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

Survey Methodology

# Are probability surveys bound to disappear for the production of official statistics?

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

by Jean-François Beaumont

Release date: June 30, 2020

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** STATCAN.infostats-infostats.STATCAN@canada.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 |

**Depository Services Program**

| | |
|---|---|
| • Inquiries line | 1-800-635-7943 |
| • Fax line | 1-800-565-7757 |

## 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 2020
Vol. 46, No. 1, pp. 1-28
Statistics Canada, Catalogue No. 12-001-X

# Are probability surveys bound to disappear for the production of official statistics?

**Jean-François Beaumont**[1]

## Abstract

For several decades, national statistical agencies around the world have been using probability surveys as their preferred tool to meet information needs about a population of interest. In the last few years, there has been a wind of change and other data sources are being increasingly explored. Five key factors are behind this trend: the decline in response rates in probability surveys, the high cost of data collection, the increased burden on respondents, the desire for access to "real-time" statistics, and the proliferation of non-probability data sources. Some people have even come to believe that probability surveys could gradually disappear. In this article, we review some approaches that can reduce, or even eliminate, the use of probability surveys, all the while preserving a valid statistical inference framework. All the approaches we consider use data from a non-probability source; data from a probability survey are also used in most cases. Some of these approaches rely on the validity of model assumptions, which contrasts with approaches based on the probability sampling design. These design-based approaches are generally not as efficient; yet, they are not subject to the risk of bias due to model misspecification.

**Key Words:** Statistical matching; Calibration; Non-probabilistic data; Data integration; Fay-Herriot model; Propensity score.

# 1 Introduction

In 1934, Jerzy Neyman laid the foundation for probability survey theory and his design-based approach to inference with an article published in the *Journal of the Royal Statistical Society*. His article (Neyman, 1934) piqued the interest of a number of statisticians at the time, and the theory was developed further in the following years. Still today, many articles on this topic are published in statistics journals. Rao (2005) provides an excellent review of various developments in probability survey theory during the 20th century (see also Bethlehem, 2009; Rao and Fuller, 2017; Kalton, 2019). Nowadays, national statistical agencies, such as Statistics Canada and the Institut National de la Statistique et des Études Économiques (INSEE) in France, use probability surveys most often to get the information they seek on a population of interest.

The popularity of probability surveys for producing official statistics stems largely from the non-parametric nature of the inference approach developed by Neyman (1934). In other words, probability surveys allow for valid inferences about a population without having to rely on model assumptions. This is appealing – even fundamental, according to Deville (1991) – to national statistical agencies that produce official statistics. In fact, these agencies have historically been reluctant to take unnecessary risks, which are unavoidable for approaches that depend on the validity of model assumptions, especially when it is difficult to check the underlying assumptions.

[1] Jean-François Beaumont, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@canada.ca.

However, estimates from probability surveys can prove inefficient, even to the point of being unusable, particularly when the sample size is small (see, for example, Rao and Molina, 2015). Furthermore, they are based on the assumption that non-sampling errors, such as measurement, coverage or non-response errors, are negligible. To minimize these errors, national statistical agencies often invest considerable resources. For example, questionnaires are tested to ensure that respondents fully understand them; survey data are validated using various edit rules; respondents are contacted again, if necessary, to confirm the data collected; non-respondent follow-ups are conducted to minimize the impact of non-response on the estimates, etc. Despite all these efforts, non-sampling errors persist in practice. There are, of course, adaptations of the theory for taking these errors into account. These adaptations necessarily come with model assumptions and thus with the risk of bias resulting from inadequate assumptions. Probability surveys are not a panacea but they are generally recognized as providing a reliable source of information about a population, except when non-sampling errors become dominant. Brick (2011) takes the argument further and defends the idea that a probability survey with a low response rate – if properly designed – usually provides estimates with smaller bias than those obtained from a volunteer non-probability survey. Dutwin and Buskirk (2017) show empirical results that corroborate this argument.

For the past few years, a wind of change has been blowing over national statistical agencies, and other data sources are being increasingly explored. Five key factors are behind this trend: i) the decline in response rates in probability surveys in recent years; ii) the high cost of data collection; iii) the increased burden on respondents; iv) the desire for access to “real-time” statistics (Rao, 2020), in other words, having the ability to produce statistics practically at the same time or very shortly after the information needs are expressed; and v) the proliferation of non-probability data sources (Rancourt, 2019) such as administrative sources, social media, web surveys, etc. To control data collection costs of probability surveys and reduce the adverse effects of non-response on the quality of estimates, a number of authors have proposed and evaluated responsive data collection methods (e.g., Laflamme and Karaganis, 2010; Lundquist and Särndal, 2013; Schouten, Calinescu and Luiten, 2013; Beaumont, Haziza and Bocci, 2014; Särndal, Lumiste and Traat, 2016). Tourangeau, Brick, Lohr and Li (2017) review various methods and point out their limited success in reducing non-response bias and costs. Särndal et al. (2016) also reach the same conclusion regarding bias. Some surveys conducted by national statistical agencies still have very low response rates, and it becomes risky to rely solely on data collection and estimation methods to correct potential non-response biases. Indeed, a number of authors (e.g., Rivers, 2007; Elliott and Valliant, 2017) pointed out the similarity between a probability survey with a very low response rate and a non-probability survey. Yet, a non-probability survey has the advantages of having a usually much larger sample size and being less costly. Given the above discussion, some have come to believe that probability surveys could gradually disappear (see Couper, 2000; Couper, 2013; Miller, 2017).

However, data from non-probability sources are not without challenges, as noted by Couper (2000), Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013), and Elliott and Valliant (2017), among others. For example, it is well known that non-probability surveys that collect data from volunteers can often lead to estimates with significant selection bias (or participation bias). Bethlehem (2016) provides a bias expression and argues that the potential for bias is usually higher for a non-probability survey than for a probability survey affected by non-response. Meng (2018) illustrates that bias becomes dominant as the non-probability sample size increases, which significantly reduces the effective sample size. Therefore, the acquisition of large non-probability samples alone cannot ensure the production of estimates with an acceptable quality. The pre-election poll conducted by the *Literary Digest* magazine for predicting the outcome of the 1936 U.S. presidential election is a prime example of this (Squire, 1988; Elliott and Valliant, 2017). Despite a huge sample size of over two million people, the poll was unable to predict Franklin Roosevelt's overwhelming victory. Instead, it incorrectly predicted a convincing victory for his opponent, Alfred Landon. The set of poll respondents, who were highly unrepresentative of the voting population, was made up mainly of car and phone owners as well as the magazine's subscribers. Couper (2000) and Elliott and Valliant (2017) cite other more recent examples of non-probability surveys that led to erroneous conclusions.

Selection bias is not the only challenge that must be overcome when using data from a non-probability source. Another major challenge is the presence of measurement errors (e.g., Couper, 2000). They can significantly impact the estimates, especially when data are collected without relying on an experienced interviewer. This is the case for most non-probability sources, in particular volunteer web surveys.

The current context leads to the following question: How can data from a non-probability source be used to minimize, even eliminate, the data collection costs and respondent burden of a probability survey, all the while preserving a valid statistical inference framework and acceptable quality? That is the main question this article attempts to answer.

Most of the methods we present integrate data from a probability survey and a non-probability source. Zhang (2012) discusses the concept of statistical validity when integrated data are used to make inferences. We contend that establishing a statistical framework that can be used to make valid inferences is essential for the production of official statistics, a point that also seems to be shared by Rancourt (2019). Without such a framework, the usual properties of estimators, such as bias and variance, are not defined. It then becomes impossible to select estimators based on an objective criterion such as, for example, choosing the linear unbiased estimator with the smallest possible variance. Without a valid statistical inference framework, estimates can be calculated, but all the usual tools for determining the quality of those estimates and drawing accurate conclusions about the population's characteristics of interest are lost.

In the rest of this article, we differentiate design-based approaches to inference, described in Section 3, from model-based approaches to inference, described in Section 4. For each approach, we consider two scenarios: In the first one, the data from the non-probability source match exactly the concepts of interest and are not fraught with measurement errors. Those data can therefore be used to replace the data from a probability survey. In the second scenario, the data from the non-probability source do not reflect concepts of interest or are subject to measurement errors. Although these data cannot be used to directly replace data from a probability survey, they can still be used as auxiliary information to enhance it. In Section 5, we provide some additional thoughts. Let us first begin with some background in Section 2.

## 2 Background

One of the first steps to meet information needs is to define the target population for which that information is sought. We denote this target population by $U$. Then, it is necessary to define the parameters of interest, i.e. what it is desired to know about the target population. In practice, it is often desired to estimate many parameters. To simplify the discussion, we suppose that only one parameter is of interest: the total of the variable $y$, $\theta = \sum_{k \in U} y_k$, where $y_k$ is the value of the variable $y$ for unit $k$ of the population $U$. We use $\mathbf{Y}$ to denote the vector containing the values $y_k$ for $k \in U$. Lastly, a set of procedures must be established for the estimation of the parameter $\theta$ while taking into account various factors, such as the available budget, the respondent burden, the desired precision, etc. During this process, it is necessary to identify the data sources that will be used – probabilistic or not – and a statistical inference framework that will allow for assessing the properties of the estimates produced, such as bias and variance.

The above sequence, which starts with defining the target population and parameters of interest, followed by the data sources and estimation procedures, is consistent with the proposal by Citro (2014). She suggests that national statistical agencies first determine the information needs along with potential users. Next, they can work at identifying the data source(s) that will meet those needs while preserving an acceptable quality of estimates, keeping costs within the established budget and controlling for respondent burden. It seems preferable to avoid the reverse procedure, however tempting it is, of first identifying available data sources and then artificially determining the needs based on what can be produced by these sources. In general, this kind of procedure cannot adequately meet users' actual needs.

We assume that we have access to data from a non-probability source (e.g., administrative data, web survey data, etc.). Values are observed for a few variables, including a variable $y^*$, for all units of a subset of $U$, denoted as $s_{\text{NP}}$. The variable $y^*$ is not necessarily equal to $y$ because of conceptual differences and/or measurement errors. At least, it is hoped that there is a strong association between the two variables. We denote the inclusion indicator in $s_{\text{NP}}$ as $\delta_k$; in other words, $\delta_k = 1$ if unit $k$ is in $s_{\text{NP}}$ and $\delta_k = 0$, otherwise. The vector of the inclusion indicators $\delta_k$ for $k \in U$ is denoted by $\boldsymbol{\delta}$.

Data from a probability survey may also be available. In that case, a sample $s_P$ of the population $U$ is randomly selected with probability $p(s_P \mid \mathbf{Z})$. The matrix $\mathbf{Z}$ contains information available on the sampling frame that is used to define the sampling design, such as stratum identifiers for each unit of the population. The sample inclusion indicators, $I_k$, $k \in U$, are defined as follows: $I_k = 1$ if unit $k$ is selected in the sample $s_P$; otherwise, $I_k = 0$. We use $\mathbf{I}$ to denote the vector containing the sample inclusion indicators for $k \in U$. The probability that unit $k$ of the population $U$ is chosen in the sample is denoted by $\pi_k = E(I_k \mid \mathbf{Z})$. Most of the time it is known or can be approximated. We assume that

$\pi_k > 0, k \in U$. For each unit $k \in s_P$, the values of certain variables are collected, which may or may not include the variable $y$.

We use $\mathbf{\Omega}$ to denote the set of all the auxiliary data used to make inferences. Among other things, $\mathbf{\Omega}$ includes the design information, $\mathbf{Z}$, if a probability sample is used, and potentially other auxiliary variables such as calibration variables, matching variables or explanatory variables of a model (see Sections 3 and 4). The inclusion indicator $\delta_k$ can also be used as an auxiliary variable either for stratifying the population or for calibration (see Section 3). The vector $\boldsymbol{\delta}$ can thus be included in $\mathbf{Z}$ and $\mathbf{\Omega}$.

The following two assumptions are used throughout the article:

*Assumption* 1: $\mathbf{I}$ is independent of $\mathbf{\Omega}$ and $\mathbf{Y}$ after conditioning on $\mathbf{Z}$.
*Assumption* 2: $\boldsymbol{\delta}$ and $\mathbf{I}$ are independent after conditioning on $\mathbf{\Omega}$ and $\mathbf{Y}$.

Assumption 1 implies that the values of the variables included in $\mathbf{\Omega}$ and $\mathbf{Y}$ are not affected by whether or not a unit is included in the sample $s_P$. This is implicit in the literature on probability surveys and results from the very definition of the sampling design, which depends only on $\mathbf{Z}$. Assumption 2 is automatically satisfied if the non-probability source (and thus $\boldsymbol{\delta}$) is available prior to selecting the probability sample. Note that if $\delta_k$ is used as an auxiliary variable to stratify the population, then $\boldsymbol{\delta}$ is included in $\mathbf{\Omega}$ and assumption 2 is still satisfied. It will not be satisfied if being selected in $s_P$ impacts the provision of data to the non-probability source. For example, being selected in $s_P$ (and contacted) could be an indirect reminder for the selected individual to fill out forms required by the government (non-probability source). It can be expected that assumptions 1 and 2 are satisfied in most cases.

The union of $\mathbf{\Omega}, \boldsymbol{\delta}, \mathbf{I}$ and $\mathbf{Y}$ contains all the information used for making inferences. The various approaches to inference set out in Sections 3 and 4 differ in what they treat as fixed and what they treat as random. For example, in the design-based approach to inference, everything is considered fixed except for the vector $\mathbf{I}$; in other words, design-based inferences are conditional on $\mathbf{\Omega}, \mathbf{Y}$ and $\boldsymbol{\delta}$. To simplify the notation, we use $\mathbf{\Omega}_P$ to denote the union of $\mathbf{\Omega}, \mathbf{Y}$ and $\boldsymbol{\delta}$. Thus, design expectations are denoted as $E(\cdot \mid \mathbf{\Omega}_P)$ rather than $E(\cdot \mid \mathbf{\Omega}, \mathbf{Y}, \boldsymbol{\delta})$. In the design-based approach to inference, an estimator $\hat{\theta}$ of $\theta$ is usually chosen so that the design bias, $E(\hat{\theta} - \theta \mid \mathbf{\Omega}_P)$, is zero or negligible. Under assumptions 1 and 2, we note that $E(I_k \mid \mathbf{\Omega}_P) = E(I_k \mid \mathbf{Z}) = \pi_k$. For estimating the total $\theta = \sum_{k \in U} y_k$, an estimator of the form $\hat{\theta} = \sum_{k \in s_P} w_k y_k$ is frequently used, where $w_k$ is a survey weight for unit $k$. The standard basic weight is $w_k = \pi_k^{-1}$. This weight ensures that the estimator $\hat{\theta}$ is exactly design-unbiased for $\theta$. The basic weight can then be modified using calibration techniques (e.g., Deville and Särndal, 1992; Haziza and Beaumont, 2017). The advantage of this approach is its non-parametric nature: no model assumption is needed for making valid inferences about the population because the first two design moments are controlled by the statistician and are usually known. Yet, the approach is not free of assumptions, for example to ensure the consistency and asymptotic normality of estimators, but it does not require any parametric model.

In practice, non-response is often observed in probability surveys as well as other non-sampling errors. Non-response of some sample units is often viewed as an additional phase of sampling that is not controlled by the statistician. In other words, the non-response mechanism is not known, unlike the sampling design. Assuming an adequate model for the non-response mechanism, estimators with little or no bias can be obtained, for example by weighting the responding units by the inverse of their estimated response probability. However, this requires careful modelling of the response indicators. In the rest of this paper, we ignore non-sampling errors and assume that the estimates from the probability survey are not biased or, at least, that their bias is small compared to the bias of the estimates from the non-probability source alone. This assumption may not always be satisfied in practice, but it is reasonable in many contexts (see Brick, 2011), especially in large surveys conducted by national statistical agencies.

The acquisition of data from non-probability sources is generally inexpensive compared to the cost of collecting data from a probability survey. Therefore, they would ideally be used to replace data from a probability survey. This data replacement is valid only if $y_k^* \approx y_k, k \in U$. This assumption will not be satisfied with all non-probability data sources, but may be realistic with some administrative data sources. In Sections 3 and 4, we will differentiate the methods based on the assumption that $y_k^* = y_k$ from the methods not requiring this assumption. Several methods described in Sections 3 and 4 are also reviewed in the upcoming article by Rao (2020) that was presented to Statistics Canada in summer 2018.

# 3 Design-based approaches

Design-based approaches yield design-consistent estimators of $\theta$ even when the non-probability source produces estimates with a significant selection bias. In this context, the purpose of using a non-probability sample is to reduce the variance of estimators of $\theta$. The efficiency gains achieved can be used to justify a reduction of the probability sample size, thereby a reduction of the data collection costs and respondent burden. The methods that we consider in Sections 3.1 and 3.2 require collecting the values of the variable of interest $y$ in the probability sample, just like small area estimation methods described in Section 4.4. However, the efficiency gains are usually expected to be more modest than those obtained using small area estimation methods. In Section 3.1, we consider the scenario $y_k^* = y_k$ whereas in Section 3.2, we consider the scenario $y_k^* \neq y_k$.

## 3.1 Weighting by the inverse of the probability of inclusion in the combined sample

The ideal case occurs when the non-probability sample is a census, i.e., $s_{\mathrm{NP}} = U$. In that case, the value of the parameter of interest $\theta = \sum_{k \in U} y_k$ can be directly calculated without worrying about bias or variance since $y_k^* = y_k$ is assumed in this section. In general, we expect under-coverage in the sense that $s_{\mathrm{NP}}$ is smaller than the population $U$. In a design-based approach, the potential under-coverage bias can be addressed by selecting a probability sample $s_P$ from $U$ and collecting the values of the variable $y$ for the sample units. Ideally, the probability sample is drawn from $U - s_{\mathrm{NP}}$ but it is possible that the units in $s_{\mathrm{NP}}$ cannot be linked to those of the sampling frame $U$ to establish the set $U - s_{\mathrm{NP}}$. In general, the larger the non-probability sample, the more it is possible to reduce the size of the probability sample without jeopardizing the desired precision of the estimates.

It seems desirable to estimate $\theta$ using all the data collected in the combined sample $s = s_P \cup s_{\mathrm{NP}}$. The inclusion indicator in $s$ can be defined as $\tilde{I}_k = \delta_k + (1 - \delta_k) I_k$. To obtain a design-unbiased estimator of $\theta$, each unit $k \in s$ is weighted by $\tilde{w}_k = \tilde{\pi}_k^{-1}$ where $\tilde{\pi}_k = E\left(\tilde{I}_k \mid \mathbf{\Omega}_P\right)$. Under assumptions 1 and 2, $E\left(I_k \mid \mathbf{\Omega}_P\right) = \pi_k$ and we obtain

$$\tilde{\pi}_k = E\left(\tilde{I}_k \mid \mathbf{\Omega}_P\right) = \delta_k + (1 - \delta_k)\, \pi_k.$$

The resulting estimator is written:

$$\hat{\theta} = \sum\nolimits_{k \in s} \tilde{w}_k y_k = \sum\nolimits_{k \in s_{\mathrm{NP}}} y_k + \sum\nolimits_{k \in s_P} \frac{1}{\pi_k} (1 - \delta_k)\, y_k. \tag{3.1}$$

Note that estimator (3.1) requires the indicator $\delta_k$ to be available for all units in the sample $s_P$. For the units $k \in s_P \cap s_{\mathrm{NP}}$, we have two values: $y_k$ and $y_k^*$. In principle, we should have $y_k^* = y_k$, but it is possible that this relationship is not exactly satisfied. These units can be used to validate the assumption $y_k^* \approx y_k$. If significant differences are observed, it may be preferable to not consider this approach and to rely on the methods in Section 3.2 that use data from the non-probability source as auxiliary data. If we trust the data quality of the non-probability source, it may be advisable not to collect the variable $y$ in the probability sample for the units also present in the non-probability sample in order to reduce the data collection costs and respondent burden.

We can view the problem as if we had two sampling frames: $U$ and $s_{\mathrm{NP}}$. A sample $s_P$ is drawn randomly from $U$ and a census is taken from $s_{\mathrm{NP}}$. The probability of selection in the sample $s$, $\Pr\left(k \in s \mid \mathbf{\Omega}_P\right)$, can then be calculated for each unit $k \in U$, and the estimator (3.1) is recovered by weighting each unit $k \in s$ by the inverse of that probability. This approach was proposed by Bankier (1986) to address the problem of multiple sampling frames. In the context of integrating a probability and non-probability sample, estimator (3.1) was proposed by Kim and Tam (2020).

The last sum of (3.1) is a design-unbiased estimator of $\sum\nolimits_{k \in U} (1 - \delta_k)\, y_k = \sum\nolimits_{k \in U - s_{\mathrm{NP}}} y_k$. If a vector of auxiliary variables, $\mathbf{x}_k$, is available for $k \in s_P$ as well as the total $\mathbf{T}_{\mathbf{x}} = \sum\nolimits_{k \in U} \mathbf{x}_k$ then the weight $1 / \pi_k$ in (3.1) can be replaced with a calibrated weight $w_k$ (e.g., Deville and Särndal, 1992; Haziza and Beaumont, 2017). The calibrated weights minimize a distance function between $w_k$ and $1 / \pi_k$, $k \in s_P$, under the constraint of satisfying the calibration equation $\sum\nolimits_{k \in s_P} w_k \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}$. Ideally, the calibration is done only on the portion not covered by the non-probability sample, $U - s_{\mathrm{NP}}$; i.e., the calibration vector $(1 - \delta_k)\, \mathbf{x}_k$ is used, and the calibration equation becomes: $\sum\nolimits_{k \in s_P} w_k (1 - \delta_k)\, \mathbf{x}_k = \sum\nolimits_{k \in U - s_{\mathrm{NP}}} \mathbf{x}_k$. This is not possible when $\sum\nolimits_{k \in U - s_{\mathrm{NP}}} \mathbf{x}_k$ is unknown.

*Remark*: If assumption 2 is not appropriate, then $E\left(I_k \mid \mathbf{\Omega}_P\right) \neq E\left(I_k \mid \mathbf{Z}\right) = \pi_k$. To get around this problem, all the units for which the data were collected after selecting the sample $s_P$ can be removed from $s_{\mathrm{NP}}$. Assumption 2 is then satisfied, but a lot of available data may be omitted. To take advantage of the full set $s_{\mathrm{NP}}$, it is necessary to make a few assumptions and partially depart from the design-based approach. Assuming that $E(I_k \mid \mathbf{\Omega}_P) = \Pr(I_k = 1 \mid \delta_k, \mathbf{Y}, \mathbf{\Omega})$, we can use Bayes' theorem to show that

$$\Pr(I_k = 1 \mid \delta_k = 0, \mathbf{Y}, \mathbf{\Omega}) = \frac{1 - \Pr(\delta_k = 1 \mid I_k = 1, \mathbf{Y}, \mathbf{\Omega})}{1 - \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{\Omega})} \pi_k,$$

for the units $k \in U - s_{\mathrm{NP}}$. Therefore, estimating $E(I_k \mid \mathbf{\Omega}_P)$ requires postulating a model for $\delta_k$. Under some assumptions, $\Pr(\delta_k = 1 \mid I_k = 1, \mathbf{Y}, \mathbf{\Omega})$ can be estimated using the data from the probability sample and, for example, a logistic regression model. Estimating $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{\Omega})$ can be done using the methods described in Section 4.3 that do not rely on the validity of assumption 2, such as the method by Chen, Li and Wu (2019). These methods require that the auxiliary variables used to model this probability be available for all units of the combined sample $s = s_P \cup s_{\mathrm{NP}}$. Unlike in Section 4.3, here we can take advantage of the availability of $y_k$ for all units of both samples, and we can use the variable of interest as an auxiliary variable. Then, $\theta$ is estimated by replacing $\pi_k$ in (3.1) with an estimate of $\Pr(I_k = 1 \mid \delta_k = 0, \mathbf{Y}, \mathbf{\Omega})$. Similar approaches were proposed by Beaumont, Bocci and Hidiroglou (2014) to take into account late respondents in Statistics Canada's National Household Survey, i.e., households that responded to the initial questionnaire after the follow-up probability sample of non-respondents was drawn.

## 3.2 Calibration of the probability sample to the non-probability source

Data from non-probability sources, such as those provided by web panel respondents, can be fraught with measurement errors large enough to cast doubt on the assumption that $y_k^* \approx y_k$. Therefore, such data cannot be used to directly replace the values of the variable $y$. However, they can be used as auxiliary data to enhance the probability survey using the calibration technique. The non-probability source contains the values $y_k^*$ for $k \in s_{\mathrm{NP}}$ and potentially the values of other variables. From all these variables, it is possible to form a vector of auxiliary variables $\mathbf{x}_k^*$, available for $k \in s_{\mathrm{NP}}$, that could include an intercept. Its total is denoted as $\mathbf{T}_{\mathbf{x}^*} = \sum_{k \in s_{\mathrm{NP}}} \mathbf{x}_k^* = \sum_{k \in U} \delta_k \mathbf{x}_k^*$. Another vector of auxiliary variables, $\mathbf{x}_k$, may also be available for $k \in s_P$, as well as its total for the entire population $U$, $\mathbf{T}_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$. The calibrated weights $w_k$, $k \in s_P$, are obtained by minimizing a distance function between $w_k$ and $1/\pi_k$, $k \in s_P$, under the constraint of satisfying the calibration equation

$$\sum_{k \in s_P} w_k \begin{pmatrix} \mathbf{x}_k \\ \delta_k \mathbf{x}_k^* \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{x}} \\ \mathbf{T}_{\mathbf{x}^*} \end{pmatrix}.$$

Note that this calibration can be done only if $\mathbf{x}_k^*$ is available in the probability sample for all units $k \in s_P \cap s_{\mathrm{NP}}$. The estimator of $\theta$ is again written as $\hat{\theta} = \sum_{k \in s_P} w_k y_k$, where $w_k$ is the calibrated weight satisfying the above calibration equation. No model assumption is required for the validity of the approach, and the resulting estimates remain design-consistent regardless of the strength of the relationship between $y_k$ and the auxiliary variables $\mathbf{x}_k$ and $\mathbf{x}_k^*$. A strong relationship will help reduce the design variance of $\hat{\theta}$, $\operatorname{var}(\hat{\theta} \mid \mathbf{\Omega}_P)$. Kim and Tam (2020) discuss the use of such calibration.

Canada's Labour Force Survey (LFS) provides an example of a potential application for this calibration method. The unemployment rate, defined as the number of unemployed persons divided by the number of persons in the labour force, is a key parameter of interest that the LFS estimates. To improve the precision of the LFS estimates, a calibration variable indicating whether an individual is receiving employment insurance could be effective because there is definitely a connection between receiving employment insurance and being unemployed. The total of this calibration variable, the number of employment insurance beneficiaries, is needed for implementing this calibration and is available from an administrative source. However, applying this method would require adding a question to the LFS to identify LFS respondents who are receiving employment insurance. This information could also be obtained through a linkage between the LFS and the administrative source. It remains to be determined whether such a calibration variable could yield significant gains in the LFS.

# 4 Model-based approaches

Model-based approaches can eliminate the selection bias of the non-probability source and enable valid statistical inferences, provided that their underlying assumptions hold. The objective of the methods in Sections 4.1, 4.2 and 4.3 is to reduce respondent burden and costs by eliminating data collection for some variables of interest in a probability sample. The greater the number of variables of interest for which the values are not collected, the greater the reduction in data collection costs and respondent burden. However, these methods assume that the variables of interest are measured without error in the non-probability sample $\left(y_k^* = y_k\right)$.

From the non-probability sample $s_{\mathrm{NP}}$, we can obtain the naive estimator $\hat{\theta}^{\mathrm{NP}} = N \sum_{k \in s_{\mathrm{NP}}} y_k \,/\, n^{\mathrm{NP}}$ of the total $\theta$, where $n^{\mathrm{NP}}$ is the number of units in $s_{\mathrm{NP}}$ and $N$ is the size of the population $U$. It is well known that the selection bias of the naive estimator may be significant (see, for example, Bethlehem, 2016). The objective of the methods in Sections 4.1, 4.2 and 4.3 is to reduce the bias of the naïve estimator by using a vector of auxiliary variables, $\mathbf{x}_k$. We use $\mathbf{X}$ to denote the matrix that contains the values of vector $\mathbf{x}_k$, $k \in U$. We assume that $\mathbf{x}_k$ is measured without error in both samples $s_{\mathrm{NP}}$ and $s_P$.

Section 4.4 briefly discusses small area estimation and the area-level model of Fay and Herriot (1979). Small area estimation methods are generally used to improve the precision of estimates for population sub-groups (domains) that have a small probability sample size. They require collecting the variable $y$ in the probability sample, but not in the non-probability sample. Therefore, they do not require the condition $y_k^* = y_k$. Ideally, the non-probability sample contains variables correlated to $y$.

## 4.1 Calibration of the non-probability sample

The most natural approach to correcting the selection bias of a non-probability source is to model the relationship between the variable of interest $y_k$ and the auxiliary variables $\mathbf{x}_k$ and then predict the total $\theta$ by predicting the variable $y_k$ for each unit outside the non-probability sample. This prediction approach is described in Royall (1970) and generalized in Royall (1976); see also Elliott and Valliant

(2017). Readers are referred to Valliant, Dorfman and Royall (2000) for more details. With this approach, inferences are conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$. As a result, $\mathbf{Y}$ is considered random as well as $\boldsymbol{\Omega}$ (unless $\boldsymbol{\Omega} = \mathbf{X}$). If a probability sample is used, $\mathbf{I}$ is also considered random. It is usually assumed that the nonprobability sample selection mechanism is not informative:

*Assumption* 3: $\mathbf{Y}$ and $\boldsymbol{\delta}$ are independent after conditioning on $\mathbf{X}$.

Assumption 3 is the key to eliminating the selection bias. The more access we have to auxiliary variables that are strongly related to both $y_k$ and $\delta_k$, the more plausible assumption 3 becomes. In other words, the richer $\mathbf{X}$ is, the more the conditional independence between $\mathbf{Y}$ and $\boldsymbol{\delta}$ becomes a realistic assumption. This assumption, called the *exchangeability* assumption, is discussed in Mercer, Kreuter, Keeter and Stuart (2017). Schonlau and Couper (2017) also discuss the selection of auxiliary variables and emphasize their key role in reducing selection bias.

Often, a linear model is considered where it is assumed that the observations $y_k$ are mutually independent with $E(y_k \mid \mathbf{X}) = \mathbf{x}'_k\boldsymbol{\beta}$ and $\operatorname{var}(y_k \mid \mathbf{X}) \propto v_k$, where $\boldsymbol{\beta}$ is a vector of unknown model parameters and $v_k$ is a known function of $\mathbf{x}_k$. The best linear unbiased predictor of $\theta$ (see, for example, Valliant, Dorfman and Royall, 2000) is given by

$$\hat{\theta}^{\text{BLUP}} = \sum\nolimits_{k \in s_{\text{NP}}} y_k + \sum\nolimits_{k \in U - s_{\text{NP}}} \mathbf{x}'_k \hat{\boldsymbol{\beta}} = \mathbf{T}'_{\mathbf{x}} \hat{\boldsymbol{\beta}} + \sum\nolimits_{k \in s_{\text{NP}}} \left( y_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}} \right), \tag{4.1}$$

where

$$\hat{\boldsymbol{\beta}} = \left( \sum\nolimits_{k \in s_{\text{NP}}} v_k^{-1} \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum\nolimits_{k \in s_{\text{NP}}} v_k^{-1} \mathbf{x}_k y_k.$$

The predictor $\hat{\theta}^{\text{BLUP}}$ can also be re-written in the weighted form $\hat{\theta}^{\text{BLUP}} = \sum\nolimits_{k \in s_{\text{NP}}} w_k^C y_k$, where

$$w_k^C = 1 + v_k^{-1} \mathbf{x}'_k \left( \sum\nolimits_{k \in s_{\text{NP}}} v_k^{-1} \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \left( \mathbf{T}_{\mathbf{x}} - \sum\nolimits_{k \in s_{\text{NP}}} \mathbf{x}_k \right). \tag{4.2}$$

It can easily be shown that $w_k^C$ is a calibrated weight that satisfies the calibration equation $\sum\nolimits_{k \in s_{\text{NP}}} w_k^C \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}$. Therefore, the prediction approach is equivalent to calibration when a linear model is used to describe the relationship between $y_k$ and $\mathbf{x}_k$. The calibration equation satisfies what Mercer et al. (2017) call the *composition* assumption. This approach requires knowing the vector of control totals $\mathbf{T}_{\mathbf{x}}$. If it is unknown, an alternative is to replace it in (4.1) or (4.2) with an estimate, $\hat{\mathbf{T}}_{\mathbf{x}} = \sum\nolimits_{k \in s_P} w_k \mathbf{x}_k$, from a probability survey (Elliott and Valliant, 2017). If assumptions 1 to 3 are satisfied, it can be shown that the predictor $\hat{\theta}^{\text{BLUP}}$ is unbiased, i.e., $E\left( \hat{\theta}^{\text{BLUP}} - \theta \,\middle|\, \boldsymbol{\delta}, \mathbf{X} \right) = 0$, whether $\mathbf{T}_{\mathbf{x}}$ or $\hat{\mathbf{T}}_{\mathbf{x}}$ is used, provided that the latter is design-unbiased, i.e., $E\left( \hat{\mathbf{T}}_{\mathbf{x}} \,\middle|\, \boldsymbol{\Omega}_P \right) = \mathbf{T}_{\mathbf{x}}$. Of course, the unbiasedness property of the predictor $\hat{\theta}^{\text{BLUP}}$ requires the linear model to be valid.

*Remark*: In practice, auxiliary variables for which the population total is known are usually few in number and not sufficiently predictive of the variable $y$ for eliminating the selection bias. These may be supplemented with other auxiliary variables for which the total can be estimated using an existing probability survey. Therefore, the vector of population totals may be a blend of known and estimated totals. If the probability survey itself is calibrated to known population totals, then only the estimated totals $\hat{\mathbf{T}}_{\mathbf{x}}$ from the probability survey can be used.

A linear model is not always appropriate. This is the case when the variable $y$ is categorical. Another typical example occurs when it is desired to estimate the total of a quantitative variable in a domain of interest. The variable $y$ is then defined as the product of that quantitative variable and a binary variable indicating domain membership. To model such a variable, it is natural to consider a mixture of a degenerate distribution at 0 and a continuous distribution. When the relationship between $y_k$ and $\mathbf{x}_k$ is not linear, model-assisted calibration of Wu and Sitter (2001) can be used to preserve the weighted form of the predictor $\theta$ while taking into account the non-linearity of the relationship. Suppose that we replace the above linear model with a non-linear (or non-parametric) model such that $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k)$, where $h(\cdot)$ is some function. The Wu and Sitter (2001) calibration first involves predicting $y_k$ by $\hat{y}_k = \hat{h}(\mathbf{x}_k),\, k \in U,$ where $\hat{h}(\mathbf{x}_k)$ is a model-based estimate of $h(\mathbf{x}_k)$. Then, the total $T_{\hat{y}} = \sum_{k \in U} \hat{y}_k$ is calculated, and weights, $w_k^{\text{MC}},\, k \in s_{\text{NP}}$, are found that satisfy the calibration equation:

$$\sum_{k \in s_{\text{NP}}} w_k^{\text{MC}} \begin{pmatrix} 1 \\ \hat{y}_k \end{pmatrix} = \begin{pmatrix} N \\ T_{\hat{y}} \end{pmatrix}.$$

In other words, the equation (4.2) can be used, where $\mathbf{x}_k'$ is replaced with $(1,\, \hat{y}_k)$. This method requires knowing the population size $N$ as well as the vector $\mathbf{x}_k$ for all units in the population $U$. If $N$ and $T_{\hat{y}}$ are unknown, they can be replaced with estimates from a probability survey. For example, we can replace $N$ with $\hat{N} = \sum_{k \in s_P} w_k$ and $T_{\hat{y}}$ with $\hat{T}_{\hat{y}} = \sum_{k \in s_P} w_k \hat{y}_k$. The approach can also be extended to the case of multiple variables of interest.

We mentioned that the selection bias may be considerably reduced if $\mathbf{x}_k$ is rich and contains variables that are related to both $\delta_k$ and $y_k$, which makes assumption 3 more realistic. It can therefore be useful in practice to consider a large number of potential auxiliary variables and select the most relevant ones using a variable selection technique. Chen, Valliant and Elliott (2018) suggest the LASSO technique for selecting auxiliary variables and show its good properties.

It should be noted that the predictor $\hat{\theta}^{\text{BLUP}}$ reduces to the naive estimator, $\hat{\theta}^{\text{NP}}$, in the simplest case possible where only one constant auxiliary variable is used: $x_k = 1,\, k \in U$. The naive estimator is usually highly biased. Its bias can be significantly reduced if the population $U$ can be subdivided into $H$ disjoint and exhaustive post-strata, $U_h,\, h = 1, \ldots, H,$ of size $N_h$. The post-stratification model, $E(y_k \mid \mathbf{X}) = \beta_h,\, k \in U_h$, is then postulated, which is an important special case of the above linear model. Assuming that the variance $\operatorname{var}(y_k \mid \mathbf{X})$ is constant for $k \in U_h$, the predictor $\hat{\theta}^{\text{BLUP}}$ is written: $\hat{\theta}^{\text{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h$, where $\hat{\beta}_h = \sum_{k \in s_{\text{NP},h}} y_k / n_h^{\text{NP}}$, $s_{\text{NP},h}$ is the set of units in $U_h$ that are part of the sample $s_{\text{NP}}$ and $n_h^{\text{NP}}$ is the size of $s_{\text{NP},h}$. If the population sizes $N_h$ are unknown, they can be replaced with estimates, $\hat{N}_h = \sum_{k \in s_{P,h}} w_k$, from a probability survey, where $s_{P,h}$ is the set of units in $U_h$ that are part of the sample $s_P$. Regression trees could prove to be an interesting approach for forming post-strata, especially when the auxiliary variables are categorical.

If multiple categorical auxiliary variables are available, it can be useful to form a large number of post-strata to reduce the selection bias. If many auxiliary variables are crossed, the sample sizes $n_h^{\text{NP}}$ could become very small, thereby making the estimators $\hat{\beta}_h$ very unstable. Gelman and Little (1997) suggest using a multi-level regression model to obtain estimators $\tilde{\beta}_h$ more stable than $\hat{\beta}_h$. They then consider the post-stratified predictor: $\hat{\theta}^{\text{MRP}} = \sum_{h=1}^{H} N_h \tilde{\beta}_h$. Nowadays, this method is known as Mr.P or MRP (Multilevel Regression and Poststratification); see, for example, Mercer et al. (2017). A similar approach would use small area estimation methods (Rao and Molina, 2015) to stabilize the estimators $\hat{\beta}_h$. Although such methods are likely to produce much more precise estimates of the average of variable $y$ over the population $U_h$, it remains to be determined whether such methods can produce significant efficiency gains for estimating the overall total $\theta$ compared to the simple post-stratified predictor $\hat{\theta}^{\text{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h$. It seems that regression trees provide another way to control the instability of the estimators $\hat{\beta}_h$ since a criterion is generally used to prevent an overly narrow subdivision of the population. These various methods warrant further investigation in future research. Precise estimation of population sizes $N_h$, if not known, is also a problem not to be overlooked when the population is divided into a large number of post-strata.

## 4.2 Statistical matching

Statistical matching, or data fusion, is an approach developed for combining data from two different sources that contain both source-specific variables and common variables. Readers are referred to D'Orazio, Di Zio and Scanu (2006) or Rässler (2012) for a review of statistical matching methods. In the context of this article, statistical matching involves modelling the relationship between $y_k$ and the auxiliary variables $\mathbf{x}_k$, which are common to both sources, using data from the non-probability sample. As with calibration, the non-probability sample selection mechanism is assumed to be non-informative, and the auxiliary variables must be chosen carefully in order to make assumption 3 as plausible as possible. Once a model has been determined, it is used to predict the $y$ values in a probability sample. Statistical matching can be viewed as an imputation problem with an imputation rate of 100%. The predictor of $\theta$, obtained from the probability sample, takes the form: $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$, where $y_k^{\text{imp}}$ is the imputed value for the unit $k \in s_P$. As in calibration, inferences are conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$. Assumption 3, in a statistical matching context, can be viewed as analogous to the Population Missing At Random (PMAR) assumption introduced by Berg, Kim and Skinner (2016) in a non-response context.

If the linear regression model $E(y_k \mid \mathbf{X}) = \mathbf{x}_k' \boldsymbol{\beta}$ is used, the imputed value for the unit $k \in s_P$ is $y_k^{\text{imp}} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ and the resulting predictor is given by $\hat{\theta}^{\text{SM}} = \hat{\mathbf{T}}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}$. If assumptions 1 to 3 are satisfied and $E(\hat{\mathbf{T}}_{\mathbf{x}} \mid \boldsymbol{\Omega}_P) = \mathbf{T}_{\mathbf{x}}$, statistical matching produces an unbiased predictor, $\hat{\theta}^{\text{SM}}$, i.e., $E(\hat{\theta}^{\text{SM}} - \theta \mid \boldsymbol{\delta}, \mathbf{X}) =$ 0. Also, if $v_k = \mathbf{x}_k' \boldsymbol{\lambda}$, for a certain known vector $\boldsymbol{\lambda}$, it can be shown that $\sum_{k \in s_{\text{NP}}} (y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}) = 0$, and the predictor $\hat{\theta}^{\text{SM}}$ is equivalent to the predictor $\hat{\theta}^{\text{BLUP}}$ if we replace $\mathbf{T}_{\mathbf{x}}$ in (4.1) with $\hat{\mathbf{T}}_{\mathbf{x}}$. It can also be shown that, for a post-stratification model where we impute $y_k, k \in s_{P,h},$ with $y_k^{\text{imp}} = \hat{\beta}_h,$ the predictor $\hat{\theta}^{\text{SM}}$ reduces to $\hat{\theta}^{\text{SM}} = \sum_{h=1}^{H} \hat{N}_h \hat{\beta}_h.$ Therefore, statistical matching and calibration produce similar predictors, even identical in some cases, when a linear model is postulated and the totals $\mathbf{T}_\mathbf{x}$ are estimated.

Choosing between statistical matching or calibration can depend on the user's perspective. For example, if it is the content of the non-probability source, in terms of variables of interest, that is relevant to the user, then it seems natural to weight the non-probability sample in the hopes of reducing the selection bias for all variables of interest. The calibration technique or the methods in Section 4.3 are obvious choices for such weighting. Conversely, if instead it is the content of the probability survey that is relevant, then statistical matching is the appropriate choice. This method enriches the probability survey by imputing the missing variables of interest.

Statistical matching is easily generalized to non-linear or non-parametric models such that $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k).$ The imputed values $y_k^{\text{imp}}$ are simply obtained by predicting the missing values $y_k,$ $k \in s_P,$ using the chosen model. The predictor $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$ remains unbiased if assumptions 1 to 3 are satisfied and if $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0.$ Donor or nearest neighbour imputation is a non-parametric imputation method commonly used for handling non-response (see, for example, Beaumont and Bocci, 2009) that does not require a linear relationship between $y_k$ and $\mathbf{x}_k.$ In the context of matching non-probability and probability samples, donor imputation was popularized by Rivers (2007). For a given unit $k \in s_P,$ the method involves finding the nearest donor, with respect to the auxiliary variables $\mathbf{x}$, among the units of the non-probability sample and replacing the missing value $y_k$ with the $y$ value from this donor. For donor imputation, the condition $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0$ is satisfied if, for each recipient $k \in s_P,$ the donor has exactly the same values of $\mathbf{x}$ as the recipient. When one or more auxiliary variables are continuous, this condition is satisfied only asymptotically in general. A very large non-probability sample provides a large pool of donors, which should help to approximately satisfy this condition.

*Remark*: In some applications, a very large non-probability panel of volunteers, $s_{\text{NP}},$ is available, which contains a few auxiliary variables for matching, $\mathbf{x}$, but no variable of interest. Ideally, the variables of interest would be collected for all units of the panel $s_{\text{NP}},$ but that is impossible due to the cost and the burden on the panel members. Therefore, in practice, a sub-sample $s_{\text{NP}}^*$ of $s_{\text{NP}}$ is selected using random or non-random sampling methods. Quota sampling (e.g., Deville, 1991) is often considered in this context. In addition to collecting the variables of interest for all units of $s_{\text{NP}}^*,$ there may also be interest in collecting other auxiliary variables for matching in order to enhance the vector $\mathbf{x}$. The matching can then be done to the probability sample, often much smaller in size, as long as the latter contains the same auxiliary variables as those of the non-probability sub-sample $s_{\text{NP}}^*.$ By carefully choosing the auxiliary variables for the matching, the potential for bias reduction is increased (Schonlau and Cooper, 2017). The implementation proposed by Rivers (2007) is slightly different. Rivers (2007) suggests conducting the matching between the probability sample and the panel $s_{\text{NP}}$ using the auxiliary variables available in both sources. The variables of interest are collected only for the set of donors in $s_{\mathrm{NP}}$ who have been matched to a unit in the probability sample, which allows for a significant reduction of data collection costs and burden. The implicit assumption is that the panel members, initially volunteers, are more likely to respond than individuals chosen at random in the population. Obviously, non-response is unavoidable, and this problem must be dealt with, potentially through imputation. The advantage of this method is that the matching is carried out using the panel $s_{\mathrm{NP}}$ rather than a sub-sample of this panel; the pool of donors is larger. However, the matching cannot be done using the enhanced vector of auxiliary variables because it is not available for the units in $s_{\mathrm{NP}}$, which limits the potential for bias reduction.

Lavallée and Brisbane (2016) point out the connection between statistical matching and indirect sampling (Lavallée, 2007; Deville and Lavallée, 2006). They propose an estimator obtained by imputing each missing value $y_k, k \in s_P$, by a weighted average of the $y$ values of nearest donors. In reality, their estimator can also be obtained equivalently by imputing the missing values using fractional donor imputation (for example, Kim and Fuller, 2004). The use of more than one donor to impute the missing values yields a typically modest variance reduction.

Several imputation methods used in practice can be considered linear (Beaumont and Bissonnette, 2011). This is the case for linear regression imputation, donor imputation and fractional donor imputation. An imputation method is said to be linear if the imputed value $y_k^{\mathrm{imp}}, k \in s_P$, can be written as $y_k^{\mathrm{imp}} = \sum_{l \in s_{\mathrm{NP}}} \omega_{kl} y_l$, where $\omega_{kl}$ is a function of $\boldsymbol{\delta}$ or $\mathbf{X}$ but not of $\mathbf{Y}$. For example, for donor or nearest-neighbour imputation, $\omega_{kl} = 1$ if the unit $l \in s_{\mathrm{NP}}$ is the donor for the recipient $k \in s_P$; otherwise $\omega_{kl} = 0$. For a linear imputation method, the estimator $\hat{\theta}^{\mathrm{SM}} = \sum_{k \in s_P} w_k y_k^{\mathrm{imp}}$ can be rewritten as a weighted sum over the non-probability sample: $\hat{\theta}^{\mathrm{SM}} = \sum_{l \in s_{\mathrm{NP}}} W_l y_l$, where $W_l = \sum_{k \in s_P} w_k \omega_{kl}$. Therefore, for linear imputation methods, statistical matching is an alternative to calibration and to the methods in Section 4.3 if the objective is to properly weight the non-probability sample.

So far, we have considered only the estimation of the total $\theta = \sum_{k \in U} y_k$. However, the probability sample contains other variables, and there may be interest in the relationship between two or more variables, some from the probability survey and others imputed from the non-probability sample. As an example, suppose that the estimation of the total $\theta = \sum_{k \in U} \tilde{y}_k y_k$ is of interest, where $\tilde{y}_k$ is a variable collected in the probability survey, but not available in the non-probability sample. It could, for example, define membership in a domain of interest. Statistical matching can be used to estimate this parameter by $\hat{\theta}^{\mathrm{SM}} = \sum_{k \in s_P} w_k \tilde{y}_k y_k^{\mathrm{imp}}$. We use $\tilde{\mathbf{Y}}$ to denote the vector that contains the values of the variable $\tilde{y}_k$, $k \in U$. It can be shown that $\hat{\theta}^{\mathrm{SM}}$ is unbiased, $E\left(\hat{\theta}^{\mathrm{SM}} - \theta \,\middle|\, \boldsymbol{\delta}, \mathbf{X}, \tilde{\mathbf{Y}}\right) = 0$, if assumptions 1 to 3 are satisfied in addition to the following assumption:

*Assumption* 4: $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ are independent after conditioning on $\boldsymbol{\delta}$ and $\mathbf{X}$.

Assumption 4 is known as the conditional independence assumption in the statistical matching literature.

## 4.3 Inverse propensity score weighting

Instead of modelling the relationship between $y_k$ and $\mathbf{x}_k$, the relationship between $\delta_k$ and $\mathbf{x}_k$ could be modelled. The main advantage of this approach is to simplify the modelling effort when there are multiple variables of interest since there is always only one variable $\delta_k$. With this approach, inferences are conditional on $\mathbf{Y}$ and $\mathbf{X}$. Also, it is usually assumed that assumption 3 is valid and thus $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) = \Pr(\delta_k = 1 \mid \mathbf{X})$. The probability of participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{X})$ is then estimated by $\hat{p}_k$, and the estimate $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$ is calculated, where $w_k^{\text{PS}} = 1 / \hat{p}_k$. The assumption that $p_k > 0, k \in U$, must be made. It is called the *positivity* assumption by Mercer et al. (2017). It may also be required in the calibration and statistical matching approaches. For example, empty post-strata $(n_h^{\text{NP}} = 0)$ may occur if it is not satisfied. To fix this issue, these empty post-strata are usually collapsed with other non-empty post-strata. This collapsing may jeopardize the validity of assumption 3 if the collapsed post-strata are different.

The estimation of $p_k$ can be achieved by postulating a parametric model $p_k = g(\mathbf{x}_k;\ \boldsymbol{\alpha})$, where $g$ is some function, normally bounded by 0 and 1, and $\boldsymbol{\alpha}$ is a vector of unknown model parameters. The logistic function $g(\mathbf{x}_k;\ \boldsymbol{\alpha}) = \exp\left(\mathbf{x}_k'\boldsymbol{\alpha}\right) \Big/ \left[1 + \exp\left(\mathbf{x}_k'\boldsymbol{\alpha}\right)\right]$ predominates in the applications (see Kott, 2019, for a recent application). The estimator of $\boldsymbol{\alpha}$ is denoted by $\hat{\boldsymbol{\alpha}}$ and the estimated probability by $\hat{p}_k = g(\mathbf{x}_k;\ \hat{\boldsymbol{\alpha}})$. Ideally, $\boldsymbol{\alpha}$ would be estimated using $\mathbf{x}_k$ for all the units in the population $U$ similar to what would be done in a non-response context. For example, assuming the logistic function is used, $\boldsymbol{\alpha}$ could be estimated by solving the maximum likelihood equation:

$$\sum\nolimits_{k \in U} [\delta_k - p_k(\boldsymbol{\alpha})]\, \mathbf{x}_k = \sum\nolimits_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum\nolimits_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}. \tag{4.3}$$

This is impossible when $\mathbf{x}_k$ is not known for all units $k \in U - s_{\text{NP}}$, which is almost always the case in practice. Iannacchione, Milne and Folsom (1991) proposed another unbiased estimation equation for $\boldsymbol{\alpha}$ (see also Deville and Dupont, 1993):

$$\sum\nolimits_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum\nolimits_{k \in U} \mathbf{x}_k = \mathbf{0}. \tag{4.4}$$

The main advantage of equation (4.4) is that it does not require knowing $\mathbf{x}_k$ for each unit $k \in U - s_{\text{NP}}$. However, it is necessary to have access to the vector of totals $\sum_{k \in U} \mathbf{x}_k$ from an external source. An interesting property of equation (4.4) is that the resulting weights $w_k^{\text{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ satisfy the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$, just like the weights $w_k^C$ given in (4.2). Indeed, it can be shown that solving (4.4) yields $w_k^{\text{PS}} = w_k^C$ if the model $p_k(\boldsymbol{\alpha}) = \left(1 + v_k^{-1}\mathbf{x}_k'\boldsymbol{\alpha}\right)^{-1}$ is used. However, this is a less natural model than the above logistic model for modelling a probability.

To get around the problem of missing values $\mathbf{x}_k, k \in U - s_{\text{NP}}$, Chen et al. (2019) suggest estimating $\sum_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k$ in (4.3) using a probability survey. The equation to be solved becomes:

$$\sum\nolimits_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum\nolimits_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}. \tag{4.5}$$

Equation (4.5) is unbiased conditionally on $\mathbf{Y}$ and $\mathbf{X}$ provided that the probability survey allows for unbiased estimation, conditionally on $\mathbf{Y}$ and $\mathbf{\Omega}$, of any population total that is not a function of $\boldsymbol{\delta}$ such as $\sum_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k$. Assumptions 1 and 3 are required, but not assumption 2. Using the idea of Iannacchione et al. (1991), an alternative to (4.5) is obtained by solving:

$$\sum_{k \in s_{\mathrm{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{0}. \tag{4.6}$$

Equation (4.6) produces weights $w_k^{\mathrm{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ that satisfy the calibration equation $\sum_{k \in s_{\mathrm{NP}}} w_k^{\mathrm{PS}} \mathbf{x}_k = \sum_{k \in s_P} w_k \mathbf{x}_k$ (see also Lesage, 2017; Rao, 2020). The estimators of $\boldsymbol{\alpha}$ obtained using (4.5) or (4.6) are likely less efficient than those obtained using (4.3) or (4.4). If $\mathbf{x}_k, k \in U - s_{\mathrm{NP}}$, or the vector $\sum_{k \in U} \mathbf{x}_k$ is known, then using (4.3) or (4.4) is preferable. Otherwise, the estimating equations (4.5) or (4.6) can be used provided that $\mathbf{x}_k$ is collected in a probability survey. Note that the indicators $\delta_k$ do not need to be observed in the probability sample.

Equations (4.5) and (4.6) may be more difficult to solve than equations (4.3) and (4.4) and may not have a solution. Consider, for example, the case where there is only one auxiliary variable: $x_k = 1$. Using (4.5) or (4.6), it can be seen that the estimated probability reduces to: $\hat{p}_k = n^{\mathrm{NP}} \big/ \sum_{k \in s_P} w_k$. If the size of the probability sample is sufficiently large, it is expected that $0 < \hat{p}_k < 1$. For small sample sizes, it may happen that $\hat{p}_k > 1$ due to the variability of $\sum_{k \in s_P} w_k$. In that case, equations (4.5) and (4.6) would not have a solution if the logistic function is used since it requires that $0 < \hat{p}_k < 1$. To avoid this issue, it may be helpful to consider other functions not bounded by 1, such as $g(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp(\mathbf{x}_k' \boldsymbol{\alpha})$.

Kim and Wang (2019) suggest using the probability sample to estimate the participation probability. Assuming the logistic function is used, the equation to be solved is:

$$\sum_{k \in s_P} w_k [\delta_k - p_k(\boldsymbol{\alpha})]\, \mathbf{x}_k = \sum_{k \in s_P} w_k \delta_k \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}.$$

The method requires knowing the indicators $\delta_k$ in the probability sample and the validity of assumptions 1, 2 and 3 to ensure the estimating equation is unbiased. Also, the probability sample size is usually small relative to the non-probability sample size, and it can be numerically difficult to estimate $\boldsymbol{\alpha}$, especially when $\mathbf{x}_k$ contains a large number of variables and the overlap between the two samples is small.

Lee (2006), see also Rivers (2007), Valliant and Dever (2011) and Elliott and Valliant (2017), proposes to combine the two samples and then estimate $p_k$ using logistic regression. It seems that the author implicitly assumes that the two samples do not overlap, i.e., that $\delta_k = 0$ for all units in $s_P$. Using again the logistic function, the resulting estimating equation is:

$$\sum_{k \in s_{\mathrm{NP}}} \eta_k^{\mathrm{NP}} [1 - p_k(\boldsymbol{\alpha})]\, \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}, \tag{4.7}$$

where $\eta_k^{\mathrm{NP}}$ is a certain weight for the units in the non-probability sample. The method is somewhat similar to the one proposed by Chen et al. (2019), but the estimating equation (4.7) is not unbiased, conditionally on $\mathbf{Y}$ and $\mathbf{X}$, unlike equations (4.5) and (4.6). However, if we assume $\eta_k^{\mathrm{NP}} = 1$ and if $\max\{p_k;\ k \in U\}$ is small, equation (4.7) becomes approximately equivalent to equation (4.5). Yet Lee (2006) does not directly use the estimated probabilities resulting from (4.7). The author uses them only to order the union of the two samples and then create homogeneous classes. Using homogeneous classes brings some robustness to model misspecification and can help prevent very small estimated probabilities and thus very large weights. In the context of non-response, forming homogeneous imputation or reweighting classes was studied by Little (1986), Eltinge and Yansaneh (1997), and Haziza and Beaumont (2007), among others. Haziza and Lesage (2016) illustrate the robustness of the method when the function $g(\mathbf{x}_k;\ \boldsymbol{\alpha})$ is misspecified. The method is used regularly in Statistics Canada surveys for dealing with non-response.

Rather than using (4.7), homogeneous classes could be formed by starting with the unbiased equations (4.5) or (4.6). These initial estimated probabilities are denoted by $\hat{p}_k^0 = g(\mathbf{x}_k;\ \hat{\boldsymbol{\alpha}})$. The sample $s = s_P \cup s_{\mathrm{NP}}$ can then be sorted by $\hat{p}_k^0$ and divided into $C$ homogeneous classes of equal or unequal sizes. The set of units in $s_P$ that are part of class $c$ is denoted by $s_{P,\,c}$ whereas the set of units in $s_{\mathrm{NP}}$ that are part of class $c$ is denoted by $s_{\mathrm{NP},c}$. The weight $w_k^{\mathrm{PS}}$ for a unit $k \in s_{\mathrm{NP},c}$ is equal to the inverse of the estimated participation rate in class $c$ and is given by $w_k^{\mathrm{PS}} = \hat{N}_c \,/\, n_c^{\mathrm{NP}}$, where $\hat{N}_c = \sum_{k \in s_{P,c}} w_k$ and $n_c^{\mathrm{NP}}$ is the number of units in $s_{\mathrm{NP},c}$ This weight ensures the calibration property: $\sum_{k \in s_{\mathrm{NP},c}} w_k^{\mathrm{PS}} = \hat{N}_c$. The number of classes must be large enough to capture a high percentage of the variability of the initial probabilities $\hat{p}_k^0$, thereby reducing the bias. On the other hand, it must not be too large to prevent the occurrence of empty classes since the weights $w_k^{\mathrm{PS}} = \hat{N}_c \,/\, n_c^{\mathrm{NP}}$ cannot be calculated if $n_c^{\mathrm{NP}} = 0$. Regression trees can prove to be an effective alternative for forming classes. In a non-response context, they have been studied by Phipps and Toth (2012). The estimator $\hat{\theta}^{\mathrm{PS}} = \sum_{k \in s_{\mathrm{NP}}} w_k^{\mathrm{PS}} y_k$ obtained after forming homogeneous classes has exactly the same form as the post-stratified estimator described in the calibration approach in Section 4.1; the only difference is that the classes are built by modelling $\delta_k$ rather than $y_k$.

Assumption 3 may not be realistic in some contexts so that $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) \neq \Pr(\delta_k = 1 \mid \mathbf{X})$. In this case, the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X})$ might be modelled using a vector of explanatory variables $\mathbf{x}_k^*$, defined using the variable of interest $y_k$ (or variables of interest if there are several) and potentially other auxiliary variables $\mathbf{x}_k$. A parametric model, $p_k = g(\mathbf{x}_k^*;\ \boldsymbol{\alpha})$, can be considered for modelling the participation probability. Equations (4.5) and (4.6) cannot be used to estimate $\boldsymbol{\alpha}$ because $y_k$ (and therefore $\mathbf{x}_k^*$) is not available in the probability sample. However, an equation similar to (4.6) can be used:

$$\sum_{k \in s_{\mathrm{NP}}} \frac{\mathbf{x}_k^I}{g(\mathbf{x}_k^*;\ \boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k^I = \mathbf{0}. \tag{4.8}$$

The vector $\mathbf{x}_k^I$, of the same size as $\boldsymbol{\alpha}$, contains calibration variables, also called instrumental variables in the econometric literature. We use $\mathbf{X}^I$ to denote the matrix that contains the values of vector $\mathbf{x}_k^I$, $k \in U$. Equation (4.8) requires knowing the calibration variables $\mathbf{x}_k^I$ for both samples. However, the explanatory variables $\mathbf{x}_k^*$ can be observed only for the units in the non-probability sample. Equation (4.8) produces weights $w_k^{\text{PS}} = 1 / g(\mathbf{x}_k^*;\ \hat{\boldsymbol{\alpha}})$ that satisfy the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k^I = \sum_{k \in s_P} w_k \mathbf{x}_k^I$. An equation similar to (4.8) was originally proposed by Deville (1998) to deal with nonresponse (see also Kott, 2006; Haziza and Beaumont, 2017). Equation (4.8) is unbiased, conditionally on $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{X}^I$, if the instrumental variables $\mathbf{x}_k^I$ can be selected such that the following assumption is satisfied:

*Assumption* 5: $\boldsymbol{\delta}$ and $\mathbf{X}^I$ are independent after conditioning on $\mathbf{Y}$ and $\mathbf{X}$.

Assumption 3 is no longer required, but is replaced with another assumption. The choice of instrumental variables $\mathbf{x}_k^I$ that satisfy assumption 5 is not always obvious in practice. They must not be predictive of $\delta_k$ after conditioning on $\mathbf{x}_k^*$. Ideally, for efficiency reasons, the instrumental variables are selected so as to be predictive of $\mathbf{x}_k^*$ without compromising assumption 5. Unlike equations (4.5) and (4.6), equation (4.8) cannot be used to form homogeneous classes because the participation probabilities $\hat{p}_k = g(\mathbf{x}_k^*;\ \hat{\boldsymbol{\alpha}})$ cannot be calculated for the units in the probability sample. As such, the property of robustness that comes with homogeneous classes is lost. Because of these drawbacks, equation (4.8) should be considered only when there are strong reasons to believe that assumption 3 is not appropriate.

Once weights $w_k^{\text{PS}}$ have been calculated using one of the methods in this section, they can still be adjusted through calibration. The objective of this calibration is to improve the precision of the estimator $\hat{\theta}^{\text{PS}}$ and also obtain a double robustness property (see Chen et al., 2019).

In general, the variable $y$ is observed for the entire non-probability sample, and the inverse propensity-score weighted estimator, $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$, or a weighted estimator obtained by calibration or statistical matching can be used. Sometimes, the non-probability sample is too large and the variable $y$ can only be collected for a sub-sample of $s_{\text{NP}}$. Quota sampling (e.g., Deville, 1991) is a commonly used method for drawing the sub-sample if auxiliary variables are available for $k \in s_{\text{NP}}$. An alternative to quota sampling is to calculate the weights $w_k^{\text{PS}}$ for the entire non-probability sample and use them to select a random sub-sample with probabilities proportional to the weights. The variable $y$ is then collected only for the sub-sample, and the estimates are obtained as if the sub-sample was drawn from the population using an equal probability design. This approach is called inverse sampling in the literature on probability surveys (see, for example, Hinkins, Oh and Scheuren, 1997; or Rao, Scott and Benhin, 2003) and was proposed by Kim and Wang (2019) for non-probability samples.

## 4.4 Small area estimation

In most surveys, it is desired to estimate the total of the variable $y$ not just for the entire population $U$, but also for different subgroups of the population, called domains. Probability surveys conducted by national statistical agencies generally produce reliable estimates for domains with a sufficient number of sample units. Their bias is controlled through the various sampling and data collection procedures, and their variance is typically small enough to draw accurate conclusions. When the domain of interest contains few sample units, the survey estimates may become unstable to the point of being unusable even when their bias stays under control. To remedy a lack of data in a domain of interest, small area estimation methods may be considered. These methods offset the lack of observed data in a domain through model assumptions that link auxiliary data to survey data. Two types of models are commonly used: unit-level models and area-level models. The area-level model of Fay and Herriot (1979) is undoubtedly the most popular. It requires auxiliary data to be available at the domain level only, unlike unit-level models, which require auxiliary variables for each unit of the population $U$. Readers are referred to Rao and Molina (2015) for an excellent coverage of the various approaches. Below, we focus on the Fay-Herriot model.

Suppose it is desired to estimate $D$ totals, $\theta_d = \sum_{k \in U_d} y_k$, $d = 1, \ldots, D$, where $U_d$ are $D$ disjoint subsets of the population. Using a probability survey, $\theta_d$ can be estimated by $\hat{\theta}_d = \sum_{k \in s_{P,d}} w_k y_k$, where $s_{P,d}$ is the set of sample units that fall within domain $d$. The estimator $\hat{\theta}_d$ is called the direct estimator of $\theta_d$ because it only uses $y$ values of units belonging to domain $d$. Small area estimation techniques generally lead to indirect estimators that combine the sample $y$ values of domain $d$ with $y$ values of units outside domain $d$. We assume that a vector of auxiliary variables is available at the area level, and these variables come from sources independent of the probability sample. This vector for domain $d$ is denoted by $\mathbf{x}_d$. For example, the vector $\mathbf{x}'_d = \left(N_d, N_d \hat{\mu}_d^{\mathrm{NP}}\right)$ could be considered, where $N_d$ is the population size in domain $d$, $\hat{\mu}_d^{\mathrm{NP}} = \sum_{k \in s_{\mathrm{NP},d}} y_k^* / n_d^{\mathrm{NP}}$ is the average of variable $y^*$ in a non-probability sample, $s_{\mathrm{NP},d}$ is the set of units in the non-probability sample that are in domain $d$ and $n_d^{\mathrm{NP}}$ is the size of the non-probability sample in domain $d$. If the population size $N_d$ is unknown, it can be replaced with an estimate independent of the probability survey. We use $\mathbf{X}$ to denote the matrix that contains the values of vector $\mathbf{x}_d$, $d = 1, \ldots, D$. Note that the vector $\boldsymbol{\delta}$ is hidden in the matrix $\mathbf{X}$ in this section.

The Fay-Herriot model has two components: the sampling model and the linking model. The sampling model is based on the assumption that, conditionally on $\boldsymbol{\Omega}_P$, the direct estimators $\hat{\theta}_d$ are independent and unbiased, i.e., $E\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right) = \theta_d$. Their design variance is denoted by $\psi_d = \operatorname{var}\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right)$. The sampling model is usually written in the form:

$$\hat{\theta}_d = \theta_d + e_d, \tag{4.9}$$

where $e_d$ is the sampling error such that $E\left(e_d \mid \boldsymbol{\Omega}_P\right) = 0$ and $\operatorname{var}\left(e_d \mid \boldsymbol{\Omega}_P\right) = \psi_d$. The independence assumption of the estimators $\hat{\theta}_d$ (and therefore of the sampling errors $e_d$) can be questioned when the strata do not coincide with the domains of interest. Section 8.2 of Rao and Molina (2015) discusses methods that take into account correlated sampling errors. In practice, it is often assumed that these correlations are weak, and they are ignored.

The linking model assumes that, conditionally on $\mathbf{X}$, the totals $\theta_d$ are independent, $E(\theta_d \mid \mathbf{X}) = \mathbf{x}'_d\boldsymbol{\beta}$ and $\operatorname{var}(\theta_d \mid \mathbf{X}) = b_d^2\sigma_v^2$, where $b_d$ are known constants used for controlling heteroscedasticity and $\boldsymbol{\beta}$ and $\sigma_v^2$ are unknown model parameters. The linking model is usually written in the form:

$$\theta_d = \mathbf{x}'_d\boldsymbol{\beta} + b_d v_d, \tag{4.10}$$

where $v_d$ is the model error such that $E(v_d \mid \mathbf{X}) = 0$ and $\operatorname{var}(v_d \mid \mathbf{X}) = \sigma_v^2$. When the parameters of interest, $\theta_d$, are totals, it is often appropriate to let $b_d = N_d$. From (4.9) and (4.10), we obtain the combined model:

$$\hat{\theta}_d = \mathbf{x}'_d\boldsymbol{\beta} + a_d, \tag{4.11}$$

where $a_d = b_d v_d + e_d$ is the combined error. When using the Fay-Herriot model (4.11), inferences are usually made conditionally on $\mathbf{X}$. It can easily be shown that $E(a_d \mid \mathbf{X}) = 0$ and $\operatorname{var}(a_d \mid \mathbf{X}) = b_d^2\sigma_v^2 + \tilde{\psi}_d$, where $\tilde{\psi}_d = E(\psi_d \mid \mathbf{X})$ is called the smooth design variance (Beaumont and Bocci, 2016; and Hidiroglou, Beaumont and Yung, 2019).

Now suppose that it is desired to predict the total $\theta_d$ using a linear predictor $\hat{\theta}_d^{\text{LIN}} = \sum_{i=1}^{D} \lambda_{di}\hat{\theta}_i$, where $\lambda_{di}$ are constants to be determined. A linear predictor uses all the data from the probability sample for predicting $\theta_d$, not just the data from domain $d$. This explains how it derives its efficiency. However, not all linear predictors are appropriate for predicting $\theta_d$. A strategy often used for determining the constants $\lambda_{di}$ is to minimize the variance of the prediction error, $\operatorname{var}\left(\hat{\theta}_d^{\text{LIN}} - \theta_d \,\middle|\, \mathbf{X}\right)$, subject to the constraint that the predictor must be unbiased, $E\left(\hat{\theta}_d^{\text{LIN}} - \theta_d \,\middle|\, \mathbf{X}\right) = 0$. The resulting predictor, called the Best Linear Unbiased Predictor (BLUP), is denoted by $\hat{\theta}_d^{\text{BLUP}}$, and can be written in the form (see, for example, Rao and Molina, 2015):

$$\hat{\theta}_d^{\text{BLUP}} = \gamma_d\hat{\theta}_d + (1 - \gamma_d)\,\mathbf{x}'_d\hat{\boldsymbol{\beta}}, \tag{4.12}$$

where $\gamma_d = b_d^2\sigma_v^2 \,/\, \left(b_d^2\sigma_v^2 + \tilde{\psi}_d\right)$ is bounded by 0 and 1, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d\mathbf{x}'_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\,\hat{\theta}_d.$$

The predictor (4.12) is a weighted average of the direct estimator $\hat{\theta}_d$ and a prediction, $\mathbf{x}'_d\hat{\boldsymbol{\beta}}$, often called the synthetic estimator. More weight is given to the direct estimator when the smooth design variance, $\tilde{\psi}_d$, is small relative to the variance of the linking model, $b_d^2\sigma_v^2$. The predictor $\hat{\theta}_d^{\text{BLUP}}$ is then similar to the direct estimator. This situation normally occurs when the sample size in the domain is large. Conversely, if the direct estimator is unstable and has a large smooth design variance, more weight is given to the synthetic estimator. If the number of domains is large, the prediction variance of $\hat{\theta}_d^{\text{BLUP}}$, $\operatorname{var}\left(\hat{\theta}_d^{\text{BLUP}} - \theta_d \,\middle|\, \mathbf{X}\right)$, is approximately equal to $\gamma_d\tilde{\psi}_d$. Since $\operatorname{var}\left(\hat{\theta}_d - \theta_d \,\middle|\, \mathbf{X}\right) = \tilde{\psi}_d$, the constant $\gamma_d$ can be interpreted as being a variance reduction factor resulting from using $\hat{\theta}_d^{\text{BLUP}}$ instead of $\hat{\theta}_d$. Therefore, the variance reduction is greater when $\gamma_d$ is small, i.e., when the direct estimator is not precise. On the other hand, if the linking model is not properly specified, there is greater risk of significant bias when $\gamma_d$ is small. To better understand this point, suppose that the real linking model is such that $E(\theta_d \mid \mathbf{X}) = \mu(\mathbf{x}_d)$ for some function $\mu(\cdot)$. Under this model, it can be shown that the bias of the predictor $\hat{\theta}_d^{\text{BLUP}}$ is given by

$$E\left(\hat{\theta}_d^{\text{BLUP}} - \theta_d \,\middle|\, \mathbf{X}\right) = -(1 - \gamma_d)\left(\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0\right), \tag{4.13}$$

where

$$\boldsymbol{\beta}_0 = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d\mathbf{x}_d'}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\,\mu(\mathbf{x}_d).$$

If the linear model $\mu(\mathbf{x}_d) = \mathbf{x}_d'\boldsymbol{\beta}$ is valid, the bias disappears. Otherwise, the bias is not zero and increases as $\gamma_d$ decreases or as the specification error of the linking model, $\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0$, increases. When $\gamma_d$ is close to 1, the bias is usually negligible, but so is the variance reduction.

*Remark*: Note that the predictor $\hat{\theta}_d^{\text{BLUP}}$ and the bias (4.13) depend on the variance $\sigma_v^2$. If the linear model (4.10) is not valid, the parameters $\boldsymbol{\beta}$ and $\sigma_v^2$ no longer exist. Yet, the linking model (4.10) can still be postulated and its parameters can be estimated from the observed data as if the model were valid. The model variance $\sigma_v^2$, which enters in the calculation of the predictor $\hat{\theta}_d^{\text{BLUP}}$ and the bias (4.13), can be viewed as being the value towards which an estimator of $\sigma_v^2$ converges.

The predictor (4.12) cannot be calculated because it depends on the unknown variances $\sigma_v^2$ and $\tilde{\psi}_d$. When $\sigma_v^2$ and $\tilde{\psi}_d$ in (4.12) are replaced with estimators $\hat{\sigma}_v^2$ and $\hat{\tilde{\psi}}_d$, the BLUP (4.12) becomes the empirical best linear unbiased predictor, denoted as $\hat{\theta}_d^{\text{EBLUP}}$. There are a number of methods for estimating $\sigma_v^2$ (see Rao and Molina, 2015). One of the most commonly used methods is restricted maximum likelihood. To estimate $\tilde{\psi}_d$, we assume that a design-unbiased estimator of $\psi_d$ is available, denoted by $\hat{\psi}_d$. This assumption is formally written: $E(\hat{\psi}_d \mid \boldsymbol{\Omega}_P) = \psi_d$. It follows that $E(\hat{\psi}_d \mid \mathbf{X}) = \tilde{\psi}_d$. Therefore, the estimator $\hat{\psi}_d$ is unbiased for $\tilde{\psi}_d$, but can be very unstable when the domain sample size is small. A more efficient approach for estimating $\tilde{\psi}_d$ involves modelling $\hat{\psi}_d$ given the auxiliary variables $\mathbf{x}_d$. In practice, a linear model is often used for $\log(\hat{\psi}_d)$, and it is assumed that the model errors follow a normal distribution (for example, Rivest and Belmonte, 2000). Beaumont and Bocci (2016), see also Hidiroglou et al. (2019), provide a method of moments for estimating $\tilde{\psi}_d$ that does not require the normality assumption.

The Fay-Herriot model requires the availability of auxiliary data only at the domain level. The variable $y$ must be measured without error in the probability survey, but it is not essential for the auxiliary source to be perfect. This leaves the door open to all kinds of files external to the probability survey such as big data files. Kim, Wang, Zhu and Cruze (2018) is a recent example where an extension of the Fay-Herriot model was used with auxiliary data from satellite images. Small area estimation methods often achieve significant and sometimes impressive variance reductions (see, for example, Hidiroglou et al., 2019). The trade-off for obtaining these gains is the introduction of model assumptions and the risk that these assumptions do not hold. Therefore, model validation is a critical step in producing small area estimates, as in any model-based approach.

Small area estimation methods are generally used to improve the efficiency of estimators for domains with a small sample size. They could also be used to reduce the data collection costs and respondent burden by reducing the overall sample size of a probability survey for a few, if not all, survey variables. The estimates obtained from the reduced sample and the Fay-Herriot model, for example, could thus have a precision similar to the direct estimates from the probability survey obtained from the full sample. In this context, small area estimation methods would not be used to improve the precision for domains containing few units, but instead to reduce the overall data collection effort while preserving the quality of the estimates.

# 5 Conclusion

In this paper, we presented several methods that use data from a non-probability source while preserving a statistical framework that allows for valid inferences. This, in our view, is essential for national statistical agencies because, without this framework, the usual measures of the quality of the estimates, such as variance or mean square error estimates, disappear and it becomes difficult to draw accurate conclusions. Using data from a non-probability source is not without risk. For model-based approaches, it seems unavoidable to plan enough time and resources for modelling. The literature on classical statistics is replete with tools for validating model assumptions. Although this topic was not adequately covered in the previous sections, careful validation of the assumptions is still a critical step in the success of these approaches (Chambers, 2014) and is one of the recommendations made by Baker et al. (2013).

Estimating the variance or mean square error of the estimators described in the previous sections is also an important topic that we omitted. Yet, this problem does not pose any particular difficulties, in general, and a number of methods exist for variance or mean square error estimation. For design-based approaches, the topic has been extensively covered in the literature (see, for example, Wolter, 2007). This is also true for small area estimation methods (see Rao and Molina, 2015) and for the calibration approach (see Valliant, Dorfman and Royall, 2000). Nevertheless, it might be useful that research be undertaken to adequately address this issue in some specific cases, such as weighting by inverse propensity score or statistical matching by nearest donor.

We assumed that the non-probability source was a subset of the population of interest and that it may be subject to measurement errors. However, there are other potential flaws with non-probability sources. For example, they may contain duplicates or units outside the population. This could make some of the methods discussed in this article unusable, especially the design-based methods. Therefore, it might be useful to tackle these problems in the future.

We mainly limited ourselves to describing several methods that use data from a non-probability sample, whether or not combined with data from a probability survey, once all the data have been collected and processed. There are a number of other methods that use data from non-probability sources during the various stages of a probability survey. For example, one or more non-probability sources can be used to create a sampling frame or improve its coverage. These sources can also be used in a multi-frame sampling context, to replace data collection for certain variables, or to impute the missing values in a probability survey. These topics were not covered in this article, but are reviewed in Lohr and Raghunathan (2017).

The literature on integrating data of a probability and non-probability sample is quite recent. However, there are a number of methods that combine data from two probability surveys (e.g., Hidiroglou, 2001; Merkouris, 2004; Ybarra and Lohr, 2008; Merkouris, 2010; and Kim and Rao, 2012). Such methods may be used to first combine two probability surveys before integrating them with a non-probability source using one of the methods in Section 4. For example, if the total $\mathbf{T}_{\mathbf{x}}$ is unknown, it may be possible to estimate it using more than one probability survey and then use this estimated total in the calibration approach. It still needs to be assessed whether such a strategy would yield significant efficiency gains.

Are probability surveys bound to disappear for the production of official statistics? The question is relevant in the current context of surveys conducted by national statistical agencies where high data collection costs and increasingly lower response rates are observed. In our opinion, the time has not yet come because the alternatives are not reliable and general enough to eliminate the use of probability surveys without severely sacrificing the quality of the estimates. In Section 4, we mentioned that calibration and weighting by inverse propensity score could eliminate the use of a probability survey, provided that a vector of population totals $\mathbf{T}_{\mathbf{x}}$ is available from a census or a comprehensive administrative source. In general, these known totals will not be numerous and effective enough to sufficiently reduce the selection bias of a non-probability sample. To get around this problem, the suggestion has been made in the literature to complement $\mathbf{T}_{\mathbf{x}}$ with other totals estimated using a good-quality probability survey. It seems to us that this is the way to significantly reduce bias and to really take advantage of calibration and weighting by inverse propensity score methods presented in Section 4. Of course, some probability surveys with very low response rates and/or data of questionable quality could occasionally be eliminated in favour of data from non-probability sources. In our view, most surveys conducted by Statistics Canada do not fall into this category. Although they are not perfect, they continue to provide reliable information to meet users' needs and to make informed decisions. The complete elimination of probability surveys seems highly unlikely in the short or medium term. However, it can be expected that their use will be reduced in the future in order to control costs and respondent burden.

# References

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Beaumont, J.-F., and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, 37, 2, 171-179. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11605-eng.pdf.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., and Bocci, C. (2016). *Small Area Estimation in the Labour Force Survey*. Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

Beaumont, J.-F., Bocci, C. and Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Paper presented at the Advisory Committee on Statistical Methods, May 2014, Statistics Canada.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Berg, E., Kim, J.-K. and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4, 436-462.

Bethlehem, J. (2009). The rise of survey sampling. Discussion paper (09015), Statistics Netherlands, The Hague.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching. *Social Science Computer Review*, 34, 59-77.

Brick, J.M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75, 872-888.

Chambers, R. (2014). Survey sampling in official statistics – Some thoughts on directions. *Proceedings of the 2014 International Methodology Symposium*, Statistics Canada, Ottawa, Canada.

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

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 1, 117-144. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54963-eng.pdf.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, 2, 137-161. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

Deville, J.-C. (1991). A theory of quota surveys. *Survey Methodology*, 17, 2, 163-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14504-eng.pdf.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Sherbrooke, Canada.

Deville, J.-C., and Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, 53-69, December 15 and 16, 1993, INSEE, Paris.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 2, 165-176. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9551-eng.pdf.

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

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. New York: John Wiley & Sons, Inc.

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-249.

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.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A., and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23, 2, 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

Haziza, D., and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75, 25-43.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

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

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 2, 143-154. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6091-eng.pdf.

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

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-22. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3101-eng.pdf.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642, Alexandria, VA.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87, S10-S30.

Kim, J.K., and Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

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

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87, S177-S191.

Kim, J.K., Wang, Z., Zhu, Z. and Cruze, N.B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics*, 23, 175-189.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf.

Kott, P.S. (2019). A partially successful attempt to integrate a Web-recruited cohort into an address-based sample. *Survey Research Methods*, 13, 95-101.

Laflamme, F., and Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. *Proceedings of the European Conference on Quality in Official Statistics*, Helsinki, Finland, May 2010.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., and Brisbane, J. (2016). Sample matching: Towards a probabilistic approach for web surveys and big data? Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

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

Lesage, É. (2017). Combiner des données d'enquêtes probabilistes et des données massives non probabilistes pour estimer des paramètres de population finie. Unpublished manuscript.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Lohr, S., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design with applications to the swedish living conditions survey. *Journal of Official Statistics*, 29, 557-582.

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

Mercer, A.W., Kreuter, F., Keeter, S. and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Miller, P.V. (2017). Is there a future for surveys? *Public Opinion Quarterly*, 81, 205-212.

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

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

Rancourt, E. (2019). Admin-first as a statistical paradigm for Canadian statistics: Meaning, challenges and opportunities. *Proceedings of Statistics Canada's 2018 International Methodology Symposium* (to appear).

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K. (2020). Making inference by combining data from multiple sources: An appraisal. *Sankhyā* (under review).

Rao, J.N.K., and Fuller, W. (2017). Sample survey theory and methods: Past, present and future directions. *Survey Methodology*, 43, 2, 145-160. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017002/article/54888-eng.pdf.

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

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling (with discussion). *Survey Methodology*, 29, 2, 107-128. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6787-eng.pdf.

Rässler, S. (2012). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics, New York: Springer, 168.

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

Rivest, L.-P., and Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology*, 26, 1, 67-78. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000001/article/5179-eng.pdf.

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

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Särndal, C.-E., Lumiste, K. and Traat, I. (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? *Survey Methodology*, 42, 2, 219-238. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14663-eng.pdf.

Schonlau, M., and Couper, M.P. (2017). Options for conducting Web surveys. *Statistical Science*, 32, 279-292.

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11824-eng.pdf.

Squire, P. (1988). Why the 1936 *Literary Digest* Poll failed. *Public Opinion Quarterly*, 52, 125-133.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society*, 180, 201-223.

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

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

Wolter, K.M. (2007). *Introduction to Variance Estimation*. Second Edition, New-York: Springer.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Ybarra, L.M., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919-931.

Zhang, L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.