Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data

by Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu

Release date: June 25, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 77-106
Statistics Canada, Catalogue No. 12-001-X

Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data

Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu[1]

Abstract

Non-probability samples are being increasingly explored in National Statistical Offices as an alternative to probability samples. However, it is well known that the use of a non-probability sample alone may produce estimates with significant bias due to the unknown nature of the underlying selection mechanism. Bias reduction can be achieved by integrating data from the non-probability sample with data from a probability sample provided that both samples contain auxiliary variables in common. We focus on inverse probability weighting methods, which involve modelling the probability of participation in the non-probability sample. First, we consider the logistic model along with pseudo maximum likelihood estimation. We propose a variable selection procedure based on a modified Akaike Information Criterion (AIC) that properly accounts for the data structure and the probability sampling design. We also propose a simple rank-based method of forming homogeneous post-strata. Then, we extend the Classification and Regression Trees (CART) algorithm to this data integration scenario, while again properly accounting for the probability sampling design. A bootstrap variance estimator is proposed that reflects two sources of variability: the probability sampling design and the participation model. Our methods are illustrated using Statistics Canada's crowdsourcing and survey data.

Key Words: Akaike Information Criterion; Classification and Regression Trees; Logistic model; Participation probability; Statistical data integration; Variable selection.

1. Introduction

Non-probability samples are being increasingly explored at Statistics Canada and in other statistical agencies around the world. Indeed, Statistics Canada has recently conducted several non-probability surveys to evaluate the impacts of the COVID-19 pandemic on different aspects of life of the Canadian population. Data of these non-probability surveys were collected from visitors of Statistics Canada's website who responded voluntarily to an online survey questionnaire. The main motivation for considering this non-probability approach, called crowdsourcing at Statistics Canada, over probability surveys is the significant reduction in time and cost that can be achieved in the production of survey statistics. Another important advantage is the non-intrusive nature of crowdsourcing since participation is made on a voluntary basis. However, it is well known that the use of a non-probability sample alone, such as a crowdsourcing sample, may produce estimates with significant bias due to the unknown nature of the underlying selection (or participation) mechanism. To reduce this participation bias, data from a non-probability sample can be combined with data from a probability sample, ideally a large one. Estimation methods that combine data from probability and non-probability samples fall under the area of statistical data integration.

We consider the data integration scenario for which the variables of interest are available only in the non-probability sample. However, a vector of auxiliary variables is observed in both samples and used to

1. Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca, keven.bosa@statcan.gc.ca, andrew.brennan@statcan.gc.ca, joanne.charlebois@statcan.gc.ca and kenneth.chu@statcan.gc.ca.

reduce bias. A possible approach to inference under this scenario relies on a model for the variables of interest along with the assumption that the non-probability sample is not informative with respect to the model. The prediction approach for finite populations (e.g., Royall, 1970; Valliant, Dorfman and Royall, 2000) is one possible avenue for data integration. If a linear model between the variables of interest and the auxiliary variables holds, it can be implemented by weighting the non-probability sample through calibration on known population totals or totals estimated from the probability survey (e.g., Elliott and Valliant, 2017; Valliant, 2020). Another model-based method is statistical matching (see Yang, Kim and Hwang, 2021, for a recent reference). It consists of imputing the missing values of the variables of interest in the probability sample using non-probability sample data. The method is called sample matching (e.g., Rivers, 2007) when donor imputation is used to fill in the missing values. The prediction approach with estimated totals and statistical matching lead to identical estimators under linear models with error variance linearly related to auxiliary variables (Beaumont, 2020). Since both methods rely on a model for the variables of interest, they may become impractical when there are multiple variables of interest as a model needs to be determined and validated for each of them.

An alternative approach to inference relies on a model for the participation indicator rather than a model for the variables of interest. This approach is more appealing when there are multiple variables of interest as there is only one participation indicator, and thus only one model to choose and validate. Estimates are obtained by weighting each participant in the non-probability sample by the inverse of its estimated participation probability. This is often called inverse probability weighting or propensity score weighting in the literature. We focus on this approach. If the values of the auxiliary variables are observed for the entire population, the problem is basically identical to weighting for survey nonresponse, and nonresponse weighting methods can be applied directly to weight the non-probability sample.

In general, the auxiliary variables are observed only for the participants in the non-probability sample. Chen, Li and Wu (2020) proposed a simple and attractive method to address this issue. It requires the auxiliary variables to be also observed in a probability sample and assumes that the logistic function is used to model the participation probability. An alternative to Chen, Li and Wu (2020) consists of creating a pooled sample from the probability and nonprobability samples and modelling the participation indicator under the assumption that there is no overlap between the two samples (e.g., Lee, 2006; Valliant and Dever, 2011; and Ferri-Garcia and Rueda, 2018). Chen, Li and Wu (2020) noted that this pooling method leads to a biased estimator of the participation probability. However, Beaumont (2020) pointed out that it yields estimated participation probabilities approximately equivalent to those of Chen, Li and Wu (2020) when all the participation probabilities are small and the probability sample is properly weighted. Wang, Valliant and Li (2021) proposed an extension of the pooling method to account for a non-negligible overlap between the probability and non-probability samples. Elliott and Valliant (2017) proposed another inverse probability weighting method based on the pooled sample. It also assumes no overlap between both samples and requires the probability survey weights to be available in the non-probability sample. Recent reviews of statistical data integration methods are given in Beaumont (2020), Lohr (2021), Rao (2021), Valliant (2020), Wu (2022) and Yang and Kim (2020).

The choice of auxiliary variables is key for bias reduction. They should ideally be related to both the participation indicator and the variables of interest. Chen, Li and Wu (2020) supposed that the auxiliary variables were given. In practice, there may be a number of auxiliary variables available in both samples, often categorical, and it may not be obvious to determine the relevant ones along with proper interactions. Variable selection tools could be useful but need to be adapted to the data integration scenario considered in this paper. In particular, they need to account for the sampling design used to select the probability sample and for any adjustments to the design weights, such as nonresponse and calibration adjustments. We propose a stepwise selection procedure that achieves this goal. It is based on a modification of the Akaike Information Criterion (AIC) similar to the one Lumley and Scott (2015) developed for the estimation of model parameters from probability survey data. The Least Absolute Shrinkage and Selection Operator (LASSO) is an alternative that is considered by Bahamyirou and Schnitzer (2021). This technique usually involves cross-validation for the determination of the penalty parameter. The development of cross-validation methods that handle a combination of a probability and non-probability sample, and that properly account for the probability sampling design, requires further research.

The logistic model may sometimes produce a few estimated probabilities that are very small leading to very large weights and potentially unstable estimates. A common solution to this problem in the context of survey nonresponse is to create homogeneous groups and weight each respondent (participant) in a given group by the inverse of the estimated response (participation) rate in the group. The resulting weights possess a calibration property (see Section 3.3), which tends to limit the magnitude of the largest weights. The creation of homogeneous groups also provides some robustness to model misspecifications, as illustrated by Haziza and Lesage (2016) in the context of survey nonresponse.

A possible avenue to the creation of homogeneous groups is to adapt the Classification and Regression Trees (CART) algorithm, developed by Breiman, Friedman, Olshen and Stone (1984), to the data integration scenario studied in this paper. A nice advantage of tree-based methods is that auxiliary variables and their interactions are chosen automatically. Chu and Beaumont (2019) developed an algorithm for growing a tree that accounts for the survey weights. They called the algorithm "nppCART" because it integrates data from both a non-probability and probability sample. Pruning is an important aspect of CART that is used to avoid overfitting and to improve the efficiency of the resulting estimates. Pruning is often based on cross-validation techniques but, as pointed out above, these techniques have yet to be extended to the data integration scenario studied in this paper. Instead, we consider a modification of the AIC, similar to Lumley and Scott (2015), that properly accounts for the probability sampling design and any design weight adjustments, and use it to develop a pruning procedure.

In Section 2, we introduce the data integration problem along with some notation. The estimation of participation probabilities is discussed in Sections 3 and 4. In Section 3, we consider more specifically the logistic model and describe our proposed variable selection procedure as well as a simple rank-based method, called the Frank method, for the creation of homogeneous groups. In Section 4, we describe nppCART and our proposed pruning procedure. Bootstrap estimation of the variance of our estimators is discussed in Section 5. An empirical evaluation of our methods using real data is shown in Section 6. The last section contains some concluding remarks.

2. Data integration scenario

Let us consider the estimation of the population total $\theta = \sum_{k \in U} y_k$, where U is the set of population units and y_k is the value of a variable of interest y for population unit k. We assume that y_k is observed without error in a non-probability sample $s_{\text{NP}} \subset U$. Along with y_k, a vector of auxiliary variables $\mathbf{x}_k$ is also observed for each unit $k \in s_{\text{NP}}$. The indicator of participation in the non-probability sample is denoted by δ_k, i.e., $\delta_k = 1$, if $k \in s_{\text{NP}}$, and $\delta_k = 0$, otherwise. A probability sample s_P, drawn using some probability sampling design, is also available. The auxiliary variables $\mathbf{x}_k$ are observed for each unit $k \in s_P$, but the variable of interest y_k and the participation indicator δ_k are missing in the probability sample.

The objective is to estimate θ under the above data integration scenario, i.e., using the y values observed in the non-probability sample along with the **x** values observed in both samples. Inverse probability weighting involves modelling the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$, which is assumed to be strictly greater than 0. The estimator of θ under this approach is $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, where $\hat{w}_k^{\text{NP}} = \hat{p}_k^{-1}$ is the non-probability survey weight, also called the pseudo survey weight, of participant k, and $\hat{p}_k$ is a consistent estimator of p_k. A critical assumption for the validity of this approach is that the participation mechanism is not informative, i.e., $\Pr(\delta_k = 1 \mid \mathbf{x}_k, y_k) = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$. The availability of auxiliary variables associated with both δ_k and y_k is key to making this assumption plausible and reducing the participation bias.

The non-probability survey weight $\hat{w}_k^{\text{NP}}$ can then be calibrated (e.g., Deville and Särndal, 1992) to achieve greater efficiency gains as well as a double robustness property (e.g., Chen, Li and Wu, 2020; Valliant, 2020). Calibration of the non-probability survey weight $\hat{w}_k^{\text{NP}}$ may be particularly efficient when auxiliary variables strongly predictive of y_k are available, which were excluded from the modelling of p_k. We focus next on the modelling and estimation of the participation probability p_k.

3. Estimation of the participation probability using a logistic model

The most common model for the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$ is the logistic model $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k' \boldsymbol{\alpha})\right]^{-1}$, where $\boldsymbol{\alpha}$ is a vector of unknown model parameters. Assuming $\mathbf{x}_k$ is observed for all $k \in U$, and δ_k are mutually independent, an estimator of $\boldsymbol{\alpha}$ can be found by solving the unbiased maximum likelihood estimating equation:

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{k \in U} \left[\delta_k - p_k(\boldsymbol{\alpha})\right] \mathbf{x}_k = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}. \tag{3.1}$$

The resulting maximum likelihood estimator is denoted by $\tilde{\boldsymbol{\alpha}}$ and satisfies $\mathbf{U}(\tilde{\boldsymbol{\alpha}}) = \mathbf{0}$. The estimated participation probability is denoted by $\tilde{p}_k = p_k(\tilde{\boldsymbol{\alpha}})$.

The estimating equation (3.1) cannot be used when the vector of auxiliary variables $\mathbf{x}_k$ is only observed for $k \in s_{\text{NP}}$ and missing for $k \in U - s_{\text{NP}}$. Chen, Li and Wu (2020) proposed to estimate $\sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k$ in (3.1) using a probability survey. The resulting pseudo maximum likelihood estimating equation is

$$\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}, \tag{3.2}$$

where w_k is the probability survey weight for unit $k \in s_P$. For simplicity, we assume in our theoretical developments that $w_k = \pi_k^{-1}$, where π_k is the probability that population unit k is selected in s_P. This weight ensures that $E_d[\hat{\mathbf{U}}(\boldsymbol{\alpha})] = \mathbf{U}(\boldsymbol{\alpha})$, where the subscript d indicates that the expectation is taken with respect to the probability sampling design. As a result, the estimating equation (3.2) is unbiased with respect to both the participation model and the probability sampling design. In practice, the survey weight w_k is often obtained after adjusting the basic design weight, π_k^{-1}, for nonresponse and calibration. The estimating equation (3.2) requires knowing the vector $\mathbf{x}_k$ for all $k \in s_{\text{NP}}$ and all $k \in s_P$ but not for all $k \in U$. Its solution yields the pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$, which satisfies $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$. The resulting estimated participation probability is denoted by $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$. Note that the estimating equation (3.2) may not have a solution. This is more likely to occur when n^{NP}/N is large and the probability sample is small (see Beaumont, 2020). This was not an issue in our experimentations since n^{NP}/N was smaller than 1%. Beaumont (2020) argued that the occurrence of inexistent solutions may be reduced by replacing the logistic model with the exponential model.

Chen, Li and Wu (2020) considered the case where the auxiliary variables are given. In practice, it may be necessary to choose relevant auxiliary variables and their interactions among a large set of candidate auxiliary variables. In the applications we have experimented with so far, the candidate auxiliary variables are often categorical (e.g., education, marital status, etc.). Blindly crossing all these variables may lead to a huge number of groups with many small groups, even empty. This was our motivation for finding methods that could select relevant auxiliary variables and their interactions.

We consider a stepwise selection procedure that attempts to minimize a modified version of the AIC, which properly accounts for the probability sampling design used to draw s_P. The justification for this modified AIC is provided in Section 3.1, and our selection procedure is described in Section 3.2. Section 3.3 considers an important special case of the logistic model: the homogeneous group model. In Section 3.4, a simple rank-based method for creating homogeneous groups is proposed. Finally, in Section 3.5, the recent method of Wang, Valliant and Li (2021) is discussed and contrasted with the method of Chen, Li and Wu (2020).

3.1 A modified AIC for the logistic model that accounts for the probability sampling design

Let us first consider the case where $\mathbf{x}_k$ is known for all the population units $k \in U$. Assuming δ_k are mutually independent, we can write the log likelihood function as

$$\begin{aligned} l(\boldsymbol{\alpha}) &= \sum_{k \in U} \delta_k \log[p_k(\boldsymbol{\alpha})] + (1-\delta_k)\log[1-p_k(\boldsymbol{\alpha})] \\ &= \sum_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum_{k \in U} \log[1-p_k(\boldsymbol{\alpha})]. \end{aligned}$$

Let us define $l_0(\boldsymbol{\alpha}) = E_m[l(\boldsymbol{\alpha})]$, where the subscript m indicates that the expectation is taken with respect to the true unknown participation model. The maximum likelihood estimator $\tilde{\boldsymbol{\alpha}}$ maximizes $l(\boldsymbol{\alpha})$ and we denote by $\boldsymbol{\alpha}_0$, the value of $\boldsymbol{\alpha}$ that maximizes $l_0(\boldsymbol{\alpha})$. Under regularity conditions, the maximum likelihood estimator $\tilde{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}_0$ under the model, i.e., $\sqrt{N}\left(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) = O_p(1)$, where N is the population size.

The AIC is an estimator of $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})]$. It is well known that a consistent estimator of $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})]$ is

$$\text{AIC} = -2l(\tilde{\boldsymbol{\alpha}}) + 2q, \tag{3.3}$$

where q is the number of model parameters (or the number of auxiliary variables). Equation (3.3) is the original AIC expression and the most widespread in practice.

Let us now consider the case where $\mathbf{x}_k$ is known only for $k \in s_{\text{NP}}$ and $k \in s_P$. Chen, Li and Wu (2020) proposed the pseudo log likelihood function

$$\hat{l}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k \in s_P} w_k \log[1 - p_k(\boldsymbol{\alpha})]. \tag{3.4}$$

Using $w_k = \pi_k^{-1}$ ensures that $E_d[\hat{l}(\boldsymbol{\alpha})] = l(\boldsymbol{\alpha})$ and $E_{md}[\hat{l}(\boldsymbol{\alpha})] = l_0(\boldsymbol{\alpha})$. Under regularity conditions, the pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$, which maximizes $\hat{l}(\boldsymbol{\alpha})$ in (3.4), is consistent for $\boldsymbol{\alpha}_0$ under both the model and the sampling design, i.e., $\sqrt{n^P}\,(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = O_p(1)$, where n^P is the size of the probability sample.

Under pseudo maximum likelihood estimation, the AIC can be defined as an estimator of

$$-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})] = -2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}})] + 2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})].$$

In Appendix 1, we provide a sketch of the proof that

$$E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})] \approx q + \text{tr}\left[E_m\{\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}[-\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\right], \tag{3.5}$$

where the function $\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \partial\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ is given in (3.2) for the logistic model, and $\mathbf{H}_0(\boldsymbol{\alpha}) = \partial^2 l_0(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}'$. Our derivations follow closely those of Lumley and Scott (2015). From (3.5) and (A.3) in Appendix 1, a consistent estimator of $-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})]$ is

$$\text{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2q + 2\text{tr}\left\{\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)][-\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})]^{-1}\right\}, \tag{3.6}$$

where $\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ is any design-consistent estimator of $\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ and $\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \partial^2\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}'$. For the logistic model,

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = -\sum\nolimits_{k \in s_P} w_k p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]\,\mathbf{x}_k\mathbf{x}_k'. \tag{3.7}$$

The AIC expression (3.6) is similar to the one given in Lumley and Scott (2015) but they omitted the term $2q$. This term is negligible compared with the third term on the right-hand side of (3.6) when the sampling fraction n^P/N is negligible. However, the term $2q$ may not always be negligible compared with the third term of (3.6), even when n^P/N is small. This would tend to occur when the participation

probabilities $p_k(\boldsymbol{\alpha})$ are small, which is typically the case of online volunteer-based surveys like Statistics Canada's crowdsourcing surveys. Therefore, the term $2q$ should generally not be neglected unless the non-probability sample size is significantly larger than the probability sample size. Another reason for keeping $2q$ in the expression (3.6) is that it reduces to the standard AIC expression (3.3) when the probability sample is a census. The last term on the right-hand side of (3.6) can thus be interpreted as a penalty for using a probability sample instead of a complete census in the estimating equation (3.2). The smaller the probability sample, the larger the effect of the penalty on the AIC (3.6).

3.2 Stepwise selection of auxiliary variables and pairwise interactions

In the empirical Section 6, we use a stepwise procedure based on the AIC (3.6) to select auxiliary variables (main effects) and pairwise interactions. Our procedure starts with the naïve model, which only includes the intercept. At each step of the procedure, a variable (main effect or pairwise interaction) is either included in the model or, if it was previously included, removed from the model. The inclusion or removal of the variable that yields the largest reduction of the AIC (3.6) is selected. An interaction is only eligible for inclusion when both main effects have already been selected, and a main effect is only eligible for removal when it is not supporting any interaction. The procedure stops when no variable can be added or removed from the model, i.e., no further reduction of the AIC (3.6) is possible.

One issue with the selection of auxiliary variables in a participation model is that it ignores the relationships between auxiliary variables and the variables of interest. As a result, an auxiliary variable that would be weakly associated with participation but strongly associated with some of the variables of interest could be discarded from the final participation model. This could have a negative effect on the bias reduction of the estimator $\hat{\theta}_{\text{NP}}$ of the finite population parameter θ. It is thus advisable to consider variable selection methods that lean towards overfitting, such as the AIC, to reduce the risk of omitting a relevant auxiliary variable. Moderate overfitting may better control for bias at the expense of a possible increase in variance. Our intent is to avoid gross overfitting so as to stabilize $\hat{\theta}_{\text{NP}}$. As pointed out in Section 2, the above variable selection issue can also be dealt with by calibrating inverse probability weights $\hat{w}_k^{\text{NP}}$ using calibration variables that are predictive of the variables of interest.

3.3 The homogeneous group model

Consider a partition of the population U into G groups, U_g, $g = 1,...,G$, and let $s_{\text{NP},g}$ and $s_{P,g}$ be the sets of units $k \in U_g$ that fall in the non-probability and probability samples, respectively. In the homogeneous group model, the participation probability is assumed to be constant for all units $k \in U_g$, i.e., $p_k \equiv p_g$, $k \in U_g$, $g = 1,...,G$. The homogeneous group model can be viewed as a special case of the logistic model with $q = G$, $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_g, \ldots, \alpha_G)$ and $\mathbf{x}_k' = (x_{1k}, \ldots, x_{gk}, \ldots, x_{Gk})$, where x_{gk} is a binary variable that equals 1 if $k \in U_g$, and that equals 0, otherwise. Therefore, for a unit $k \in U_g$, $p_k(\boldsymbol{\alpha}) = p(\alpha_g) \equiv p_g = [1 + \exp(-\alpha_g)]^{-1}$, and thus $\alpha_g = \log[p_g/(1-p_g)]$. For this model, the pseudo log likelihood function (3.4) reduces to

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{g=1}^{G} n_g^{\mathrm{NP}} \log\left[\frac{p(\alpha_g)}{1-p(\alpha_g)}\right] + \hat{N}_g \log[1-p(\alpha_g)], \tag{3.8}$$

where n_g^{NP} is the size of $s_{\mathrm{NP},g}$ and $\hat{N}_g = \sum_{k \in s_{p,g}} w_k$ is the estimated population size in group g obtained from the probability sample. The pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}' = (\hat{\alpha}_1, \ldots, \hat{\alpha}_g, \ldots, \hat{\alpha}_G)$, which maximizes $\hat{l}(\boldsymbol{\alpha})$ in (3.8), is such that $\hat{\alpha}_g = \log[\hat{p}_g/(1-\hat{p}_g)]$, $g = 1, \ldots, G$, where

$$\hat{p}_g = \frac{n_g^{\mathrm{NP}}}{\hat{N}_g}. \tag{3.9}$$

From (3.8), we can write $\hat{l}(\hat{\boldsymbol{\alpha}})$ as

$$\hat{l}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G} \hat{N}_g [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g) \log(1-\hat{p}_g)]. \tag{3.10}$$

For the homogeneous group model, the estimating function $\hat{\mathbf{U}}(\boldsymbol{\alpha})$ in (3.2) reduces to $[\hat{\mathbf{U}}(\boldsymbol{\alpha})]' = [\hat{U}_1(\alpha_1), \ldots, \hat{U}_g(\alpha_g), \ldots, \hat{U}_G(\alpha_G)]$, where

$$\hat{U}_g(\alpha_g) = n_g^{\mathrm{NP}} - \hat{N}_g p(\alpha_g). \tag{3.11}$$

Also, from (3.7), the matrix $\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})$ reduces to a diagonal matrix with the g^{th} element on the diagonal given by

$$\hat{H}_g(\hat{\alpha}_g) = -\hat{N}_g \hat{p}_g (1-\hat{p}_g). \tag{3.12}$$

Let $\boldsymbol{\alpha}_0' = (\alpha_{0,1}, \ldots, \alpha_{0,g}, \ldots, \alpha_{0,G})$. Using (3.11) and (3.12), the AIC (3.6) becomes

$$\mathrm{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} \frac{\hat{v}_d[\hat{U}_g(\alpha_{0,g})]}{\hat{N}_g \hat{p}_g (1-\hat{p}_g)}, \tag{3.13}$$

where $\hat{v}_d[\hat{U}_g(\alpha_{0,g})]$ is a design-consistent estimator of $\mathrm{var}_d[\hat{U}_g(\alpha_{0,g})]$. Using (3.11), a consistent variance estimator is

$$\hat{v}_d[\hat{U}_g(\alpha_{0,g})] = \hat{p}_g^2 \hat{v}_d(\hat{N}_g), \tag{3.14}$$

where $\hat{v}_d(\hat{N}_g)$ is a design-consistent estimator of $\mathrm{var}_d(\hat{N}_g)$. Using (3.14), the AIC (3.13) can be rewritten as

$$\mathrm{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} n_g^{\mathrm{NP}} \frac{[\mathrm{cv}_d(\hat{N}_g)]^2}{1-\hat{p}_g}, \tag{3.15}$$

where $\mathrm{cv}_d(\hat{N}_g) = \sqrt{\hat{v}_d(\hat{N}_g)}/\hat{N}_g$ is the estimated coefficient of variation of $\hat{N}_g$. Again, the last term on the right-hand side of (3.13) or (3.15) can be interpreted as a penalty for estimating the unknown population sizes N_g, $g = 1, \ldots, G$, using a probability sample.

Using (3.9), we obtain the non-probability survey weight of a unit $k \in s_{\text{NP},g}$ as

$$\hat{w}_k^{\text{NP}} = \hat{p}_k^{-1} = \frac{\hat{N}_g}{n_g^{\text{NP}}}. \tag{3.16}$$

The non-probability survey weight (3.16) shows the importance of avoiding groups for which n_g^{NP} is very small, even zero, so as to reduce the occurrence of extreme weights. Using (3.16), the inverse probability weighted estimator of the population total θ can be written as

$$\hat{\theta}_{\text{NP}} = \sum\nolimits_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k = \sum_{g=1}^{G} \hat{N}_g \bar{y}_g^{\text{NP}}, \tag{3.17}$$

where $\bar{y}_g^{\text{NP}} = \sum\nolimits_{k \in s_{\text{NP},g}} y_k \Big/ n_g^{\text{NP}}$ is the average of variable of interest y over units in $s_{\text{NP},g}$. The estimator (3.17) is simply a post-stratified estimator and satisfies the calibration equations $\sum\nolimits_{k \in s_{\text{NP},g}} \hat{w}_k^{\text{NP}} = \hat{N}_g$, $g = 1, \ldots, G$. The groups (post-strata) are constructed to be homogeneous with respect to the participation indicator. If they are also homogeneous with respect to the variable of interest then the post-stratified estimator (3.17) has a double robustness property (e.g., see Chen, Li and Wu, 2020; and Valliant, 2020).

We have assumed so far that the group membership is pre-determined for every population unit. In practice, homogeneous groups are often defined after observing sample data. There are several methods of constructing sample-dependent homogeneous groups. In Section 3.4, we propose a simple rank-based method that partitions the non-probability sample with respect to estimated participation probabilities from a logistic model. An extension of CART, nppCART, is described in Section 4. Once the non-probability and probability samples have been partitioned into sample-dependent groups, weights can be computed using (3.16) as if the group memberships were fixed.

3.4 A rank-based method for creating homogeneous groups

The first step of this method consists of estimating participation probabilities using a logistic model (with or without stepwise selection). We denote by $\hat{p}_k^{\text{logistic}} = p_k(\hat{\boldsymbol{\alpha}})$ these estimated participation probabilities, which are computed for each $k \in s_{\text{NP}}$ and $k \in s_P$. The idea is then to form G groups that are homogeneous with respect to $\hat{p}_k^{\text{logistic}}$ so as to make the homogeneous group model plausible. Once the groups are formed, the estimated probabilities $\hat{p}_k^{\text{logistic}}$ are discarded and the non-probability survey weights are computed using (3.16).

There are many methods for partitioning s_{NP} into homogeneous groups. A simple and popular method is to form groups with an equal number of participants (e.g., Eltinge and Yansaneh, 1997, formed groups with an equal number of sample units in the context of survey nonresponse). This method is equivalent to determining group boundaries from equal-width intervals in the range of r_k, $k \in s_{\text{NP}}$, where r_k is the rank of $\hat{p}_k^{\text{logistic}}$. We propose below a generalization of this method that retains the simplicity of assigning units based on their rank, but allows some flexibility so that the classes do not need to be equal-sized.

Rather than making equal-width bins in the range of r_k, we propose to form G equal-width bins in the range of $f(r_k)$, a monotone function of the rank r_k. We call it the Frank method. All the non-probability sample units that fall in a given bin are assigned to the same group. Any non-linear function f would thus

make smaller groups (fewer units) where the slope is steeper and larger groups where the slope is flatter. We propose the function

$$f(r_k) = \log\left(1 + a\frac{r_k}{n^{\text{NP}}}\right),$$

$k \in s_{\text{NP}}$, where n^{NP} is the size of the non-probability sample and a is a non-negative pre-specified constant that determines the degree of non-linearity. This function is concave down, with a larger slope and smaller groups for the lower-ranked units. The constant a determines the size of this effect, with a large value (e.g., $a = 100$) providing groups that are more unequal in size. The limit as a approaches 0 from above renders this function linear and so returns the equal-sized groups. The rank can be defined in ascending order of $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ for the smallest $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ for the second smallest $\hat{p}_k^{\text{logistic}}$, etc.), in which case the units with smaller estimated probabilities will be in the smaller groups, or in descending order of $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ for the largest $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ for the second largest $\hat{p}_k^{\text{logistic}}$, etc.), in which case the units with larger estimated probabilities will be in the smaller groups. The Frank method is somewhat similar to forming equal-width groups but with the groups bunched toward one end or the other, depending on whether $\hat{p}_k^{\text{logistic}}$ are sorted in ascending or descending order. Figure 1A in Appendix 2 illustrates the Frank method for $a = 10$, $G = 15$ and $n^{\text{NP}} = 31{,}415$, which is the size of the non-probability sample used in our empirical study in Section 6.

Once the non-probability sample has been partitioned into groups, each probability sample unit must then be assigned to one of the groups. Because the function f is monotone, each group contains non-probability sample units with values of $\hat{p}_k^{\text{logistic}}$ within a certain interval, and the intervals of any two different groups do not overlap so that the groups can be sorted based on their average value of $\hat{p}_k^{\text{logistic}}$. The boundary between any two consecutive groups is taken as the midpoint between the largest $\hat{p}_k^{\text{logistic}}$ from the group with the smaller average and the smallest $\hat{p}_k^{\text{logistic}}$ from the other group. Once all the boundaries have been determined, each probability sample unit $k \in s_P$ is assigned to the group with boundaries that cover $\hat{p}_k^{\text{logistic}}$.

The application of the Frank method requires determining suitable values of a and G as well as sorting $\hat{p}_k^{\text{logistic}}$, $k \in s_{\text{NP}}$, in ascending or descending order before computing the ranks r_k. Each possible choice leads to a different set of groups. We propose to determine the values of a and G, and the sorting order, by looking at different options and choosing the one that yields the smallest value of the AIC (3.15). This is investigated empirically in Section 6.3.

3.5 Adjusted logistic propensity weighting

As pointed out in the introduction, Wang, Valliant and Li (2021) proposed an extension of the pooling method to account for a non-negligible overlap between the probability and non-probability samples. The justification of their method, called Adjusted Logistic Propensity (ALP) weighting, is not based on a true likelihood approach, but still yields an *md*-unbiased estimating equation given by

$$\hat{\mathbf{U}}^{\text{ALP}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \frac{1}{1 + p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k - \sum_{k \in s_P} w_k \frac{p_k^{\text{ALP}}(\boldsymbol{\alpha})}{1 + p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k = \mathbf{0}, \tag{3.18}$$

where $p_k^{\text{ALP}}(\boldsymbol{\alpha}) = \exp(\mathbf{x}_k'\boldsymbol{\alpha})$. The estimating equation (3.18) is not equivalent to (3.2). However, if all the participation probabilities are small, both estimating equations should yield similar estimates of the participation probabilities.

An important difference between Wang, Valliant and Li (2021) and Chen, Li and Wu (2020) is the choice of the participation model. Chen, Li and Wu (2020) modelled the participation probability using a logistic function whereas Wang, Valliant and Li (2021) considered an exponential function. The logistic model is more natural as it ensures that estimated participation probabilities are always within the $(0,1)$ interval. This is to be contrasted with the exponential model, which may produce estimated probabilities greater than 1. Wang, Valliant and Li (2021) conducted a simulation study to evaluate their method. Their results show that (3.18) yields estimates of population means that are more robust to model failure than (3.2). This robustness could be explained by the use of the exponential model.

For the homogeneous group model, we have seen in Section 3.3 that the solution of (3.2) yields $p_k(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, for every unit $k \in U_g$. It is straightforward to show that the solution of (3.18) for the homogeneous group model also yields $p_k^{\text{ALP}}(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, for every unit $k \in U_g$. The equivalence between (3.2) and (3.18) for the homogeneous group model suggests that, in general, the two methods may produce similar estimates of θ, particularly when estimated probabilities are used only for the purpose of creating homogeneous groups (e.g., using the Frank method described in Section 3.4).

Wang, Valliant and Li (2021) also proposed a scaled version of their ALP method. Although the scaled estimating equation is not *md*-unbiased anymore, the authors showed its effectiveness in a simulation study for the estimation of population means. We tested the ALP method, including its scaled version, in our empirical experiments. The resulting estimates (not reported) were close to the pseudo maximum likelihood estimates of Chen, Li and Wu (2020), particularly after creating homogeneous groups. This observation is not surprising considering that the non-probability sample size is smaller than 1% of the population size in our experiments and that the estimated participation probabilities tend to be quite small. A thorough comparison of ALP and pseudo maximum likelihood estimation is left for future research.

One of the objectives of this paper was to develop a variable selection procedure applicable to the data integration scenario described in Section 2. Wang, Valliant and Li (2021) did not tackle the problem of variable selection. An AIC based on Lumley and Scott (2015) is not appropriate with ALP (or its scaled version) because the underlying estimating equation is not justified through a true likelihood approach. However, if ALP were preferable in a given context, variable selection could first be based on the pseudo likelihood method of Chen, Li and Wu (2020) and then ALP could be applied using the selected auxiliary variables.

4. Estimation of the participation probability using nppCART

The CART tree-growing procedure, developed by Breiman, Friedman, Olshen and Stone (1984), is a recursive binary partitioning algorithm that minimizes a certain objective function. For a binary dependent variable such as δ_k, a suitable objective function is the entropy impurity. For a given partition, U_g, $g = 1, \ldots, G$, the entropy impurity is given by

$$I = -\sum_{g=1}^{G} \frac{N_g}{N} [\tilde{p}_g \log(\tilde{p}_g) + (1-\tilde{p}_g) \log(1-\tilde{p}_g)],$$

where N_g is the size of U_g, $N = \sum_{g=1}^{G} N_g$ and $\tilde{p}_g = n_g^{\text{NP}} / N_g$. The entropy impurity cannot be computed when N_g is unknown. We propose to replace N_g with the survey-weighted estimator $\hat{N}_g$. This yields the computable objective function

$$\hat{I} = -\sum_{g=1}^{G} \frac{\hat{N}_g}{\hat{N}} [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g) \log(1-\hat{p}_g)], \tag{4.1}$$

where $\hat{p}_g$ is given in (3.9) and $\hat{N} = \sum_{g=1}^{G} \hat{N}_g$. The estimated entropy impurity (4.1) is proportional to the pseudo log likelihood function (3.10) under the homogeneous group model since $\hat{I} = -\hat{l}(\hat{\boldsymbol{\alpha}}) / \hat{N}$.

The recursive binary partitioning algorithm starts by examining all the possible splits of the non-probability sample s_{NP} into two groups. A split is any binary partition of s_{NP} based on the categories or numerical values of one of the candidate auxiliary variables. For instance, a split could be "SEX = male" and "SEX = female" or "AGE < 25" and "AGE ≥ 25". For each split of s_{NP}, the probability sample s_P is also split using the same binary partition. A split is said to be inadmissible and is rejected if it satisfies any of the following three stopping criteria:

i) $n_g^{\text{NP}} < C_{\text{NP}}$, for $g = 1$ or $g = 2$, where $C_{\text{NP}} \geq 1$ is a pre-determined constant specifying the minimum number of participants in a group;

ii) $n_g^{\text{NP}} \geq \hat{N}_g$, for $g = 1$ or $g = 2$;

iii) $n_g^{P} < C_P$, for $g = 1$ or $g = 2$, where n_g^{P} is the size of $s_{P,g}$ and $C_P \geq 1$ is a pre-determined constant specifying the minimum number of probability sample units in a group.

Then, the estimated entropy impurity (4.1) with $G = 2$ is computed for each admissible split, and the best of those admissible splits, i.e., the one that has the smallest value of (4.1), is selected to form the first two groups. If all the splits are inadmissible or the best split does not decrease the objective function (4.1) then partitioning into two groups is not done.

After the determination of the first two initial groups, the same splitting operation is repeated for each of the two groups, and so on and so forth, layer by layer, until all the groups cannot be split further based on the stopping criteria. We say that this process results in a fully grown tree although it is a slight abuse of language as there are stopping criteria that limit its growth. The above procedure, the earlier version of which was called nppCART by Chu and Beaumont (2019), is essentially identical to the original CART algorithm, except for the use of the estimated entropy (4.1) and the three stopping criteria above. The stopping criterion (i) ensures that the non-probability survey weight $\hat{w}_k^{\text{NP}}$ in (3.16) does not become extreme. The stopping criterion (ii) ensures that the estimated probability $\hat{p}_g$ is always smaller than 1. The last criterion is added to ensure that the estimator $\hat{N}_g$ is not too unstable.

Chu and Beaumont (2019) developed an R program that implements the nppCART algorithm. They showed in a simulation study that this algorithm was effective for reducing the participation bias although the resulting post-stratified estimator (3.17) had a variance somewhat larger than its competitors. This

instability might be explained by overfitting, i.e., the creation of too many groups. The usual recommendation to avoid overfitting is to prune the tree after it has been grown. Pruning is usually applied in two steps. In the first step, a finite sequence of nested subtrees of decreasing size and increasing impurity is determined, starting with the fully grown tree that has the maximum number of groups and ending with the degenerate subtree that contains only one group. In the second step, the best of these nested subtrees is selected, often through K-fold cross-validation. This pruning approach is equivalent to penalizing the objective function with an additive penalty term defined as the product of a positive penalty parameter and the number of groups. Cross-validation is then typically used to determine an optimal value for the penalty parameter. Greater detail on pruning can be found in Breiman, Friedman, Olshen and Stone (1984); see also Izenman (2008, Chapter 9). In the context of survey nonresponse, classification and regression trees have been explored by Phipps and Toth (2012) and Lohr, Hsu and Montaquila (2015).

However, as pointed out in the introduction, classical cross-validation methods cannot be directly applied to the data integration scenario studied in this paper, and this topic requires further research. As an alternative to cross-validation for the selection of the best subtree, among a set of nested subtrees of decreasing size and increasing impurity, we propose to choose the subtree that minimizes the AIC (3.15). This AIC takes the probability sampling design into account through the estimation of the design variance of $\hat{N}_g$ (see Section 5). This variance could be readily estimated in our experiments in Section 6 using available bootstrap weights. Similar to variable selection, discussed in Section 3.2, pruning is intended to avoid gross overfitting so as to stabilize $\hat{\theta}_{\text{NP}}$.

5. Bootstrap variance estimation

It is not enough to produce inverse probability weighted estimates of finite population parameters; it is also important to provide users with indicators of the quality of those estimates. We propose a bootstrap procedure to estimate the variance of inverse probability weighted estimators with a focus on the post-stratified estimator (3.17). The variance may be useful but has some limitations since it is derived under the assumption that the participation model is correctly specified and that the inverse probability weighted estimators are unbiased. The absence of bias depends critically on the availability and proper choice of auxiliary variables so as to make the non-informative participation assumption reasonable. Although some amount of bias seems unavoidable in practice, the computation of variance estimates may nonetheless provide some useful information for comparison and evaluation purposes, as illustrated in Section 6.

The bootstrap variance estimator that we propose accounts for two sources of variability: the probability sampling design and the participation model. We suppose that B bootstrap weights $w_k^{(b)}$, $b = 1, \ldots, B,$ are available for each unit $k \in s_P,$ and that these bootstrap weights properly capture the variability due to the probability sampling design. For instance, we assume that these bootstrap weights can be used to obtain a design-consistent estimator of $\text{var}_d(\hat{N}_g)$ as

$$\hat{v}_d^{\text{boot}}(\hat{N}_g) = \frac{1}{B}\sum_{b=1}^{B}(\hat{N}_g^{(b)} - \hat{N}_g)^2, \tag{5.1}$$

where $\hat{N}_g^{(b)} = \sum_{k \in s_{P,g}} w_k^{(b)}$ is the b^{th} bootstrap replicate of $\hat{N}_g$. The Rao, Wu and Yue (1992) bootstrap weights are often used in social surveys conducted by Statistics Canada. They are applicable for stratified multistage designs when the first-stage sampling fractions are small and can incorporate weight adjustments, such as nonresponse adjustments and calibration. Beaumont and Émond (2022) proposed an extension of the method that removes the requirement of small first-stage sampling fractions.

The unknown participation mechanism is modelled as a Poisson sampling design, where population units are assumed to participate independently of one another with probability p_k, $k \in U$. For Poisson sampling, Beaumont and Patak (2012) pointed out that valid bootstrap weights for sample units $k \in s_{\text{NP}}$ can be written as $p_k^{-1} a_k^{(b)}$, $b = 1, \ldots, B$, provided that the bootstrap factors $a_k^{(b)}$ are generated independently of one another using a distribution that is not too heavily skewed with a mean of one and a variance of $1 - p_k$. For a non-probability sample, the true participation probability p_k is unknown but can be replaced with a consistent estimator $\hat{p}_k$. Following Beaumont and Émond (2022), who studied bootstrap under survey nonresponse, we thus suggest generating the bootstrap factors $a_k^{(b)}$, $k \in s_{\text{NP}}$ and $b = 1, \ldots, B$, independently of one another using the gamma distribution with a mean of one and a variance of $1 - \hat{p}_k$. The choice of the gamma distribution is to ensure non-negative bootstrap factors $a_k^{(b)}$.

The bootstrap estimator of the variance of the inverse probability weighted estimator $\hat{\theta}_{\text{NP}}$, $\text{var}_{md}(\hat{\theta}_{\text{NP}})$, is given by

$$\hat{v}_{md}^{\text{boot}}(\hat{\theta}_{\text{NP}}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{\text{NP}}^{(b)} - \hat{\theta}_{\text{NP}})^2, \tag{5.2}$$

where $\hat{\theta}_{\text{NP}}^{(b)}$ is the b^{th} bootstrap replicate of $\hat{\theta}_{\text{NP}}$. Assuming the logistic model is used with fixed auxiliary variables, the b^{th} bootstrap replicate of $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, with $\hat{w}_k^{\text{NP}} = [p_k(\hat{\boldsymbol{\alpha}})]^{-1}$, is $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k$, where $\hat{w}_k^{\text{NP},(b)} = a_k^{(b)} / p_k(\hat{\boldsymbol{\alpha}}^{(b)})$, and $\hat{\boldsymbol{\alpha}}^{(b)}$ is the solution of the b^{th} bootstrap replicate of estimating equation (3.2):

$$\hat{\mathbf{U}}^{(b)}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} a_k^{(b)} \mathbf{x}_k - \sum_{k \in s_P} w_k^{(b)} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}.$$

Assuming now that the homogeneous group model is used, the b^{th} bootstrap replicate of the post-stratified estimator (3.17) can be written as

$$\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k = \sum_{g=1}^{G} \hat{N}_g^{(b)} \bar{y}_g^{\text{NP},(b)}, \tag{5.3}$$

where $\hat{w}_k^{\text{NP},(b)} = a_k^{(b)} \hat{N}_g^{(b)} / n_g^{\text{NP},(b)}$, for $k \in s_{\text{NP},g}$, $n_g^{\text{NP},(b)} = \sum_{k \in s_{\text{NP},g}} a_k^{(b)}$ and $\bar{y}_g^{\text{NP},(b)} = \sum_{k \in s_{\text{NP},g}} a_k^{(b)} y_k / n_g^{\text{NP},(b)}$. The bootstrap replicate (5.3) is valid provided that the homogeneous groups are fixed. This simplification is often made when estimating the variance of estimators adjusted for survey nonresponse, even when the homogeneous groups are determined adaptively from the observed sample data. In our context, it would not be straightforward to develop a bootstrap procedure that correctly accounts for variable selection or pruning. In particular, a double bootstrap might be required if the design variance estimators involved in the AIC (3.6) or (3.15) were obtained through bootstrap weights. Treating auxiliary variables or homogeneous

groups as fixed, when they are not, should tend to underestimate the variance $\text{var}_{md}(\hat{\theta}_{\text{NP}})$. Although the magnitude of the underestimation is expected to be small to moderate, further research is needed on this topic.

6. Empirical evaluation of methods using real data

We evaluated and compared inverse probability weighting methods, discussed in Sections 3 and 4, using real data. In Section 6.1, we present the three data sources used in our investigations. Methods are described in Section 6.2 and results are given in Sections 6.3 and 6.4.

6.1 Data sources and variables

After the beginning of the COVID-19 lockdown in March 2020, Statistics Canada conducted a series of crowdsourcing surveys to respond to urgent information needs about the life of the Canadian population. Each crowdsourcing survey collected data from visitors of Statistics Canada's website who responded voluntarily to a short online questionnaire. Renaud and Beaumont (2020) provide greater detail on crowdsourcing experiments conducted by Statistics Canada.

We investigated the use of the Labour Force Survey (LFS) as a means of reducing the participation bias of crowdsourcing estimates. Except for the Census, the LFS is the most important social probability survey conducted by Statistics Canada with a sample containing around 56,000 selected households each month. Data are collected for all eligible persons within responding households. The household response rate was around 90% before the pandemic but fell to around 70% in June 2020. In our empirical study, we used data of the June 2020 LFS sample, which contains responses for 87,779 persons. The LFS is based on a stratified multistage design and a regression composite estimator (see Gambino, Kennedy and Singh, 2001). Rao, Wu and Yue (1992) bootstrap weights are produced and made available to users for variance estimation.

In parallel to crowdsourcing experiments, Statistics Canada also started a series of probability web panel surveys: the Canadian Perspective Survey Series (CPSS). The CPSS sample is obtained from previous LFS respondents. The June 2020 CPSS initial probability sample was relatively large with over 30,000 selected persons but the overall recruitment/response rate was quite low at around 15%; this resulted in 4,209 respondents in June 2020. Greater detail on the CPSS can be found in Baribeau (2020).

In June 2020, participants from previous crowdsourcing experiments were also randomly chosen and sent the same questionnaire as CPSS respondents; 31,415 participants responded to the questionnaire. This allowed for a comparison of estimates from this crowdsourcing non-probability sample with those from the CPSS probability sample.

Table 6.1 shows naïve crowdsourcing estimates and CPSS estimates for nine selected proportions. For the first two proportions, LFS estimates are also available and very close to the corresponding CPSS estimates. This is not unexpected as nonresponse in the CPSS is adjusted using education and employment status. Both probability surveys show large differences with naïve crowdsourcing estimates for these two

proportions. The following five proportions also show significant differences between naïve crowdsourcing and CPSS estimates whereas estimates from both sources are similar for the last two proportions.

Table 6.1
Proportions of interest.

Proportion	Description	Naïve crowdsourcing estimate	CPSS estimate	LFS estimate
θ_1	Proportion of people having a university degree.	64.5%	30.6%	30.2%
θ_2	Proportion of people who worked at a job or business during the reference week.	65.4%	50.1%	50.3%
θ_3	Proportion of people whose usual place of work is a fixed location outside the home.	50.2%	40.2%	-
θ_4	Proportion of people who worked most of their hours at home during the reference week.	45.6%	19.3%	-
θ_5	Proportion of people who report having "more than enough" income to meet their household needs.	32.1%	15.9%	-
θ_6	Proportion of people who are "very likely" to get COVID-19 vaccine when available.	74.2%	57.3%	-
θ_7	Proportion of people who are "very concerned" about the health risk posed by gathering in large groups.	70.0%	54.4%	-
θ_8	Proportion of people who "fear being a target for putting others at risk" because they do not always wear a mask in public.	9.9%	9.8%	-
θ_9	Proportion of people who report ordering the same amount of take-out food as before.	45.6%	46.2%	-

In a first step, we used June 2020 LFS data to reduce the participation bias of naïve crowdsourcing estimates using inverse probability weighting methods discussed in Sections 3 and 4. The candidate auxiliary variables available in both the crowdsourcing and LFS samples were: age group (13 levels), sex (2 levels), economic region (56 levels), education (8 levels), immigration status (3 levels), household size (6 levels), marital status (6 levels) and employment status (3 levels). Greater detail on these eight auxiliary variables is given in Appendix 3. Then, we used non-probability survey weights to compute adjusted crowdsourcing estimates for the nine proportions defined in Table 6.1 and compared them to those obtained using the CPSS probability sample alone. These results are provided in Section 6.3. Note that a proportion is defined as $\theta = N^{-1}\sum_{k\in U} y_k$, where y_k is a binary variable of interest, and is estimated by $\hat{\theta}_{\text{NP}} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$. For the first two proportions in Table 6.1, the variable of interest y_k can be derived from auxiliary variables. We thus expect weighting methods to successfully remove the participation bias for these proportions.

In a second step, we obtained adjusted crowdsourcing estimates using June 2020 CPSS data instead of LFS data with the same candidate auxiliary variables as above. Our objective was to evaluate the effect on bias reduction of using a smaller probability sample. These results are provided in Section 6.4.

6.2 Methods

We investigated the eight methods described in Table 6.2 below. For methods 3, 5 and 6, which involve a logistic model with the stepwise selection procedure described in Section 3.2, all main effects and pairwise interactions were considered as candidate variables to be included or removed from the model. For these

methods, the estimator $\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$, required to compute the AIC (3.6), was obtained using bootstrap weights as

$$\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)] = \frac{1}{B}\sum_{b=1}^{B}[\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})][\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})]',$$

where

$$\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}}) = \sum_{k\in s_{\mathrm{NP}}}\mathbf{x}_k - \sum_{k\in s_P} w_k^{(b)} p_k(\hat{\boldsymbol{\alpha}})\,\mathbf{x}_k.$$

For methods 4, 5, 6 and 8, the estimator $\hat{v}_d(\hat{N}_g)$, required to compute the AIC (3.15), is obtained from (5.1). For methods, 6, 7 and 8, which use nppCART, we set $C_{\mathrm{NP}} = 5$ and $C_P = 5$ in the stopping criteria (i) and (iii) given in Section 4.

Table 6.2
Description of methods.

Method	Model	Stepwise selection	Homogeneous groups	Description
1	Intercept	-	-	Naïve logistic model with only the intercept (or homogeneous group model with only one group).
2	Logistic	-	-	Logistic model including all main effects but no interaction.
3	Logistic	Yes	-	Logistic model with stepwise selection of main effects and pairwise interactions by minimizing the AIC (3.6).
4	Logistic	-	Frank	Method 2 followed by creation of homogeneous groups using the Frank method, described in Section 3.4, with sorting in ascending order, $a = 10$ and the number of groups roughly minimizing the AIC (3.15).
5	Logistic	Yes	Frank	Method 3 followed by creation of homogeneous groups using the Frank method, described in Section 3.4, with sorting in ascending order, $a = 10$ and the number of groups roughly minimizing the AIC (3.15).
6	Logistic	Yes	nppCART with pruning	Method 3 followed by creation of homogeneous groups using nppCART with pruning minimizing the AIC (3.15); only one auxiliary variable is provided to nppCART: the estimated participation probability from the logistic model.
7	-	-	nppCART without pruning	nppCART based on all candidate auxiliary variables without pruning.
8	-	-	nppCART with pruning	nppCART based on all candidate auxiliary variables with pruning minimizing the AIC (3.15).

6.3 Results when integrating crowdsourcing data with the LFS probability sample

Stepwise selection results for the logistic model

Using the LFS as the probability sample, our stepwise selection procedure described in Section 3.2 resulted in the selection of all main effects along with 15 pairwise interactions for a total of 395 model parameters. Six main effects entered the model before any interaction in the following order: education, economic region, immigration status, sex, age group and household size. Together, they accounted for more than 95% of the total AIC reduction (difference between AIC of methods 1 and 3). The variable education alone accounted for more than 40% of the total AIC reduction. For these data, it thus appears that

interactions are not as important as the main effects to reduce the AIC. This suggests that a model including all the main effects but no interaction might be reasonable.

Comparisons of AIC values

Table 6.3 shows values of the Relative AIC (RAIC) for the eight methods described in Table 6.2. The Relative AIC is defined as

$$\text{RAIC} = \frac{\text{AIC}_0 - \text{AIC}}{\text{AIC}_0} \times 100\%,$$

where AIC_0 is the value of the AIC (3.6) for the naïve model containing only the intercept. For methods 1, 2 and 3, the RAIC is computed using the AIC (3.6) whereas it is computed using the AIC (3.15) for methods 4 to 8 assuming the groups are fixed. The RAIC can be interpreted similarly to the coefficient of determination in linear regression: it is 0 for the naïve model, it increases as the AIC decreases, and it is always smaller than 1. However, it can take negative values unlike the coefficient of determination. A model that has a larger RAIC than a competitor suggests that its auxiliary variables are better predictors of participation. Table 6.3 also shows the number of model parameters q or the number of groups G; q is shown for methods 1, 2 and 3, and G is shown for methods 4 to 8.

Table 6.3
RAIC values in percentage.

Method	Model	Stepwise selection	Homogeneous groups	RAIC (%)	q or G	Proportion (%) of AIC from the 1st term	Proportion (%) of AIC from the 2nd term	Proportion (%) of AIC from the 3rd term
1	Intercept	-	-	0	1	100.00	0.00	0.00
2	Logistic	-	-	10.7	90	99.90	0.04	0.06
3	Logistic	Yes	-	11.1	395	99.59	0.18	0.23
4	Logistic	-	Frank	10.7	100	99.89	0.05	0.07
5	Logistic	Yes	Frank	11.3	100	99.88	0.05	0.07
6	Logistic	Yes	nppCART with pruning	12.2	1,276	97.99	0.59	1.42
7	-	-	nppCART without pruning	11.9	3,165	96.23	1.45	2.33
8	-	-	nppCART with pruning	12.5	1,772	97.58	0.82	1.60

The RAIC varies from 10.7% to 12.5% for methods 2 to 8; thus, all these methods provide a meaningful improvement over the naïve method. Comparing methods 2 and 3, we observe that accounting for pairwise interactions yielded only a small improvement of the RAIC, as noted above. Using the Frank method to create homogeneous groups did not significantly improve the RAIC. This is an indication that the logistic model was reasonable for these data. The use of nppCART resulted in an increase of RAIC, albeit not substantial. This may indicate that nppCART has achieved some robustness. However, nppCART also resulted in a number of groups significantly larger than other methods, even after pruning. Given the AIC (3.15) assumes the groups are fixed (although they are not), this improvement of RAIC should not be over-interpreted.

Table 6.3 also shows the proportion of the AIC that comes from each of the three terms on the right-hand side of (3.6) or (3.15). Not surprisingly, the first term, $-2\hat{l}(\hat{\boldsymbol{\alpha}})$, is the dominant component of the AIC. The relative importance of the other two terms increases with q or G. Both terms have similar importance although the third term is always slightly larger than the second term. In this application, none of the terms should be omitted in the computation of the AIC.

The Frank Method

Figures 1B and 1C in the Appendix 2 illustrate the Frank method of creating homogeneous groups for method 5 in Table 6.2. Figure 1B shows a graph of $\hat{p}_k^{\text{logistic}}$ as a function of the rank r_k for both the non-probability and probability samples. It also shows the corresponding boundaries, in terms of the ranks, for $G=15$ and different values of a, and for both sorting orders. Figure 1B illustrates that the groups containing smaller values of $\hat{p}_k^{\text{logistic}}$ are under-represented in the non-probability sample, compared with the probability sample, because these units are less likely to participate. Figure 1B also illustrates that sorting in ascending order produces groups that are closer to being equal-sized in the probability sample, particularly when a is large. This has the advantage of reducing the occurrence of groups that contain too few probability sample units, which could lead to unstable weights. A value of $a=5$ or $a=10$, along with sorting in ascending order, seems to offer a suitable compromise for both samples.

Figure 1C shows the values of the AIC (3.15) as a function of the number of groups G for a few values of a and both sorting orders. It appears that the sorting order makes a significant difference on the AIC, with lower values obtained when $\hat{p}_k^{\text{logistic}}$, $k \in s_{\text{NP}}$, are sorted in ascending order. Figure 1C does not show much sensitivity to the choice of a but the best values seem to occur near $a=10$. Notably, the optimal number of classes is near 100 in this application, much larger than the value of 5 that is often recommended (e.g., Eltinge and Yansaneh, 1997). Based on these results, we chose to sort in ascending order and used $a=10$ and $G=100$ when applying the Frank method with LFS data. A smaller number of groups was chosen with the CPSS data (see Section 6.4).

With these data, forming groups with an equal number of participants $(a=0)$ was slightly inferior to $a=10$ in terms of AIC (see Figure 1C). However, both values of a led to similar estimates (results not shown).

Comparisons of estimates

Table 6.4 shows estimates and their bootstrap standard errors (in italic) for each of the nine proportions in Table 6.1 and each method described in Table 6.2. The bootstrap standard error is the square root of the bootstrap variance estimate given in (5.2). The b^{th} bootstrap replicate of the estimated proportion $\hat{\theta}_{\text{NP}} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$ is $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k \Big/ \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)}$. For methods 4 to 8, the bootstrap weights $\hat{w}_k^{\text{NP},(b)}$ are obtained under the simplification that the homogeneous groups are fixed. Bootstrap standard errors are not computed for methods 2 and 3. The CPSS estimates and their design-based standard errors are also provided for comparison purposes in the last row of Table 6.4. The CPSS estimates are believed to be less biased than adjusted crowdsourcing estimates since they are obtained from a probability

survey, albeit with a small response rate (around 15%), with nonresponse weight adjustments and calibration.

From the estimates and standard errors in Table 6.4, we make the following observations:

- Methods 2 to 8 are all roughly equivalent.
- For the first seven proportions, where the naïve estimates (method 1) are significantly different from the CPSS estimates, methods 2 to 8 yield adjusted crowdsourcing estimates closer to CPSS estimates, which suggests a non-negligible bias reduction. Indeed, for the first three proportions, the adjusted crowdsourcing estimates are not markedly different from the CPSS estimates. It is not surprising for the first two proportions since the variables of interest can be derived from auxiliary variables. This observation is particularly interesting for the third proportion. For proportions 4 to 7, the bias reduction is not so spectacular, albeit not negligible; the adjusted crowdsourcing estimates lie in between the naïve and CPSS estimates.
- For the last two proportions, the naïve, adjusted crowdsourcing and CPSS estimates are all similar. A slight but not alarming discrepancy between adjusted crowdsourcing and CPSS estimates is observed for the last proportion for methods 2 and 3, which do not use homogeneous groups. Overall, it is reassuring to observe that inverse probability weighting did not introduce significant biases for the last two proportions.
- Finally, the standard errors for the naïve method are much smaller than those for the other methods. This indicates that naïve estimates are likely more stable. However, the standard error does not account for bias and should not be the main criterion for choosing an appropriate method.

Table 6.4
Estimates and standard errors (in italic) in percentage.

Method	Model	Stepwise selection	Homogeneous groups	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Intercept	-	-	64.5	65.4	50.2	45.6	32.1	74.2	70.0	9.9	45.6
				0.27	*0.26*	*0.27*	*0.28*	*0.27*	*0.24*	*0.26*	*0.17*	*0.28*
2	Logistic	-	-	29.7	50.2	40.4	28.0	23.5	67.9	62.4	11.4	43.5
				-	-	-	-	-	-	-	-	-
3	Logistic	Yes	-	28.9	48.2	39.8	26.6	23.3	68.1	64.1	10.2	42.3
				-	-	-	-	-	-	-	-	-
4	Logistic	-	Frank	32.4	52.1	40.6	29.5	23.5	68.0	63.5	10.7	44.9
				0.41	*0.76*	*0.70*	*0.58*	*0.60*	*0.74*	*0.78*	*0.49*	*0.77*
5	Logistic	Yes	Frank	30.8	51.4	39.8	28.5	22.4	67.9	64.0	10.3	44.4
				0.35	*0.86*	*0.78*	*0.63*	*0.59*	*0.82*	*0.89*	*0.54*	*0.87*
6	Logistic	Yes	nppCART with pruning	30.9	50.7	39.5	28.4	22.9	67.8	63.7	10.4	44.5
				0.36	*0.84*	*0.78*	*0.70*	*0.79*	*1.02*	*1.00*	*0.62*	*1.02*
7	-	-	nppCART without pruning	30.2	52.7	40.6	28.0	24.3	69.3	65.4	9.4	46.8
				0.29	*0.88*	*0.91*	*0.46*	*0.82*	*0.91*	*0.96*	*0.42*	*0.74*
8	-	-	nppCART with pruning	30.2	52.5	40.5	28.0	23.8	69.4	65.2	9.3	47.0
				0.29	*0.87*	*0.91*	*0.47*	*0.81*	*0.90*	*1.03*	*0.39*	*0.78*
CPSS estimate				**30.6**	**50.1**	**40.2**	**19.3**	**15.9**	**57.3**	**54.4**	**9.8**	**46.2**
				0.87	***1.25***	***1.14***	***0.97***	***0.87***	***1.41***	***1.33***	***0.86***	***1.42***

With these data, methods 2 to 8 performed similarly. This may be due to the large size of the LFS probability sample. In order to study the behaviour of inverse probability weighting methods when the probability sample is smaller, we replaced the LFS by the CPSS probability sample. Results for this case are discussed below.

6.4 Results when integrating crowdsourcing data with the CPSS probability sample

Stepwise selection results for the logistic model

When we used the CPSS as the probability sample, our stepwise selection procedure selected again all main effects but only 10 pairwise interactions for a total of 254 model parameters. All but one main effect entered the model before any interaction in the following order: education, household size, economic region, sex, immigration status, age group and marital status. For these data, pairwise interactions were again not as important as the main effects to reduce the AIC.

Comparisons of AIC values

Table 6.5 shows values of the RAIC for the eight methods described in Table 6.2. Comparing methods 2 and 3, we observe that accounting for pairwise interactions yielded only a small improvement of the RAIC. For these data, the creation of homogeneous groups resulted in a non-negligible increase of the RAIC. In particular, when a logistic model is used along with stepwise selection, the RAIC is 12.1 and it increases to 18.5 after forming homogeneous groups with nppCART. The use of nppCART without a logistic model (methods 7 and 8) also yielded a larger RAIC than methods 2 and 3. The effect of pruning remains negligible with these data since the RAIC of methods 7 and 8 are similar. However, pruning reduced the number of groups from 600 to 451. The replacement of the LFS sample by the CPSS sample resulted in a reduction of the number of groups for methods 4 to 8; this is not surprising since the CPSS sample size is significantly smaller than the LFS sample size.

Table 6.5 also shows the proportion of the AIC that comes from each of the three terms on the right-hand side of (3.6) or (3.15). Again, the first term, $-2\hat{l}(\hat{\boldsymbol{\alpha}})$, is the dominant component of the AIC, and the relative importance of the other two terms increases with q or G. Given the small CPSS sample size, the third term, which can be viewed as a penalty for using a probability sample instead of a census, is now relatively much larger than the second term $2q$ (or $2G$). The second term could thus be omitted, as in Lumley and Scott (2015), although there is no computational advantage of neglecting it.

Comparisons of estimates

Table 6.6 shows estimates and their bootstrap standard errors (in italic) for each of the nine proportions in Table 6.1 and each method described in Table 6.2. We make the following observations:

- For the first two proportions, the variables of interest can be derived from auxiliary variables, and we expect inverse probability weighting methods to entirely remove bias. Methods 7 and 8

(nppCART without a logistic model) basically eliminated the discrepancy between the naïve and CPSS estimates. Other methods were not so successful although method 4 (logistic model with main effects followed by the Frank method) performed relatively well.

- Method 2 appeared to over-adjust the naïve estimates for the first three proportions. Forming homogeneous group (method 4) corrected for this over-adjustment.
- Methods 2 and 3 (logistic model without homogeneous groups) were somewhat erratic. This may be explained by variable and extreme non-probability survey weights, particularly for method 3. The coefficient of variation of the non-probability survey weights is provided in Table 6.7 for each method. It is 7.5 and 39.7 for methods 2 and 3, respectively, whereas it is no greater than 5.5 for all the other methods. This shows the importance of forming homogeneous groups to reduce extreme weights. By comparison, when the LFS is used as the probability sample, the coefficient of variation of the non-probability survey weights is 4.7 and 6.3 for methods 2 and 3, respectively, and it is no greater than 4.0 for all the other methods.
- Methods that use stepwise selection tended to under-adjust when homogeneous groups were formed (methods 5 and 6), particularly for the first proportion. This was not expected given their large values of RAIC in Table 6.5. However, the RAIC only indicates the strength of the association between the auxiliary variables and participation. It does not account for the strength of the association between the auxiliary variables and variables of interest, which can affect the magnitude of participation bias and variance.
- Comparing methods 5 and 6, we observe that the creation of homogeneous groups using the Frank method and nppCART yielded similar estimates with nppCART estimates tending to be slightly closer to CPSS estimates, possibly due to the larger number of groups with nppCART.
- Pruning did not show significant improvements in our experiments since methods 7 and 8 produced similar estimates.
- Overall, nppCART with or without pruning (methods 7 and 8) appeared to be the most stable and reliable method for reducing participation bias followed closely by method 4 (logistic model with main effects only along with the Frank method).

It is interesting to observe that nppCART estimates in Table 6.6 (methods 7 and 8) were not markedly different from the corresponding estimates in Table 6.4 based on the LFS probability sample. This suggests that a small probability sample can succeed at reducing bias even though it remains preferable to use a larger probability sample. For nppCART, using the LFS as the probability sample was just slightly better than using the CPSS. For other methods, the differences were sometimes much larger and using the LFS provided better estimates. This may be an argument to favour nppCART when the probability sample size is small.

Table 6.5
RAIC values in percentage.

Method	Model	Stepwise selection	Homogeneous groups	RAIC (%)	q or G	Proportion (%) of AIC from the 1st term	Proportion (%) of AIC from the 2nd term	Proportion (%) of AIC from the 3rd term
1	Intercept	-	-	0	1	100.00	0.00	0.00
2	Logistic	-	-	11.2	90	98.45	0.04	1.50
3	Logistic	Yes	-	12.1	254	96.27	0.12	3.62
4	Logistic	-	Frank	13.4	20	98.18	0.01	1.80
5	Logistic	Yes	Frank	15.9	16	99.35	0.01	0.64
6	Logistic	Yes	nppCART with pruning	18.5	384	96.43	0.19	3.38
7	-	-	nppCART without pruning	14.3	600	95.93	0.28	3.78
8	-	-	nppCART with pruning	14.4	451	96.27	0.21	3.51

Table 6.6
Estimates and standard errors (in italic) in percentage.

Method	Model	Stepwise selection	Homogeneous groups	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Intercept	-	-	64.5	65.4	50.2	45.6	32.1	74.2	70.0	9.9	45.6
				0.28	*0.28*	*0.29*	*0.29*	*0.28*	*0.25*	*0.25*	*0.17*	*0.28*
2	Logistic	-	-	21.3	44.4	34.4	24.4	22.8	69.1	61.3	10.2	44.9
				-	-	-	-	-	-	-	-	-
3	Logistic	Yes	-	29.4	43.4	28.3	29.8	27.4	78.4	71.8	10.1	27.6
				-	-	-	-	-	-	-	-	-
4	Logistic	-	Frank	34.1	50.9	39.4	30.2	25.8	70.8	66.6	9.8	45.1
				0.59	*0.61*	*0.56*	*0.51*	*0.50*	*0.55*	*0.58*	*0.36*	*0.59*
5	Logistic	Yes	Frank	43.6	54.6	41.8	34.3	27.4	71.7	67.9	9.7	44.6
				0.67	*0.54*	*0.50*	*0.55*	*0.43*	*0.44*	*0.47*	*0.30*	*0.47*
6	Logistic	Yes	nppCART with pruning	42.0	54.0	41.2	34.2	27.3	70.8	67.4	10.1	44.6
				0.81	*0.77*	*0.73*	*0.71*	*0.63*	*0.69*	*0.66*	*0.44*	*0.70*
7	-	-	nppCART without pruning	30.8	48.9	39.1	28.5	27.7	71.5	64.9	8.9	47.1
				0.98	*1.38*	*1.41*	*0.80*	*1.35*	*1.23*	*1.46*	*0.56*	*1.49*
8	-	-	nppCART with pruning	30.8	49.8	38.7	29.3	27.1	71.5	65.2	9.3	46.8
				0.98	*1.27*	*1.28*	*0.78*	*1.24*	*1.20*	*1.41*	*0.80*	*1.35*
CPSS estimate				**30.6**	**50.1**	**40.2**	**19.3**	**15.9**	**57.3**	**54.4**	**9.8**	**46.2**
				0.87	***1.25***	***1.14***	***0.97***	***0.87***	***1.41***	***1.33***	***0.86***	***1.42***

Table 6.7
Coefficients of variation of the non-probability survey weights.

Probability sample	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
CPSS	0	7.5	39.7	1.8	1.4	2.2	5.5	5.0
LFS	0	4.7	6.3	2.6	3.0	3.6	4.0	3.9

7. Conclusion

We extended the pseudo maximum likelihood method of Chen, Li and Wu (2020) that integrates data from a non-probability and probability sample: We developed a variable selection procedure for the logistic model and an extension of CART, nppCART. Inspired by Lumley and Scott (2015), our extensions use a modified AIC that properly accounts for the probability sampling design. In our investigations, we observed that the additional penalty term for using a probability sample instead of a census was not negligible.

Not surprisingly, our experimentations illustrated that inverse probability weighting methods can reduce participation bias, but sometimes a significant bias remains. For the large LFS probability sample, all the methods performed similarly. Significant differences between methods were observed when the smaller CPSS probability sample was used. In particular, our experimentations showed the importance of creating homogeneous groups to reduce the occurrence of extreme weights and improve the stability and robustness of estimates. For the small probability sample, accounting for pairwise interactions somewhat reduced the AIC but was generally not beneficial for the estimates. Main effects appeared more important than pairwise interactions to reduce the AIC with our data. Overall, the best method for bias reduction was nppCART followed closely by the use of a logistic model with main effects only along with the creation of homogeneous groups. However, different conclusions could potentially be drawn with smaller domains or other datasets.

It is well known that inverse probability weighted estimators may be inefficient, particularly when the variables of interest are weakly related to the weights. This can be addressed through calibration on known population totals or totals estimated from the probability sample. Calibration will be particularly efficient when auxiliary variables strongly related to the variables of interest are available and excluded from the participation model. This was not the case in our experimentations. Weight smoothing is an alternative aiming to improve the efficiency of inverse probability weighted estimators, which may be useful when such powerful calibration variables are not available. It consists of replacing the weights with predictions obtained by modelling the weights conditionally on the variables of interest. In the context of integrating non-probability and probability samples, weight smoothing was studied by Ferri-Garcia, Beaumont, Bosa, Charlebois and Chu (2021).

Tree-based methods more sophisticated than the CART algorithm, such as random forests, are available in the literature. Given the good performance of nppCART in our experimentations, it could be worthwhile to extend those methods to the data integration scenario considered in this paper and evaluate them. Further developments are needed on this topic.

There is most likely no inverse probability weighting method that is uniformly better than all the other methods. All the techniques are useful and can be part of the statistician's toolkit. However, there is a need for the development of bias reduction indicators that would help statisticians in choosing the best method for a given non-probability and probability sample. The relative AIC and the coefficient of variation of the non-probability survey weights are two useful indicators but they do not tell the full story as they do not say anything about the strength of the association between the auxiliary variables and variables of interest. One

idea that could be explored would be to use statistical matching methods with nonparametric models (e.g., random forests) for each variable of interest conditionally on the auxiliary variables. The resulting estimates would be expected to be more efficient than inverse probability weighting methods because they would be tailored to each variable of interest. In practice, this statistical matching strategy would be tedious to apply as a different model would need to be developed and validated for each estimate produced. However, a few statistical matching estimates could be computed and used to evaluate inverse probability weighting methods. We might expect that a better inverse probability weighting method would generally tend to yield estimates closer to statistical matching estimates. A possible procedure to reconcile the two methods would be to calibrate inverse probability weights so that the resulting estimates agree exactly with selected statistical matching estimates.

Appendix 1

Sketch of the proof of equation (3.5)

Using first-order Taylor expansions, we have

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + \left[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0) - \mathbf{U}_0(\boldsymbol{\alpha}_0)\right]' (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p\left(\frac{N}{n^P}\right) \tag{A.1}$$

and

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \hat{\mathbf{U}}(\boldsymbol{\alpha}_0) + \hat{\mathbf{H}}(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p\left(\frac{N}{\sqrt{n^P}}\right), \tag{A.2}$$

where $\mathbf{U}_0(\boldsymbol{\alpha}) = \partial l_0(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. In addition to (A.1) and (A.2), we also assume that

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \mathbf{H}_0(\boldsymbol{\alpha}) + o_p(N) \tag{A.3}$$

under the model and the sampling design. Noting that $\mathbf{U}_0(\boldsymbol{\alpha}_0) = \mathbf{0}$ and $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, we obtain from (A.1), (A.2) and (A.3),

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)' \left[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\right](\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p\left(\frac{N}{n^P}\right). \tag{A.4}$$

Ignoring the smaller order term and taking the expectation of both sides of (A.4) yield:

$$E_{md}\left[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})\right] \approx \operatorname{tr}\left[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}})\right], \tag{A.5}$$

where $\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}}) = E_{md}[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)']$. Using (A.2) and (A.3), and ignoring the smaller order terms, we can approximate this variance as

$$\begin{aligned}
\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}}) &\approx [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \operatorname{var}_{md}[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)][\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \\
&= [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \left\{\operatorname{var}_m[\mathbf{U}(\boldsymbol{\alpha}_0)] + E_m\{\operatorname{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}\right\}[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \\
&= -[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} + [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} E_m\{\operatorname{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1},
\end{aligned} \tag{A.6}$$

where $\mathbf{U}(\boldsymbol{\alpha}) = \partial l(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$ is given in equation (3.1) for the logistic model. The last equation in (A.6) results from a well-known property of the Fisher information matrix $-\mathbf{H}_0(\boldsymbol{\alpha}_0)$ (assuming the true model is in the same parametric family as the postulated model). Using (A.6) in (A.5) yields result (3.5).

Appendix 2

Illustration of the Frank method

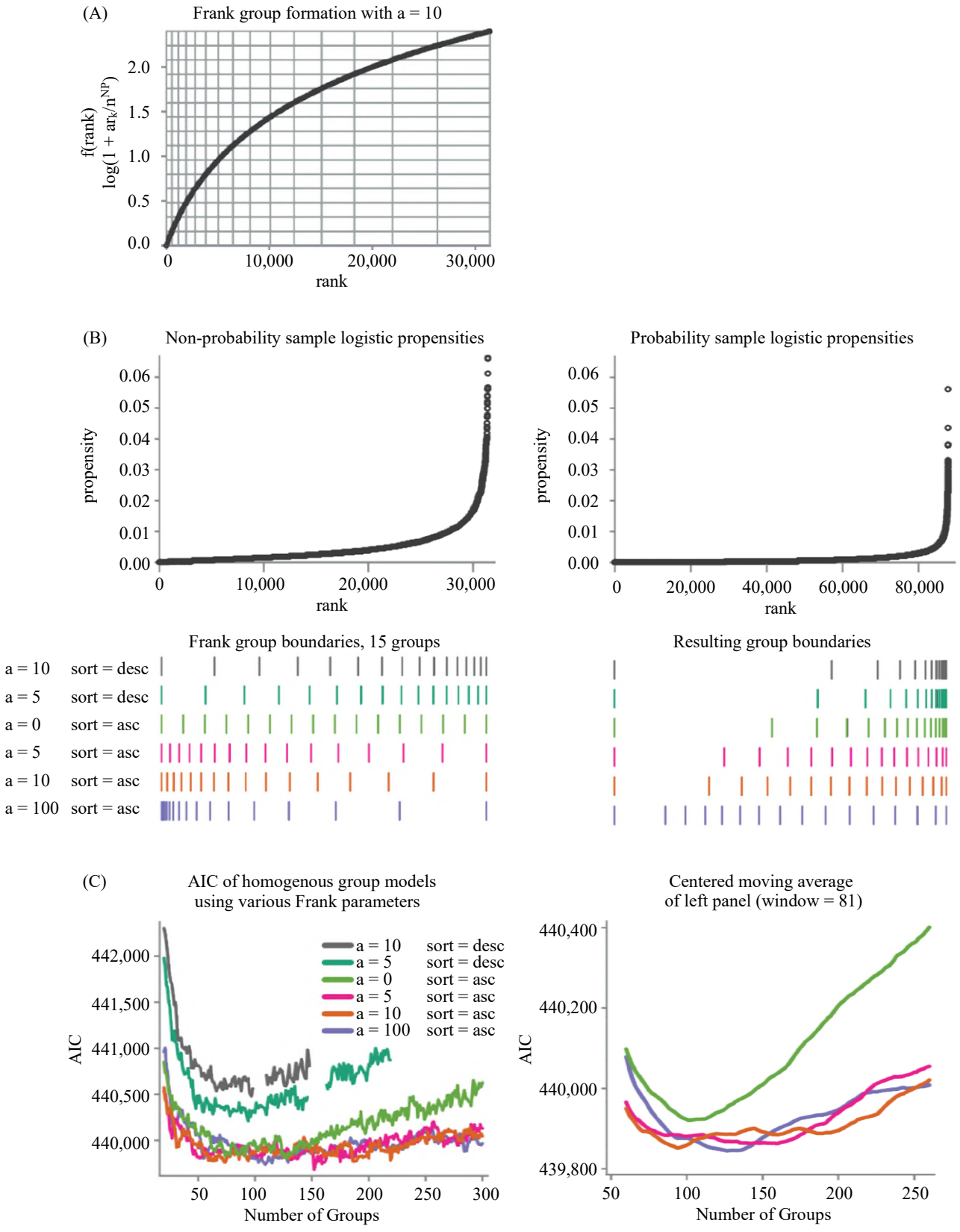
Figure 1 below contains three sub-figures, Figures 1A, 1B and 1C, that illustrate the behaviour of the Frank method for the data described in Section 6.1 and for method 5 described in Section 6.2 when the LFS is used as the probability sample. The description of each sub-figure is provided below:

(A) Frank method with $a = 10$, $G = 15$ and n^{NP} = 31,415. The rank, r_k, is on the horizontal axis and the function of the rank, $f(r_k) = \log(1 + a\, r_k / n^{\text{NP}})$, is on the vertical axis. The bins are equal-width in the range of $f(r_k)$. The constant a determines the shape of the function. As a increases, it becomes increasingly non-linear and the groups are more bunched to one side.

(B) The top panels show the sorted values of $\hat{p}_k^{\text{logistic}}$ for the non-probability (left) and probability (right) samples. Fifteen groups are formed based on the non-probability sample using Frank with different values of a and both sorting orders, resulting in different group boundaries as represented by the coloured bars in the bottom panels. For the non-probability sample (bottom left panel), when the rank is defined in ascending order of $\hat{p}_k^{\text{logistic}}$, the groups are smaller for small values of $\hat{p}_k^{\text{logistic}}$. When the rank is defined in descending order of $\hat{p}_k^{\text{logistic}}$, the groups are smaller for large values of $\hat{p}_k^{\text{logistic}}$. Increasing a increases the bunching, while $a = 0$ gives equal-sized groups.

(C) The AIC (3.15) versus the number of groups for different values of a and both sorting orders. Sorting $\hat{p}_k^{\text{logistic}}$ in ascending order leads to smaller values of AIC, without much sensitivity to changes in the value of a. The AIC is minimized with around 100 groups for all parameterizations. The right panel smooths the left panel using a centered moving average filter with window size 81. The smoothed curves show the Frank method performs slightly better than equal-sized groups $(a = 0)$, especially when the number of groups is higher than optimal, adding some robustness to the choice of the number of groups. When the number of groups is large and $\hat{p}_k^{\text{logistic}}$ are sorted in descending order, it occurs that some groups do not contain any probability sample unit. As a result, $\hat{p}_g$ is undefined for those groups, and the AIC cannot be computed.

Figure 1 Illustration of the Frank method.



Appendix 3

Auxiliary variables

Age Group: 5-year age groups, starting from 15-19 and ending with 75+.

Sex: Male/Female.

Education: 8 categories (Less than high school; High school; Some post-secondary; Trades certificate or diploma; Community college, CEGEP, etc.; University certificate below Bachelor's; Bachelor's degree; Above Bachelor's degree).

Economic Region: Sub-provincial geography partitioning the country. It contains 73 levels, but some were collapsed due to insufficient respondent counts; 56 levels were used in the models.

Immigration: 3 levels (Born in Canada; Landed immigrant; Not a landed immigrant).

Household Size: Number of people in the household, regardless of age, capped at 6.

Marital Status: 6 levels (Married; Common-law; Widow or widower; Separated; Divorced; Single, Never married).

Employment Status: 3 levels (Employed and at work at least part of the reference week; Employed but absent from work; Not employed).

References

Bahamyirou, A., and Schnitzer, M.E. (2021). Data integration through outcome adaptive LASSO and a collaborative propensity score approach. *arXiv preprint arXiv:2103.15218*.

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, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Beaumont, J.-F., and Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5, 339-357.

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127-148.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Boca Raton, FL.

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.

Chu, K., and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, May 2019.

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 Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 1, 33-40. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf.

Ferri-Garcia, R., Beaumont, J.-F., Bosa, K., Charlebois, J. and Chu, K. (2021). Weight smoothing for nonprobability surveys. *TEST* (published online).

Ferri-Garcia, R., and Rueda, M.d.M. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT*, 42, 159-182.

Gambino, J., Kennedy, B., and Singh, M.P. (2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation. *Survey Methodology*, 27, 1, 65-74. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5855-eng.pdf.

Haziza, D., and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer Science & Business Media.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

Lohr, S., Hsu, V. and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Lumley, T., and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6, 772-794.

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

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 2, 209-217. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf.

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 from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 2, 231-263.

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New-York: John Wiley & Sons, Inc.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

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, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.