration property $\sum_{i \in A} w_i x_i = X$. This property ensures that the ratio estimator agrees with the known total X when y_i is replaced by x_i.

Extensive research has been undertaken to extend calibration estimation to a vector of auxiliary variables $\mathbf{x}$ with known totals $\mathbf{X}$. A simple way of constructing calibration weights is to minimize a chi-squared distance measure between d_i and w_i for $i \in A$ with respect to w_i subject to calibration constraints $\sum_{i \in A} w_i \mathbf{x}_i = \mathbf{X}$. Post-stratification is an important special case which occurs when all

elements of $\mathbf{x}_i$ but one is equal to 0. Calibration estimation has attracted the attention of users due to its model-free property and its ability to produce a common set of weights not depending on the study variable. Van den Brakel and Bethlehem (2008) note that the calibration weights are "very attractive to produce timely official statistics in a regular production environment". The calibration approach has the potential to adjust for selection bias of non-probability samples, as noted in Section 5.

Design-consistent variance estimation under probability sampling applicable to general descriptive parameters, leading to normal theory confidence intervals on the parameters, also received a lot of attention. Methods proposed include Taylor linearization and replication methods, such as the jackknife and the bootstrap, taking account of the design features.

2.2 Unit nonresponse

Bias due to unit nonresponse in a probability sample received considerable attention, and promising remedies were proposed under a random response model. Under this model, the units in the population are assumed to respond independently if selected in the sample with unknown probabilities q_i, $i \in U$. Suppose we select a simple random sample (SRS) of size n and use the sample mean of respondent values as the estimator of the population mean $\bar{Y}$. Then, under the above design-model set up, the bias of the naïve estimator is approximately equal to $B_q = (R_{qy} S_q S_y)/\bar{Q}$, where R_{qy} is the finite population correlation between the study variable and the response probability, S_q and S_y denote the standard deviations of the response probabilities q_i and values y_i of the study variable, and $\bar{Q}$ is the population mean of the response probabilities (Bethlehem, 2020). It follows from the above bias expression that the bias increases with R_{qy}, S_q and decreasing response rate, $\bar{Q}$. The bias disappears if the response probabilities are identical ($S_q = 0$) or nonresponse is not selective ($R_{qy} = 0$). The latter case, called missing completely at random (MCAR), seldom holds in practice.

Success of any adjustment for nonresponse bias depends on the availability of auxiliary variables $\mathbf{z}$ for all the sampled units that are closely related to the study variable y. In, for instance, Scandinavian countries, population registers are often used to extract $\mathbf{z}$ for all the units in the sample. Missing at random (MAR) assumption plays a dominant role in estimating the response probabilities (propensities) q_i. Under MAR, $q_i = q(\mathbf{z}_i, \boldsymbol{\alpha})$ for a specified function $q(.)$ depending only on the observed $\mathbf{z}_i$ and a parameter $\boldsymbol{\alpha}$. A bias-adjusted expansion estimator of the total Y under MAR is constructed as $\hat{Y}_q = \sum_{i \in A(r)} (d_i / \hat{q}_i) y_i$, where $A(r)$ is the sample of respondents and $\hat{q}_i = q(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ is the estimated response probability. The parameter $\boldsymbol{\alpha}$ is estimated from the sample data $\{(\delta_i, \mathbf{z}_i), i \in A\}$, where δ_i is the response indicator. Logistic regression models, $\log\{q_i/(1-q_i)\} = \mathbf{z}_i' \boldsymbol{\alpha}$, are commonly used for this purpose. The bias-adjusted estimator $\hat{Y}_q$ is design-model consistent, provided the response propensity model is correctly specified. The estimator $\hat{Y}_q$ becomes unstable when some of the estimated response propensities become small. To avoid this problem, weighting classes can be formed, based on quantiles of the estimated response probabilities, and then use a post-stratified estimator by assuming uniform response probabilities within classes.

3 Non-probability samples: selection bias

Consider a non-probability sample B of known size N_B with data $\{(i, y_i), i \in B\}$ and a participation indicator δ_i taking the value 1 if the population unit i belongs to B and 0 otherwise. In the absence of auxiliary information $\mathbf{z}$, the estimator of $\bar{Y}$ is taken as the sample mean $\bar{y}_B = N_B^{-1} \sum_{i \in U} \delta_i y_i$. T0he estimation error $\bar{y}_B - \bar{Y}$ may be expressed as the product of three terms: (1) $corr(\delta, y) = R_{\delta,y}$, called data quality, (2) square root of $(1-f_B)/f_B$ with $f_B = N_B / N$, called data quantity and (3) square root of the population variance S_y^2, called problem difficulty (Meng, 2018). The data quality term plays the key role in determining the estimation error and it is approximately zero on the average under SRS. Note that we have not used a random participation mechanism that assumes the participation indicators δ_i are random and independent with non-zero participation probabilities $q_i = P(\delta_i = 1)$; for simplicity, we use the same notation for response propensities and participation probabilities. Under the random participation model, the bias of $\bar{y}_B$ given by $E_\delta(\bar{y}_B - \bar{Y})$ has the same expression as in the case of nonresponse. The subscript δ indicates that the expectation is taken with respect to the participation model. However, Bethlehem (2020) notes that in practical situations the two values can be substantially different. He gives an example from Netherlands where $\bar{Q}$ is around 60% for probability surveys compared to 1.5% for self-selection web surveys, even though 170,000 people completed the questionnaire in the web survey. This example suggest that self-selection surveys can suffer from the risk of a much larger bias.

The model mean squared error $MSE_\delta(\bar{y}_B)$ conditional on the sample size N_B may be expressed as the product of three terms: (1) $E_\delta(R_{\delta,y}^2)$, called the data defect index, (2) drop-out odds $(1-f_B)/f_B$ and (3) degree of uncertainty S_y^2. Note that MSE is affected by the sampling fraction f_B and not the sample size N_B. As a result, a relatively small simple random sample of size n can achieve the same MSE. For example, suppose N_B is five million and N is ten million leading to $f_B = 1/2$, and the average correlation $E_\delta(R_{\delta,y})$ is as small as 0.05. Then the "effective" sample size of the big data is less than 400. Moreover, the confidence interval, treating the big data as a simple random sample, has a small chance of covering the true mean $\bar{Y}$ because it is centered at a wrong value due to the induced bias. We know this phenomenon under probability sampling when the ratio of bias to standard error is large. For a design-consistent estimator, such as a ratio estimator, the bias ratio goes to zero as the sample size increases.

For simplicity, we assumed the absence of measurement errors in the nonprobability sample B. This is often not the case with found data from online sources, such as Facebook, where people may actively lie, and the expected bias due to measurement errors could be large. Biemer (2019) extended Meng's model to show that the bias due to measurement errors could significantly inflate the total MSE.

Unlike in the case of nonresponse in a probability sample A, auxiliary variables $\mathbf{z}$ attached to the units not participating in the nonprobability sample B are seldom available. As a result, it is not possible to estimate participation probabilities from sample B alone and make bias adjustments. In Section 5 we study some methods of estimating participation probabilities by supplementing the data $\{(i, \mathbf{z}_i), i \in B\}$ with the data $\{(i, \mathbf{z}_i), i \in A\}$ obtained from an independent probability sample A observing $\mathbf{z}$ and possibly different study variables. This set up has received a lot of attention in the recent literature, but we focus on a pseudo-likelihood method proposed by Chen, Li and Wu (2019) and a mass imputation method of Rivers (2007).

4 An illustration using real data

After the beginning of the COVID-19 lockdown in March 2020, Statistics Canada conducted a series of crowdsourcing experiments to respond to urgent information needs about the life of the Canadian population. A crowdsourcing sample can be defined as any non-probability sample of volunteers, who typically provide information through an online application. Statistics Canada's crowdsourcing data were collected by posting questionnaires on its website on different topics at regular intervals. The main advantages of crowdsourcing are its low cost and quick turnaround since estimates can be released within a couple of weeks after the information needs have been determined. This timeliness was deeply needed in a pandemic time. The first crowdsourcing experiment was viewed as a success considering that around 240,000 persons participated. However, the number of participants in the subsequent crowdsourcing experiments was smaller, but often reached over 30,000 participants. As pointed out in Section 3, ignoring possible measurement errors, the main drawback of non-probability surveys of volunteers is the selection bias, also called participation bias. To account for this bias, it was decided to apply post-stratification weighting with post-strata defined by the cross-classification of province, age group and sex. Renaud and Beaumont (2020) provide greater detail on crowdsourcing experiments conducted by Statistics Canada.

In parallel, Statistics Canada also started a shorter series of probability web panel surveys: the Canadian Perspective Survey Series (CPSS). The CPSS sample is obtained from past rotation groups of the Labour Force Survey (LFS), which is the most important social survey conducted by Statistics Canada except for the Census. The CPSS initial probability sample is relatively large with over 30 000 selected persons but the overall recruitment/response rate is usually quite low at around 15%, and the resulting number of respondents is just slightly over 4,000. Greater detail on the CPSS can be found in Baribeau (2020).

In June 2020, some participants from previous crowdsourcing experiments were randomly chosen and sent the same questionnaire as CPSS respondents. This allowed for a comparison of estimates from both the CPSS probability sample and this crowdsourcing non-probability sample. We provide some results for the variable *education* as this variable is also available in the LFS, which has generally a response rate around 80%, and is treated as our gold standard. The June 2020 CPSS contained 4,209 respondents whereas the corresponding crowdsourcing sample had 31,505 participants, and the LFS had 87,970 respondents. We computed LFS, CPSS and crowdsourcing estimates of proportions in different education categories (see Table 1 for the description of categories) for Canada (see Figure 1) and for a province of Canada (see Figure 2). Normal 95% confidence intervals were also computed for the LFS and CPSS. We considered two versions of crowdsourcing estimates: unadjusted and post-stratified. The unadjusted crowdsourcing estimates were obtained using an estimation weight equal to 1 for every participant, and the post-stratified crowdsourcing estimates were obtained by using a post-stratification weight with post-strata defined by the cross-classification of province, age group and sex. Three main conclusions can be drawn from Figure 1:

i) The crowdsourcing sample seems to significantly over-represent those with a university degree.
ii) Post-stratification by province, age group and sex have little impact on the estimates, and thereby on the selection bias.
iii) The CPSS estimates are closer to LFS estimates although with larger confidence intervals due to the smaller sample size.

The same conclusions can be drawn with the provincial estimates in Figure 2, the main difference being that confidence intervals are wider. For the province considered, the number of respondents in the LFS and CPSS are 4,734 and 231, respectively, and the number of crowdsourcing participants is 1,716.

Table 1: *Categories of education*

Education categories	Description
0	Grade 8 or lower
1	Grade 9 - 10
2	Grade 11 - 13, non graduate
3	Grade 11 - 13, graduate
4	Some post-secondary education
5	Trades certificate or diploma
6	Community college, CEGEP, etc.
7	University certificate below Bachelor's
8	Bachelor's degree
9	Above Bachelor's degree

A caveat must be mentioned about conclusion (iii): nonresponse weighting in the CPSS used the study variable, education, as one of the auxiliary variables. This may explain the small difference between the CPSS and LFS estimates for that variable, especially at Canada level. Ideally, nonresponse weighting would be done again by omitting the auxiliary variable education. This would better allow us to appreciate the accuracy of the CPSS probability sample. Unfortunately, this could not be done before the publication of this paper.

Conclusion (ii) indicates that province, age and sex are insufficient for significantly reducing the selection bias. More powerful auxiliary variables are needed for this purpose with possibly more sophisticated methods such as propensity score weighting (Chen, Li and Wu, 2019) or sample matching (Rivers, 2007) discussed in Section 5. In practice, the variable education itself can be used to reduce the selection bias for other study variables. Preliminary results suggest that it is a powerful auxiliary variable although insufficient for entirely removing the selection bias.

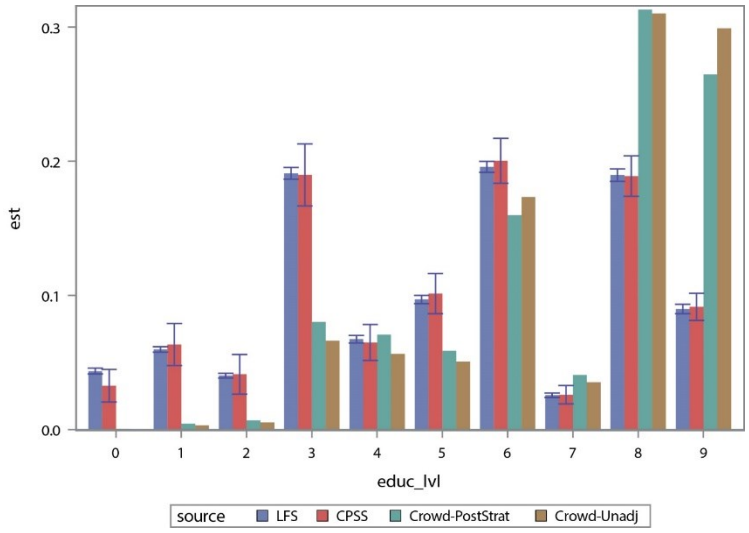

Figure 1: *Estimates of proportions in different education categories for Canada from different sources: a large probability survey (LFS), a small probability survey (CPSS), a non-probability survey (Crowd-Unadj) and a non-probability survey with post-stratification weighting by province, age group and sex (Crowd-PostStrat)*

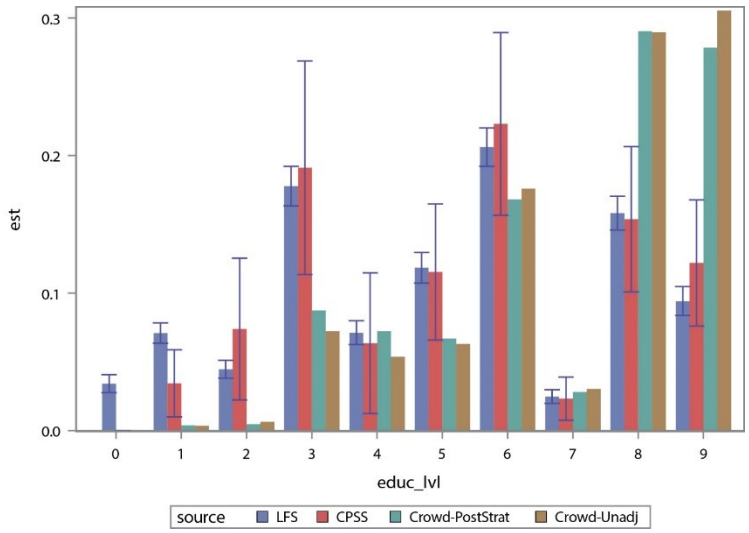

Figure 2: *Estimates of proportions in different education categories for a Canadian province from different sources: a large probability survey (LFS), a small probability survey (CPSS), a non-probability survey (Crowd-Unadj) and a non-probability survey with post-stratification weighting by province, age group and sex (Crowd-PostStrat)*

5 Methods for adjusting selection bias

Our data for estimating participation probabilities consist of the common auxiliary variables $\mathbf{z}$ observed in both samples A and B. We assume a model for the participation probabilities $q_i = q(\mathbf{z}_i, \boldsymbol{\alpha})$ for specified $q(.)$ under the MAR assumption. Chen, Li and Wu (2019) estimate the population log-likelihood $l(\boldsymbol{\alpha})$ based on the sample B by using the data $\{(i, \mathbf{z}_i), i \in A\}$ and associated design weights $d_i, i \in A$, leading to a pseudo-likelihood $\hat{l}(\boldsymbol{\alpha})$. For the commonly used logistic regression model $\log\{q_i/(1-q_i)\} = \mathbf{z}_i'\boldsymbol{\alpha}$, the corresponding pseudo-score equation reduces to

$$\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \sum_{i \in B} \mathbf{z}_i - \sum_{i \in A} d_i q(\mathbf{z}_i, \boldsymbol{\alpha}) \mathbf{z}_i = \mathbf{0}. \quad (1)$$

Equation (1) is solved using the Newton-Raphson iteration procedure with a starting value $\boldsymbol{\alpha}^{(0)} = 0$, leading to an estimator $\hat{\boldsymbol{\alpha}}$ and corresponding estimated probability $\hat{q}_i = q(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$. The resulting inverse probability weighted (IPW) estimator of the mean is $\bar{y}_{B,q} = \sum_{i \in B} \omega_i y_i / \sum_{i \in B} \omega_i$, where $\omega_i = \hat{q}_i^{-1}$. The corresponding ratio estimator of the total is $\hat{Y}_{B,q} = N\bar{y}_{B,q}$, but the population size N may not be known in practice unless the probability sample is drawn from a list frame. Conventional calibration estimation can also be used to estimate $\boldsymbol{\alpha}$ by solving

$$\sum_{i \in B} [q(\mathbf{z}_i, \boldsymbol{\alpha})]^{-1} \mathbf{z}_i = \sum_{i \in A} d_i \mathbf{z}_i. \quad (2)$$

A drawback of the IPW estimator is that it is sensitive to misspecification of the participation probability model, especially when some of the $\hat{q}_i, i \in A,$ are small. One way to achieve some robustness is to use a double robust (DR) estimator that requires modeling each study variable y. Suppose we assume a working population model, $E_m(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}), i \in U$, where E_m denotes model expectation. The choice of auxiliary variables $\mathbf{x}_i$ may include the common variables $\mathbf{z}_i$ and other available auxiliary variables in the sample B that are predictive of the study variable y_i. We further assume the population model holds for the sample B. Using the data from the sample B we

obtain an estimator $\hat{\boldsymbol{\beta}}$ which is consistent for $\boldsymbol{\beta}$ under the assumed model. Then a DR estimator of the total is given by

$$\hat{Y}_{DR.q} = \sum_{i\in B} \omega_i (y_i - \hat{m}_i) + \sum_{i\in A} d_i \hat{m}_i \ , \tag{3}$$

where $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ denote the predicted values under the model. The estimator (3) is DR in the sense it is consistent if either the model for the participation probabilities or the model for the study variable is correctly specified. The DR estimator of the mean is given by $\bar{y}_{DR,q} = \hat{Y}_{DR,q} / \hat{N}_q$, where $\hat{N}_q = \sum_{i\in B} \omega_i$ is the IPW estimator of the population size N. DR estimators have been used in the context of nonresponse in a probability sample (Kim and Haziza, 2014).

Chen, Li and Wu (2019) provided asymptotically valid variance estimators of the DR estimator and the IPW estimator under the assumed models. They also conducted a limited simulation study on IPW, DR and some alternative estimators of the mean. As expected, the DR estimator performed well in terms of relative bias (RB) and MSE when either the assumed model for participation probabilities or the model on the study variable is correctly specified. On the other hand, the IPW estimator performs poorly when the model on the participation probabilities is incorrectly specified. Chen, Li and Wu (2019) did not study the case when both models are incorrectly specified. It would be interesting to develop multiple robust (MR) estimators, along the lines of Chen and Haziza (2017) in the context of unit nonresponse, by specifying multiple models on the study variable and multiple models for the participation probability. Under this set up, an estimator is MR if it performs well when at least one of the candidate models is correctly specified. Simulation results in the context of nonresponse indicated that MR estimators tend to perform well even when the candidate models are all incorrectly specified. It would be interesting to study MR estimation in the context of nonprobability samples.

A popular method is based on non-parametric mass imputation that avoids the specification of the mean function $E_m(y_i \mid \mathbf{z}_i)$ based on common auxiliary variables $\mathbf{z}_i$ (Rivers, 2007). In this method, for each unit $i \in A$, a nearest neighbor (NN) to the associated $\mathbf{z}_i$ is found from the donor set $\{(i, \mathbf{z}_i), i \in B\}$, say $(l, \mathbf{z}_l)$, and the corresponding y_l is used as the imputed value $y_i^* (= y_l)$. Euclidean distance is commonly used to find the NN. The mass imputed estimator of the total is then given by

$$\hat{Y}_{d,I} = \sum_{i\in A} d_i y_i^* . \tag{4}$$

The estimator (4) is based on real donor values. However, it is not exactly model-design unbiased unless $E_m(y_i^* \mid \mathbf{z}_i) = E_m(y_i \mid \mathbf{z}_i)$ for $i \in A$. The NN estimator of the mean is given by $\bar{y}_{d,I} = \hat{Y}_{d,I} / \hat{N}_d$, where $\hat{N}_d = \sum_{i\in A} d_i$ is the design-unbiased estimator of the total. Under certain regularity conditions, including smoothness of the mean function and MAR assumption, Yang and Kim (2020) showed that the NN estimator (4) behaves asymptotically similar to the design-unbiased estimator with the imputed values in (4) replaced by the unobserved true values y_i of the units $i \in A$.

An advantage of the estimator (4) is that it can be readily implemented when non-probability samples are observed frequently over time leading to time-varying values of the study variable, provided the common auxiliary variables do not vary over time. We simply find the NN for the units in the probability sample from the current donor set and calculate (4).

6 Small Area Estimation

Data integration methods in Section 5 can be applied in the scenario where the study variable y is observed in the non-probability sample B, but completely missing in the probability sample A. The

reduction of the selection bias relies on the availability of a vector of auxiliary variables $\mathbf{z}$ observed in both samples. In this section, we study a different scenario. We consider the case where the study variable is observed in the probability sample A, and the non-probability sample contains a vector of auxiliary variables associated with y, such as a proxy for y, or perhaps even y itself. Design-unbiased estimators of finite population parameters can be obtained by ignoring all non-probability sample data. Therefore, the selection bias is not an issue in this context. The objective is to take advantage of the information contained in the non-probability sample to improve the efficiency of design-unbiased estimators through model assumptions that link data from both samples. This is the scope of Small Area Estimation (SAE) techniques (Rao and Molina, 2015). We focus below on the area level model of Fay and Herriot (1979). Unit level models are also used in SAE but require the auxiliary variables to be observed in the probability sample, possibly through record linkage, and this may not be feasible.

Suppose that we are interested in estimating J population parameters, θ_j, $j=1,...,J$, where the subscript j refers to J disjoint population domains of interest. For instance, θ_j could be the unemployment rate in a certain area j. Let us denote by $\hat{\theta}_j$, a design-unbiased estimator of θ_j obtained from the probability sample A. From this design-unbiasedness property, we can write the so-called sampling model as

$$\hat{\theta}_j = \theta_j + e_j,$$

where e_j is the sampling error such that $E_p\left(e_j\right)=0$, $\mathrm{var}_p\left(e_j\right)=\psi_i$ and ψ_i is the design variance of $\hat{\theta}_j$. The subscript p indicates that the expectation and variance are taken with respect to the probability sampling design. In practice, $\hat{\theta}_j$ is rarely exactly design-unbiased but is assumed to be at least approximately design-unbiased. The sampling errors are also assumed to be normally distributed and mutually independent even when the design strata do not coincide with the domains.

The above sampling model is complemented with a linking model that relates the population parameter θ_j to a vector of auxiliary variables $\mathbf{z}_{B,j}$. This vector of auxiliary variables may come from many different sources; in practice, they are often administrative sources. Here, we focus on the use of a non-probability sample as the source of auxiliary information. For instance, a proxy for θ_j could be computed from the non-probability sample B, if a variable similar to y is observed, and used as one component of $\mathbf{z}_{B,j}$. The following linking model is often used:

$$\theta_j = \mathbf{z}'_{B,j}\boldsymbol{\beta} + b_j v_j,$$

where b_j are constant to account for possible heteroscedasticity, v_j are mutually independent errors that follow the normal distribution with $E_m\left(v_j\right)=0$ and $\mathrm{var}_m\left(v_j\right)=\sigma_v^2$, and $\boldsymbol{\beta}$ and σ_v^2 are unknown model parameters. The subscript m indicates that the expectation and variance are taken with respect to the model, treating the vectors $\mathbf{z}_{B,j}$ as fixed.

The Fay-Herriot model is obtained by combining the sampling and linking model as

$$\hat{\theta}_j = \mathbf{z}'_{B,j}\boldsymbol{\beta} + a_j,$$

where $a_j = b_j v_j + e_j$. It is straightforward to show that $E_{mp}\left(a_j\right)=0$ and $\mathrm{var}_{mp}\left(a_j\right)=b_j^2\sigma_v^2+\tilde{\psi}_j$, where $\tilde{\psi}_j = E_m\left(\psi_j\right)$ is called the smooth design variance of $\hat{\theta}_j$ (Hidiroglou, Beaumont and Yung, 2019). The Empirical Best (EB) predictor of θ_j under the Fay-Herriot model is

$$\hat{\theta}_j^{EB} = \hat{\gamma}_j \hat{\theta}_j + \left(1 - \hat{\gamma}_j\right) \mathbf{z}'_{B,j} \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_j = b_j^2 \hat{\sigma}_v^2 \big/ \left(b_j^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_j\right)$, and $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ and $\hat{\tilde{\psi}}_j$ are estimators of $\boldsymbol{\beta}$, σ_v^2 and $\tilde{\psi}_j$, respectively. The vector $\boldsymbol{\beta}$ is typically estimated by generalized least squares and a number of methods exist for the estimation of σ_v^2 (see Rao and Molina, 2015), the most common being restricted maximum likelihood. Hidiroglou, Beaumont and Yung (2019) use a log-linear smoothing model and a method of moments for the estimation of $\tilde{\psi}_j$.

Chatrchi, Gauvin and Ding (2020) recently applied the Fay-Herriot model to the estimation of trip spending by visitors to Canada for 242 domains defined by the cross-classification of region and the country of origin of visitors. The direct estimator $\hat{\theta}_j$ is obtained from the Visitor Travel Survey (VTS) and a proxy for θ_j is obtained from payment processors' data. These data are subject to coverage issues and are conceptually different from the y variable measured in the VTS. Yet, they were able to observe significant precision improvements by using the Fay-Herriot model, especially for the smallest domains. Their initial analyses showed the evidence of a nonlinear relationship between $\hat{\theta}_j$ and the proxy for θ_j. A piecewise linear model appears to have addressed this issue satisfactorily. Rao (2020) describes other SAE applications, where authors have used non-probability data and big data as a source of auxiliary information in a SAE model.

7 Concluding remarks

National Statistical Offices have been increasingly considering big data and volunteer web surveys in recent years as an alternative to conducting probability surveys. The main advantages of these non-probability samples over probability samples are their low cost and respondent burden, as well as their quick turnaround, i.e., the gap between the determination of information needs and the release of estimates is quite small for non-probability samples.

It is well known that the use of a non-probability sample alone may lead to biased estimates of finite population parameters due to measurement errors and selection bias (e.g., Rao, 2020; Beaumont, 2020). As a result, it does not seem advisable to use non-probability data to produce official statistics without complementing them with probability survey data. In Section 5, we discussed data integration methods that reduce the bias by combining non-probability and probability survey data. Although these methods are useful to achieve bias reductions, they rely strongly on the MAR assumption. Practically speaking, this assumption implies that the participation indicator is independent of the study variable y after conditioning on $\mathbf{z}$. A powerful set of auxiliary variables, associated with both the participation indicator and the study variable, is key to make this assumption plausible. Before probability and non-probability surveys are conducted, it is thus useful to give some thoughts on the inclusion of proper variables to be collected in both surveys for the purpose of making the MAR assumption more realistic and dampening the selection bias. Of course, it is not always possible to do so when the non-probability sample is not managed by the same agency as the probability sample. In some cases, there may be many auxiliary variables available for adjusting estimates from the non-probability sample. Variable selection techniques can be useful for selecting relevant auxiliary variables but need to be adapted to this data integration context. This problem is currently being investigated at Statistics Canada.

Quality indicators for estimates based on integrated data is a topic that requires further research. The model variance of those estimators can be estimated, often using standard techniques, but the resulting variance estimates fail to account for the selection bias, which may not be negligible. Sensitivity analysis may provide a useful complement to variance estimates. It allows for assessing the effect of omitting a relevant auxiliary variable on the conclusions drawn from the available data.

Ding and VanderWeele (2016) is a recent paper on the topic that is currently being investigated at Statistics Canada.

In some cases, depending on the users' objectives, evidence will suggest that estimates from integrated data are not appropriate for the intended uses. Still, it might be desired to take advantage as much as possible of the available non-probability sample data. Small Area Estimation is one possible avenue that combines probability and non-probability sample data through models. Although the Fay-Herriot model discussed in Section 6 relies on the validity of model assumptions, the resulting inferences are not as dependent on the model as inferences based on methods discussed in Section 5. First, diagnostics can be used to assess the model adequacy (e.g., Hidiroglou, Beaumont and Yung, 2019). Also, when the probability sample size is large in a domain, the small area estimates are usually quite close to the reliable direct estimates. However, SAE is only feasible when the variable of interest is observed in the probability sample. Another alternative to make use of non-probability sample data is to use dual frame weighting methods (e.g., Kim and Tam, 2020; Rao, 2020; Beaumont, 2020). The advantage of this approach is that it remains design-based, thereby not dependent on any model assumptions.

We have not discussed other practical issues related to the use of big data and non-probability samples, such as privacy, access, and transparency. A report by the U. S. National Academies of Sciences, Engineering and Medicine (2017) extensively treats the privacy issue, in addition to methodology for integrating data from multiple sources.

Acknowledgements

We deeply thank Andrew Brennan and Joanne Charlebois from Statistics Canada for providing the two figures in section 4, and David Haziza for his careful reading of the paper.

References

Baribeau, B. (2020). Trial by COVID for Statistics Canada's web panel pilot. Internal document, Statistics Canada.

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

Bethlehem, J. (2020). Working with response probabilities. *Journal of Official Statistics,* **36**, 647-674.

Biemer, P. (2019). Can a survey sample of 6000 records produce more accurate estimates than an administrative data base of 100 million? (The answer may surprise you). *The Survey Statistician*, **80**, 11-15.

Chatrchi, G., Gauvin, H., Ding, A. (2020). Combining survey data with other data sources using small area estimation: a case study. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**, 439-453.

Chen, Y., Li, P., and Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (published online).

Ding, P., and VanderWeele, T.J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27, 368-377.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **85**, 398-409.

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, **45**, 1, 101-126. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, 375-394.

Kim, J.K., and Tam, S.M. (2020). Data integration by combining big data and survey data for finite population inference. Unpublished manuscript.

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. *Annals of Applied Statistics*, **12**, 685-726.

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection*: *Nest Steps.* The National Academies Press, Washington, DC. https://doi.org/10.17226/24893.

Neyman, J. (1934). On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, **97**, 558--625.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, published online April 2020.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Second Edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

Renaud, M., and Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: the Statistics Canada experience. Paper presented at the Advisory Committee on Statistical Methods, Statistics Canada, October 27, 2020.

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

Van den Brakel, J.A. and J. Bethlehem (2008), Model-based estimation in official statistics. CBS Discussion paper 08002. Statistics Netherlands, The Hague/Heerlen. https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-official-statistic.

Yang, S. and Kim, J. (2020). Statistical data integration in survey sampling: a review. *arXiv preprint arXiv:2001.03259v1.*