*Statistical Science*
2017, Vol. 32, No. 2, 206–226
DOI: 10.1214/16-STS608


# Construction of Weights in Surveys: A Review

**David Haziza and Jean-François Beaumont**

*Abstract.* Weighting is one of the central steps in surveys. The typical weighting process involves three major stages. At the first stage, each unit is assigned a base weight, which is defined as the inverse of its inclusion probability. The base weights are then modified to account for unit nonresponse. At the last stage, the nonresponse-adjusted weights are further modified to ensure consistency between survey estimates and known population totals. When needed, the weights undergo a last modification through weight trimming or weight smoothing methods in order to improve the efficiency of survey estimates. This article provides an overview of the various stages involved in the typical weighting process used by national statistical offices.

*Key words and phrases:* Calibration estimator, design-based framework, expansion estimator, propensity score adjusted estimator, unequal probability sampling, unit nonresponse, weight smoothing, weight trimming, weighting system.

## 1. INTRODUCTION

Surveys are used to gather data on a usually small subset of a population. Surveys are widely used by a variety of professions and disciplines including but not limited to national statistical offices (NSO); environmental, behavioral and social sciences; and polling market research firms. Typically, the aim is to estimate finite population parameters, which are those describing some aspect of the finite population under study. Frequently encountered parameters include means, totals and proportions. In some cases, the interest lies in estimating more complex parameters such as quantiles and poverty indicators. In most surveys, information is collected on many variables of interest (also called survey variables or characteristics of interest) and the aim is to estimate many population parameters; such surveys are thus often referred to as multipurpose surveys.

*David Haziza is Professor of Statistics, Department of Mathematics and Statistics, Université de Montréal, C.P. 6128 Succ. Centre-Ville, Montreal, Quebec, H3J 3C7 (e-mail: david.haziza@umontreal.ca). Jean-François Beaumont is Chief, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6 (e-mail: jean-francois.beaumont@canada.com).*

Surveys conducted at NSOs provide an abundance of examples of multipurpose surveys. Most often, the data collected are stored in a rectangular data file, each row corresponding to a sample unit (e.g., a business, a household, an individual, etc.) and each column corresponding to a characteristic of interest (age, gender, income, etc.). Made available on the data file is a column of final weights. This set of weights constitutes a weighting system. The idea is to construct a single weighting system applicable to all the characteristics of interest. With a unique weighting system, linearly weighted estimates of totals and other finite population parameters are readily obtained by data users. This is important for routine and timely production of statistics (Särndal and Lundström, 2005). The focus of this paper is to describe the typical weighting process in multipurpose surveys. Weighting in surveys has been discussed in Kish (1992), Kalton and Flores-Cervantes (2003), Särndal (2007), Gelman (2007) and its discussion, Rao et al. (2010), Haziza and Lesage (2016) and Lavallée and Beaumont (2016), among others.

The typical weighting process consists of three major stages:

*Stage* 1: Each sample unit is first assigned a design (or basic) weight, which is defined as the inverse of its inclusion probability in the sample. Most surveys use

some form of unequal probability sampling design that includes stratified sampling, proportional-to-size sampling and multi-stage/phase sampling as special cases. Unequal probability sampling designs are used for a number of reasons: (i) To improve the statistical efficiency of the sampling strategy or to reduce the cost of sampling. (ii) Subpopulations of high interest may be assigned a higher inclusion probability than other subpopulations. This is often referred to as disproportionate sampling. (iii) In multi-stage cluster sampling designs, proportional-to-size sampling of primary sampling units may ensure self-weighting (equal overall probabilities of selection) within strata. This approach provides approximately equal interviewer work loads which is desirable in terms of field operations (Rao, 2005). In Section 2, we describe several sampling designs frequently encountered in practice. Based on the inclusion probabilities, one can construct the basic weighting system; see Section 3.

*Stage* 2: Virtually all surveys suffer from nonresponse, which can be defined as the failure to obtain usable responses from sample units. It is customary to distinguish unit nonresponse from item nonresponse. Unit nonresponse results in a complete lack of information on a given sample unit, whereas item nonresponse occurs when some characteristics, but not all, are collected. The treatment of item nonresponse is beyond the scope of this article. The most common way of dealing with unit nonresponse is to eliminate the nonrespondents from the data file and to adjust the design weights of the respondents to compensate for the elimination of the nonrespondents. To that end, the basic weights of respondents are multiplied by a nonresponse adjustment factor. At this step, survey statisticians aim at reducing the nonresponse bias, which may be appreciable when respondents and nonrespondents are different with respect to the characteristics of interest. Whether or not one will succeed in achieving an efficient bias reduction depends on the availability of powerful auxiliary information (Särndal and Lundström, 2005), which is a set of variables observed for both respondents and nonrespondents. Weighting adjustment procedures in the context of unit nonresponse are discussed in Section 5.

*Stage* 3: The last stage consists of modifying the weights adjusted for nonresponse so that survey weighted estimates agree with known population totals available from external sources (e.g., the census or administrative data) for important variables such as gender, age group or region. This process is known as calibration. It includes post-stratification and raking as special cases. Calibration procedures are discussed in Sections 4 and 5.

In some cases, the weighting process involves an additional stage during which the final weights undergo further modification. If the latter are highly dispersed and are poorly related to the characteristics of interest, the resulting estimators may exhibit a large variance. A number of approaches, including weight trimming and weight smoothing, have been developed to improve the efficiency of survey estimates. These methods are discussed in Section 6.

In this paper, the properties (e.g., bias and variance) of estimators are studied with respect to the sampling design. This inferential approach is usually referred to as the design-based approach; see Skinner and Wakefield (2017) for a discussion of the alternative model-based approach.

## 2. SAMPLING DESIGN

Let $U$ be a finite population consisting of $N$ units. In survey sampling, the interest usually lies in estimating finite population parameters such as the population total (or the population mean) of a characteristic of interest. Let $y_1, \ldots, y_p$, denote $p$ characteristics of interest (e.g., income, age, gender, job status, etc.) and let $\mathbf{y}_k = (y_{1k}, \ldots, y_{pk})^\top$ be the $p$-vector of characteristics of interest corresponding to the $k$th unit, $k = 1, \ldots, N$. We denote by $\mathbf{Y}$ the $p \times N$ matrix whose $N$ columns are the $p$ dimensional vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$. A finite population parameter is any function $h(\mathbf{Y})$.

To estimate finite population parameters, it is common practice to select a sample $S$ from the finite population. We assume that a $q$-vector of auxiliary variables, $\mathbf{z}$, is available for all the population units prior to sampling and let $\mathbf{Z}$ denote a $q \times N$ matrix whose $N$ columns are the $q$ dimensional vectors $\mathbf{z}_1, \ldots, \mathbf{z}_N$. In practice, virtually all surveys conducted by NSO's use some form of auxiliary information in order to stratify the population and/or to select the sample with inclusion probabilities proportional to some size variable. The sample $S$ is selected according to a given sampling design $\mathcal{F}(\mathbf{I}|\mathbf{Z})$, where $\mathbf{I} = (I_1, \ldots, I_N)^\top$ is the $N$-vector of sample selection indicators such that $I_k = 1$ if unit $k$ is selected in the sample and $I_k = 0$, otherwise.

The first moment of the distribution $\mathcal{F}(\mathbf{I}|\mathbf{Z})$ is $E(\mathbf{I}|\mathbf{Z}) = (\pi_1(\mathbf{Z}), \ldots, \pi_N(\mathbf{Z}))^\top \equiv \boldsymbol{\pi}(\mathbf{Z})$, the vector of first-order inclusion probabilities. To ease notation, we write $\pi_k(\mathbf{Z}) = P(I_k = 1|\mathbf{Z}) \equiv \pi_k$ for the first-order inclusion probability attached to unit $k$. We assume that

$\pi_k > 0$ for all $k \in U$. This condition is often violated in practice, in which case we are in the presence of under-coverage.

The reader is referred to Tillé (2017) for a description of some basic sampling designs including simple random sampling without replacement, Bernoulli/Poisson sampling and proportional-to-size sampling without replacement. Below, we briefly describe two-stage and two-phase sampling designs.

### 2.1 Two-Stage Sampling

Suppose the population $U$ is partitioned into $M$ clusters or primary sampling units (p.s.u.). At the first stage, a sample $S_1$ of $m$ p.s.u.'s is selected according to a given sampling design. At the second stage, in the $i$th p.s.u. selected at the first stage, a sample $S_i$ of secondary sampling units (also called elements or ultimate units) is selected according to a given sampling design, $i = 1, \ldots, m$. The selection of elements in a p.s.u. selected at the first stage is independent of the selection of elements in any other p.s.u. selected at the first stage. The first-order inclusion probability of element $k$ in the $i$th p.s.u. is given by $\pi_k = \pi_i \times \pi_{k|i}$, where $\pi_i$ is the first-order inclusion probability of the $i$th p.s.u. in the first-stage sample, $S_1$, and $\pi_{k|i}$ is the first-order inclusion probability of element $k$ in the $i$th p.s.u. given that the $i$th p.s.u. was selected in $S_1$.

### 2.2 Two-Phase Sampling

We discuss two-phase sampling because of its close connection with unit nonresponse (see Section 5). Two-phase sampling is used in surveys when the sampling frame contains little or no auxiliary information. It consists of first selecting a large sample from the population (typically using a rudimentary sampling design) in order to collect data on variables that are inexpensive to obtain and that are related to the characteristics of interest. The idea behind two-phase sampling is to create a pseudo-sampling frame richer in auxiliary information than the original frame. Then, using the variables observed in the first phase, an efficient sampling procedure can be used to select a (typically small) subsample from the first-phase sample in order to collect the characteristics of interest.

Two-phase sampling may also be useful for surveying rare populations when an appropriate sampling frame does not exist. For instance, it is used at Statistics Canada to survey aboriginal people. There does not exist any frame containing aboriginal people in Canada. The first-phase sample consists of selecting a large sample of households and collecting basic information on all the household members such as the aboriginal status. In 2016, these data were collected through the long-form Census administered to roughly one-fourth of the Canadian population. Then, at the second phase, a stratified sample is selected among aboriginal people selected in the first phase (Cloutier and Langlet, 2014). Statistics Canada conducts other similar post-censal surveys targeting rare populations and using the long-form Census as a sampling frame for a second phase of sampling.

Two-phase sampling may be described as follows: a $N$-vector $\mathbf{I}_1$ is generated according to the sampling design $\mathcal{F}(\mathbf{I}_1|\mathbf{Z}_1)$, where $\mathbf{I}_1$ is the $N$-vector of first-phase indicators and $\mathbf{Z}_1$ is a matrix of auxiliary information available for all the population units prior to first-phase sampling. Let $S_1$ be the first-phase sample. Then a $n_1$-vector of second-phase sample indicators, $\mathbf{I}_2$, is generated according to the sampling design $\mathcal{F}(\mathbf{I}_2|\mathbf{I}_1, \mathbf{Z}_2^*)$, where $\mathbf{Z}_2^*$ denotes a matrix of auxiliary information available for all the first-phase sample units. The population version of $\mathbf{Z}_2^*$ is denoted by $\mathbf{Z}_2$ but is not observed for units not selected at the first phase. Let $\pi_{1k} = P(I_{1k} = 1|\mathbf{Z}_1)$ be the first-order inclusion probability of unit $k$ in the first-phase sample and let $\pi_{2k}(\mathbf{I}_1) = P(I_{2k} = 1|\mathbf{I}_1, \mathbf{Z}_2^*)$ be the first-order inclusion probability of unit $k$ in the second-phase sample. The second-phase sample, denoted by $S_2$, is the set of population units for which both $I_{1k} = 1$ and $I_{2k} = 1$. Note that the inclusion probabilities at the second-phase may depend on the realized sample $S_1$. For this reason, the inclusion probabilities $\pi_k$ are intractable except when the two-phase design is invariant (Beaumont and Haziza, 2016). For an arbitrary two-phase sampling design, the $\pi_k$'s are generally unknown because their calculation requires knowledge not only of $P(\mathbf{I}_1 = \mathbf{i}_1|\mathbf{Z}_1)$ for every possible realisation of $\mathbf{I}_1$ (in many cases, these are known) but also of $\pi_{2k}(\mathbf{I}_1)$ for every realisation of $\mathbf{I}_1$. The latter are generally unknown because $\pi_{2k}(\mathbf{I}_1)$ may depend on the outcome of phase 1.

### 2.3 Indirect Sampling

The selection of a sample from the finite population $U$ requires the availability of a sampling frame $U^F$ that contains a list of units that covers the entire population $U$. Ideally, the sampling frame $U^F$ and the finite population $U$ are identical so that sampling from $U^F$ is equivalent to sampling from $U$. In some sample surveys, such a frame is not available but there exists a sampling frame $U^F$ related in some way to the population of interest $U$. We assume that each unit of $U$

is linked to at least one unit of $U^F$. We denote by $l_{ik}$, the binary variable indicating whether unit $i \in U^F$ is linked to unit $k \in U$ ($l_{ik} = 1$) or not ($l_{ik} = 0$). We also denote by $U_k^F$, the set of units in $U^F$ that have a link with unit $k \in U$, and by $U_i$, the set of units in $U$ that have a link with unit $i \in U^F$.

Under the above set-up, a sample $S$ from $U$ can be obtained indirectly by drawing a sample $S^F$ from $U^F$ using standard selection methods. For each unit $i \in S^F$, the information is collected for all the units $k \in U_i$. We denote by $S_k^F$, the set of units in $S^F$ that have a link with unit $k \in U$. The sample $S$ is simply the set of units in $U$ that have a link with at least one unit in $S^F$. In other words, $S$ is the set of units $k \in U$ for which $S_k^F$ is not empty. When indirect sampling is used, the first-order inclusion probability of a unit $k \in U$, $\pi_k$, may be difficult to derive but the first-order inclusion probability of a unit $i \in U^F$, denoted by $\pi_i^F$, can be easily computed provided that standard selection methods have been used to sample from $U^F$.

The Motion Picture Production Survey conducted by Statistics Canada provides an example of indirect sampling. In that survey, the sampling frame (called the Business Register) contains an incomplete list of all the establishments for which we are interested in collecting information. A stratified simple random sample $S^F$ is drawn from that incomplete list, $U^F$. For any given selected establishment $i \in S^F$, the contact information of the company that possesses this establishment is obtained and the desired information is collected for all the establishments belonging to that company (all $k \in U_i$), including those that were not listed in the sampling frame. This is done for each selected establishment in $S^F$. This selection procedure avoids the bias due to the incomplete frame as long as each establishment of the entire population $U$ has a link with at least one establishment in the sampling frame.

Sometimes, the population $U$ can only be reached through the use of more than one overlapping sampling frames, $U_r^F$, $r = 1, \ldots, R$, with $R$ being the number of frames. A sample is then selected from each frame, each of them may be incomplete but the union of all the frames, $U^F = \bigcup_{r=1}^R U_r^F$, is assumed to cover $U$. Multiple frame surveys can be viewed as a special case of indirect sampling (see Lavallée, 2007).

## 3. BASIC WEIGHTING SYSTEM

### 3.1 Basic Set-up

For simplicity, we use the generic notation $y$ to denote a characteristic of interest but the reader should keep in mind that a typical survey involves a possibly large number of variables. We focus on an important parameter in practice: the population total, $t_y = \sum_{k \in U} y_k$, where $y_k$ denotes the $k$th value of the characteristic of interest $y$, $k = 1, \ldots, N$. We consider estimators of $t_y$ of the form

$$\hat{t}_y = \sum_{k \in S} w_k y_k, \tag{3.1}$$

where $w_k$ is a weight attached to unit $k$. The weight $w_k$ may vary from one unit to another and, for a given unit, may vary from one sample to another. The set $\{w_k; k \in S\}$ is called a weighting system. Note that the weights $w_k$ are constructed without using the values of the characteristics of interest. As a result, the weighting system $\{w_k; k \in S\}$ can be applied to any characteristic of interest.

In the case of design-based inference, the properties of estimators are studied with respect to the sampling design. That is, when taking the expectation of an estimator, the only source of randomness is the vector $\mathbf{I}$ and all the other quantities involved in the inference are treated as fixed. The latter include the design variables $\mathbf{Z}$ and the characteristics of interest $\mathbf{Y}$. To simplify the notation, the symbol $\mathbf{\Omega}$ is used to denote any quantity but the vector $\mathbf{I}$. The design-expectation of $\hat{t}_y$ is defined as $E(\hat{t}_y|\mathbf{\Omega})$. In a calibration context (see Section 4), $\mathbf{\Omega}$ will also include the calibration variables.

In this paper, we discuss the case of population totals. In practice, the interest often lies in estimating more complex parameters. For example, many parameters encountered in practice can be expressed as smooth functions of totals, for example, ratios and coefficients of correlation. These can be estimated using a plug-in type estimator, whereby each population total is replaced by its corresponding weighted estimator. The reader is referred to Särndal et al. (1992), Chapter 5, for a treatment of complex parameters.

### 3.2 Expansion Estimators

The basic (or design) weight attached to unit $k$ is defined as the inverse of its inclusion probability. That is, $d_k = 1/\pi_k$. The basic weighting system is $\{d_k; k \in S\}$. Applying the basic weighting system to the $y$-variable leads to the well-known expansion estimator or $\pi$-estimator (Narain, 1951, Horvitz and Thompson, 1952) of $t_y$:

$$\hat{t}_{y,\pi} = \sum_{k \in S} d_k y_k = \sum_{k \in U} d_k y_k I_k. \tag{3.2}$$

The $d_k$'s are often referred to as expansion factors because, when applied to the sample $y$-values, the estimator $\hat{t}_{y,\pi}$ reaches the level of the whole population. Noting that $E(I_k|\mathbf{\Omega}) = \pi_k$, the estimator (3.2) is design-unbiased for $t_y$ regardless of the characteristic of interest $y$ being considered. That is, $E(\hat{t}_{y,\pi}|\mathbf{\Omega}) = t_y$ for all $\mathbf{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$. Also, under mild regularity conditions, $\hat{t}_{y,\pi}$ is design-consistent for $t_y$ in the sense that $\hat{t}_{y,\pi} - t_y = O_p(N/\sqrt{n})$. For a discussion of the asymptotic framework as well as the regularity conditions required for establishing design-consistency, the reader is referred to Breidt and Opsomer (2016). It follows that, in the absence of nonsampling errors, the basic weighting system ensures that, when applied to any characteistic of interest $y$, the resulting estimator is design-unbiased and design-consistent. An expression of the design variance of $\hat{t}_{y,\pi}$ can be found in Breidt and Opsomer (2016), equation (2). For a fixed-size sampling design, the design-variance of $\hat{t}_{y,\pi}$, $V(\hat{t}_{y,\pi}|\mathbf{\Omega})$, is equal to zero when $y_k \propto \pi_k$. This result suggests that the expansion estimator is very efficient when $y_k$ is approximately proportional to $\pi_k$. Of course, in practice, it is not possible to satisfy the requirement $y_k \propto \pi_k$ as the inclusion probabilities are set prior to sampling, and the $y$-values are available only after the sample has been selected.

A naive estimator of $t_y$ that does not use the design information is the unweighted estimator

$$\hat{t}_{y,un} = \frac{N}{n}\sum_{k\in S} y_k = \frac{N}{n}\sum_{k\in U} y_k I_k. \tag{3.3}$$

Here, the expansion factor is equal to $N/n$ for all $k$. The unweighted estimator is generally design-biased. Assuming a sampling design with a fixed sample size $n$, its relative bias is given by

$$\begin{aligned}&\frac{E(\hat{t}_{y,un}|\mathbf{\Omega}) - t_y}{t_y}\\ &\quad = \frac{(N-1)}{N} R_{y,\pi}\mathrm{CV}(y)\mathrm{CV}(\pi),\end{aligned} \tag{3.4}$$

where

$$R_{y,\pi} = \frac{1}{(N-1)}\frac{\sum_{k\in U}(y_k - \overline{y}_U)(\pi_k - \overline{\pi}_U)}{S_y S_\pi}$$

denotes the finite population coefficient of correlation between $y$ and $\pi$ with $\overline{y}_U = t_y/N$, $\overline{\pi}_U = \sum_{k\in U}\pi_k/N$, $S_y = [(N-1)^{-1}\sum_{k\in U}(y_k - \overline{y}_U)^2]^{1/2}$ and $S_\pi$ is similarly defined, $\mathrm{CV}(y) = S_y/\overline{y}_U$ and $\mathrm{CV}(\pi) = S_\pi/\overline{\pi}_U$ are the coefficients of variation of the $y$-variable and the $\pi$-variable, respectively. From (3.4), the unweighted estimator is generally biased unless $R_{y,\pi} = 0$ (i.e., the $y$-variable is unrelated to the $\pi$-variable) and/or $\mathrm{CV}(\pi) = 0$, which corresponds to a self-weighting design (i.e., a design for which all the units have equal weights). The bias of $\hat{t}_{y,un}$ increases as the correlation between the $y$-variable and the $\pi$-variable increases. When $R_{y,\pi} = 0$, the bias of $\hat{t}_{y,un}$ vanishes and the latter may be significantly more stable than $\hat{t}_{y,\pi}$. This was illustrated in the famous circus example of Basu (1971), where the expansion estimator led to a disastrous result; see also Rao (1966) and Scott and Smith (1969) for a theoretical discussion. In general, $\hat{t}_{y,\pi}$ tends to be unstable when the $\pi_k$'s are highly dispersed and are poorly related to the characteristic of interest; see also Brewer (2002).

### 3.3 Empirical Illustration

We conducted a simulation study in order to compare the performance of $\hat{t}_{y,\pi}$ and $\hat{t}_{y,un}$ in terms of bias and efficiency. We generated a finite population of size $N = 1000$ consisting of ten variables: an auxiliary variable $z$ available for all the population units prior to sampling and nine characteristics of interest $y_1 - y_9$. The nine characteristics of interest were generated in order to illustrate the problem of the choice of weights in multipurpose surveys. First, the $z$-values were generated according to a Gamma distribution with shape parameter equal to 1 and scale parameter equal to 10, resulting in a coefficient of variation of the $z$-values, $\mathrm{CV}(z)$, equal to 1. Given the $z$-values, the values of $y_1$ to $y_9$ were generated according to the model

$$y_{tk} = \beta_{0t} + \beta_{1t} z_k + \sqrt{z_k}\varepsilon_k, \quad t = 1, \ldots, 9,$$

where the errors $\varepsilon_k$ were generated from a normal distribution with mean equal to 0 and variance $\sigma^2$. In each population, the coefficient of variation of the $y$-values, $\mathrm{CV}(y)$, and the value of $\sigma^2$ were set to 1 and 12, respectively. The values of $\beta_{0t}$ and $\beta_{1t}$ were chosen so that the coefficient of correlation between $y_t$ and $z$ was set to $0.1t$, $t = 1, \ldots, 9$.

From the population, we selected 10,000 samples of size $n = 50$ using the Rao–Sampford design (Rao, 1965, Sampford, 1967) with probability proportional to the $z$-variable, that is, $\pi_k = nz_k/\sum_{k\in U} z_k$. For such a design, we have $R_{y,z} = R_{y,\pi}$ and $\mathrm{CV}(\pi) = \mathrm{CV}(z) = 1$. This suggests that the $\pi_k$'s were fairly dispersed in our experiment.

We were interested in estimating the population total for each of the characteristics of interest $y_1 - y_9$.

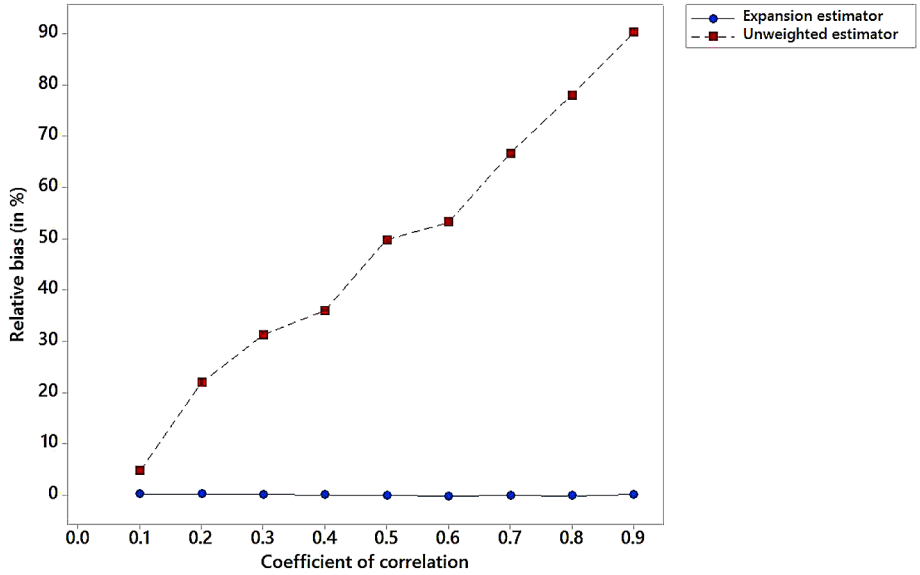


FIG. 1. *Relative bias of the Horvitz–Thompson and the unweighted estimator for several values of* $R_{y,\pi}$.

In each sample, we computed the expansion estimator given by (3.2) and the unweighted estimator given by (3.3). Define the Monte Carlo average of an estimator $\hat{t}$ by

$$E_{\mathrm{MC}}(\hat{t}) = \frac{1}{10{,}000} \sum_{r=1}^{10{,}000} \hat{t}_{(r)},$$

where $\hat{t}_{(r)}$ denotes the estimator $\hat{t}$ in the $r$th repetition, $r = 1, \ldots, 10{,}000$. As a measure of bias of an estimator $\hat{t}$, we computed the Monte Carlo percent relative bias defined as $RB_{\mathrm{MC}}(\hat{t}) = 100 \times \{E_{\mathrm{MC}}(\hat{t}) - t_y\}/t_y$. As a measure of variability of $\hat{t}$, we computed the Monte Carlo variance defined as $V_{\mathrm{MC}}(\hat{t}) = E_{\mathrm{MC}}\{\hat{t} - E_{\mathrm{MC}}(\hat{t})\}^2$. Finally, we computed the Monte Carlo mean square error defined as $\mathrm{MSE}_{\mathrm{MC}}(\hat{t}) = E_{\mathrm{MC}}(\hat{t} - t_y)^2$.

Figure 1 shows the Monte Carlo percent relative bias of $\hat{t}_{y,\pi}$ and $\hat{t}_{y,un}$ for varying values of $R_{y,\pi}$. As expected, $\hat{t}_{y,\pi}$ showed a negligible bias for every value of $R_{y,\pi}$. In contrast, the bias of $\hat{t}_{y,un}$ increased as $R_{y,\pi}$ increased. This is consistent with (3.4).

Figure 2 shows the percent relative variance and the percent relative efficiency

$$\mathrm{RV} = 100 \times \frac{V_{\mathrm{MC}}(\hat{t}_{y,\pi})}{V_{\mathrm{MC}}(\hat{t}_{y,un})},$$

$$\mathrm{RE} = 100 \times \frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{t}_{y,\pi})}{\mathrm{MSE}_{\mathrm{MC}}(\hat{t}_{y,un})}$$

for varying values of $R_{y,\pi}$. From Figure 2, $\hat{t}_{y,un}$ was more stable than $\hat{t}_{y,\pi}$ for all values of $R_{y,\pi}$ as the values of RV were all greater than 100. For example, for $R_{y,\pi} = 0.5$, the value of RV was approximately equal to 480. Turning to relative efficiency, $\hat{t}_{y,un}$ was more efficient than the expansion estimator for small values of $R_{y,\pi}$ (less than 0.4). For $R_{y,\pi} \geq 0.5$, $\hat{t}_{y,\pi}$ was more efficient than $\hat{t}_{y,un}$ with values of RE smaller than 100. For example, for $R_{y,\pi} = 0.9$, the value of RE was approximately equal to 85.

These results suggest that, although the expansion estimator is design-unbiased, it may be inefficient if the $\pi_k$'s are highly dispersed and are poorly related to the characteristic of interest.

### 3.4 The Double Expansion Estimator

As mentioned in Section 2.2, the first-order inclusion probabilities $\pi_k$ are generally unknown in the case of an arbitrary two-phase design. As a result, the basic weighting system $\{d_k; k \in S\}$ described in Section 3.2 cannot be used since the $d_k$'s are unknown. To cope with this problem, we use the weighting system $\{\tilde{d}_k; k \in S_2\}$, where $\tilde{d}_k = 1/\{\pi_{1k}\pi_{2k}(\mathbf{I}_1)\}$. Note that $\tilde{d}_k \neq d_k$, in general, unless the two-phase design is invariant. Applying the weighting system $\{\tilde{d}_k; k \in S_2\}$ to a characteristic of interest $y$ leads to double expansion estimator (Särndal et al., 1992, Chapter 9) of $t_y$:

$$\hat{t}_{y,\mathrm{DE}} = \sum_{k \in S_2} \tilde{d}_k y_k = \sum_{k \in U} \tilde{d}_k y_k I_{1k} I_{2k}. \tag{3.5}$$

Let $\boldsymbol{\Omega}$ denote any quantity involved in the inference but the vectors of sample inclusion indicators $\mathbf{I}_1$ and $\mathbf{I}_2$. Noting that $E(I_{2k}|\mathbf{I}_1, \boldsymbol{\Omega}) = \pi_{2k}(\mathbf{I}_1)$ and $E(I_{1k}|\boldsymbol{\Omega}) = \pi_{1k}$, it follows that the estimator (3.5) is

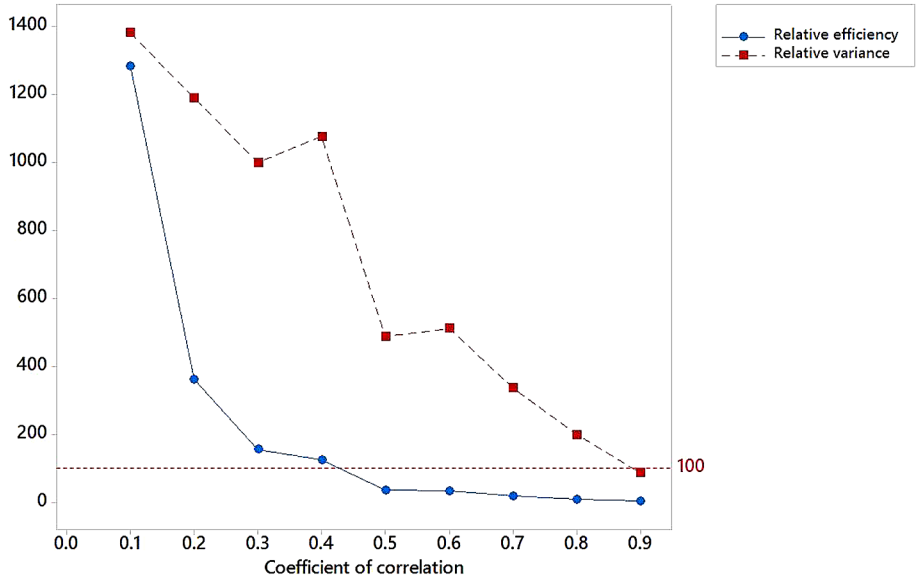


FIG. 2. *Relative variance and relative efficiency for several values of* $R_{y,\pi}$.

design-unbiased for $t_y$ regardless of the characteristic of interest $y$ being considered. That is, $E(\hat{t}_{y,\mathrm{DE}}) = E\{E(\hat{t}_{y,\mathrm{DE}}|\mathbf{I}_1, \boldsymbol{\Omega})|\boldsymbol{\Omega}\} = t_y$ for all $\mathbf{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$.

### 3.5 Generalized Weight Share Method

With indirect sampling, the $\pi_k$'s are often difficult or even impossible to compute exactly and the expansion estimator cannot be computed. One possible alternative consists of taking advantage of the links that relate the population $U$ to the sampling frame $U^F$. The basic principle behind this approach is to express a total over the population $U$ in terms of a total over $U^F$. Then standard weighting methods can be used to estimate the latter total. More formally, consider the estimation of $t_y = \sum_{k\in U} y_k$ and define constants $\alpha_{ik}$ such that

$$\sum_{i\in U^F} l_{ik}\alpha_{ik} = \sum_{i\in U_k^F} \alpha_{ik} = 1,$$

for all $k \in U$. A simple and common choice is $\alpha_{ik} = 1/M_k$, where $M_k = \sum_{i\in U^F} l_{ik}$ is the size of $U_k^F$; that is, the number of units in the sampling frame $U^F$ that have a link with unit $k \in U$. The total $t_y$ can then be rewritten as

$$t_y = \sum_{k\in U} y_k = \sum_{k\in U} y_k \sum_{i\in U^F} l_{ik}\alpha_{ik} = \sum_{i\in U^F}\sum_{k\in U} l_{ik}\alpha_{ik} y_k$$
$$= \sum_{i\in U^F}\sum_{k\in U_i} \alpha_{ik} y_k = \sum_{i\in U^F} y_i^F,$$

where $y_i^F = \sum_{k\in U} l_{ik}\alpha_{ik} y_k = \sum_{k\in U_i} \alpha_{ik} y_k$.

The estimation of $t_y = \sum_{k\in U} y_k$ can thus be achieved by estimating $\sum_{i\in U^F} y_i^F$ using standard weighting methods. For instance, the expansion estimator, $\hat{t}_{y,\pi}^F = \sum_{i\in S^F} d_i^F y_i^F$, can be used, where $d_i^F = 1/\pi_i^F$.

The estimator $\hat{t}_{y,\pi}^F$ can be used directly but it requires the creation and the use of the modified values $y_i^F$. Some users would prefer to use the original values $y_k$, for $k \in S$, along with a proper set of weights. This can be done using this approach by noting that $\hat{t}_{y,\pi}^F$ can be rewritten as

$$\hat{t}_{y,\pi}^F = \sum_{i\in S^F} d_i^F y_i^F = \sum_{i\in S^F} d_i^F \sum_{k\in U} l_{ik}\alpha_{ik} y_k$$
$$= \sum_{k\in S} y_k \sum_{i\in S_k^F} d_i^F \alpha_{ik} = \sum_{k\in S} w_k^{\mathrm{GWS}} y_k,$$

where $w_k^{\mathrm{GWS}} = \sum_{i\in S_k^F} d_i^F \alpha_{ik}$.

The weight $w_k^{\mathrm{GWS}}$ can be obtained using the so-called generalized weight share method (e.g., Lavallée, 2007; Deville and Lavallée, 2006). The condition $\sum_{i\in U_k^F} \alpha_{ik} = 1$, $k \in U$, ensures that the estimator $\hat{t}_{y,\pi}^F$ is design-unbiased for $t_y$. The choice $\alpha_{ik} = 1/M_k$ is common but may be inefficient. As an illustration, suppose that one unit $j \in U^F$ is selected with certainty, that is, $\pi_j^F = 1$. All the units $k \in U_j$ will thus be selected with certainty in $S$, that is, $\pi_k = 1$, for $k \in U_j$. As a result, efficiency considerations suggest that, for any given unit $k \in U_j$, $\alpha_{ik}$, for $i \in U_k^F$, should be chosen so that $w_k^{\mathrm{GWS}} = 1$. This can be achieved by letting $\alpha_{jk} = 1$ and $\alpha_{ik} = 0$, for all other units $i \in U_k^F$, $i \neq j$.

Intuitively, a unit $i \in U_k^F$ with a relatively large probability $\pi_i^F$ should be associated with a large value of $\alpha_{ik}$. This is not the case of $\alpha_{ik} = 1/M_k$ as it gives the same value for all units $i \in U_k^F$. Optimal or, at least, more efficient choices of $\alpha_{ik}$ are discussed in greater depth in Deville and Lavallée (2006).

## 4. CALIBRATED WEIGHTING SYSTEM

In practice, some auxiliary information is often available at the estimation stage. Let $\mathbf{x} = (x_1, \ldots, x_J)^\top$ be a $J$-vector of auxiliary variables recorded for all the sample units. Further, we assume that the vector of population totals, $\mathbf{t_x} = (t_{x_1}, \ldots, t_{x_J})^\top$, is available from external sources, where $t_{x_j} = \sum_{k \in U} x_{jk}$. External sources include the census and administrative records. While the basic weighting system ensures unbiasedness, that is, $E(\hat{\mathbf{t}}_{\mathbf{x},\pi} | \boldsymbol{\Omega}) = \mathbf{t_x}$, it does not produce an exact estimate for each of the $J$ auxiliary variables; that is, $\hat{\mathbf{t}}_{\mathbf{x},\pi} \neq \mathbf{t_x}$, in general. This lack of consistency may be embarrassing in practice. For instance, suppose that two distinct surveys collect the variable gender for each sample unit. The estimated number of males and females will almost certainly differ in both surveys. Further, these estimates will almost certainly differ from the census counts, which begs the question: which set of estimates to consider in an analysis involving the gender? To cope with this problem, the basic weights are modified so that survey estimates coincide with known population totals available from external sources (e.g., the census counts). This process is known as calibration. The reasons for using calibration are three-fold:

(i) to force consistency of certain survey estimates to known population quantities;

(ii) to reduce nonsampling errors such as nonresponse errors and coverage errors;

(iii) to improve the precision of estimates.

The use of calibration for reducing nonresponse errors is discussed in Section 5. The use of calibration for handling coverage errors is outside the scope of this paper. Its justification usually relies on the assumption that a linear model holds between the characteristics of interest and the auxiliary variables. The reader is referred to Särndal (2007), Kott (2009), Kim and Park (2010) and Wu and Lu (2016) for excellent overviews of calibration weighting.

We seek a calibrated weighting system $\{\tilde{w}_k; k \in S\}$ such that the weights $\tilde{w}_k$ are "as close as possible" to the design weights $d_k$ while satisfying the calibration constraints

$$\sum_{k \in S} \tilde{w}_k \mathbf{x}_k = \mathbf{t_x}; \tag{4.1}$$

see Deville and Särndal (1992) and Deville et al. (1993). There are infinitely many possible weighting systems $\{\tilde{w}_k; k \in S\}$ that satisfy the calibration constraints (4.1). The goal is to determine a set of calibrated weights $\tilde{w}_k$ that is close (in a sense that we define below) to the set of pre-calibrated weights $d_k$, which ensures that the resulting calibration estimator will be design-consistent.

As a measure of closeness between the pre-calibrated weights and the calibrated weights, we consider the distance function $G(\tilde{w}_k/d_k)$ such that $G(\tilde{w}_k/d_k) \geq 0$, $G(1) = 0$, differentiable with respect to $\tilde{w}_k$, strictly convex with continuous derivatives $g(u) = \partial G(u)/\partial u$ and $g(1) = 0$. The calibration problem may be formulated as follows: we seek a calibrated weighting system $\{\tilde{w}_k; k \in S\}$ that minimizes

$$\sum_{k \in S} \frac{d_k G(\tilde{w}_k/d_k)}{q_k} \tag{4.2}$$

subject to (4.1), where the coefficient $q_k$ is a scale factor indicating the importance of unit $k$ in the distance calculation. In most practical situations, the factor $q_k$ is set to 1. Using the method of Lagrange multipliers, we define

$$\phi(\tilde{w}_1, \ldots, \tilde{w}_n, \boldsymbol{\lambda}) = \sum_{k \in S} \frac{d_k G(\tilde{w}_k/d_k)}{q_k} - \boldsymbol{\lambda}^\top \Big( \sum_{k \in S} \tilde{w}_k \mathbf{x}_k - \mathbf{t_x} \Big),$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^\top$ is a $J$-vector of Lagrange multipliers. Differentiating $\phi(\tilde{w}_1, \ldots, \tilde{w}_n, \boldsymbol{\lambda})$ with respect to $\tilde{w}_k$ and setting the derivative equal to zero, we obtain

$$\tilde{w}_k = d_k F(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k), \tag{4.3}$$

where $F(u) = g^{-1}(u)$ is the inverse function of $g(\cdot)$, which is often referred to as the calibration function. Substituting (4.3) in (4.1) leads to

$$\sum_{k \in S} d_k F(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k) \mathbf{x}_k = \mathbf{t_x}. \tag{4.4}$$

It remains to solve (4.4) for $\boldsymbol{\lambda}$, which involves a system of $J$ equations in $J$ unknowns. This can be achieved using numerical methods such as the Newton–Raphson algorithm. Let $\hat{\boldsymbol{\lambda}}$ be the solution of (4.4). The calibrated weight attached to unit $k$ is given by

$$\tilde{w}_k = d_k F(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k). \tag{4.5}$$

The calibration weights (4.5) can be thought of as the product of the pre-calibrated weight, $d_k$, and a calibration factor, $F(q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$. The latter depends on the calibration function $F(\cdot)$, the characteristics of unit $k$ through $q_k$ and $\mathbf{x}_k$ as well as $\hat{\boldsymbol{\lambda}}^\top$, that can be viewed as a measure of sample imbalance. If the sample is balanced with respect to the $\mathbf{x}$-vector in the sense that $\hat{\mathbf{t}}_{\mathbf{x},\pi} = \mathbf{t}_{\mathbf{x}}$, then $\tilde{w}_k = d_k F(0) = d_k$ for all $k$. An alternative formulation of calibration has been given by Deville (2002), Estevao and Särndal (2000) and Kott (2006), among others. It consists of replacing $q_k\mathbf{x}_k$ in (4.5) by a vector of variables $\mathbf{z}_k$, which have been called instruments by these authors, and then solving the calibration constraints (4.4). The optimal choice of $\mathbf{z}_k$ ends up with the so-called optimal estimator developed by Montanari (1987); see also Rao (1994).

Applying the calibrated weighting system $\{\tilde{w}_k; k \in S\}$ to the $y$-variable leads to the calibration estimator

$$\hat{t}_{y,C} = \sum_{k\in S} \tilde{w}_k y_k = \sum_{k\in S} d_k F\big(q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k\big) y_k. \tag{4.6}$$

The estimator $\hat{t}_{y,C}$ is generally design-biased; that is, $E(\hat{t}_{y,C}|\boldsymbol{\Omega}) \neq t_y$. However, it is design-consistent for $t_y$ in the sense $\hat{t}_{y,C} - t_y = O_p(N/\sqrt{n})$; see, for example, Deville and Särndal (1992) and Kim and Park (2010). Now, suppose that the $y$-variable is linearly related to the vector of calibration variables $\mathbf{x}$ and that the relationship is perfect; that is, $y_k = \mathbf{x}_k^\top \boldsymbol{\beta}$ for all $k$, where $\boldsymbol{\beta}$ is a $J$-vector of unknown coefficients. In this case, $\hat{t}_{y,C}$ provides a perfect estimate of $t_y$ since

$$\begin{aligned}\hat{t}_{y,C} &= \sum_{k\in S} \tilde{w}_k y_k = \sum_{k\in S} \tilde{w}_k \mathbf{x}_k^\top \boldsymbol{\beta} \\ &= \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\beta} = \sum_{k\in U} y_k = t_y,\end{aligned}$$

where the third equality follows from (4.1). Therefore, we expect $\hat{t}_{y,C}$ to be very efficient if there exists a strong linear relationship between the $y$-variable and the vector $\mathbf{x}$. It is worth noting that all the calibration estimators are asymptotically equivalent to the calibration estimator obtained through the linear method; see (4.8) below. An expression of the asymptotic design variance of $\hat{t}_{y,C}$ is given in Breidt and Opsomer (2016).

### 4.1 Some Common Calibration Methods

4.1.1 *Calibration estimators that do not depend on the calibration function.* Given a vector of auxiliary variables $\mathbf{x}_k$, different calibration functions $F(\cdot)$ lead to different calibration estimators $\hat{t}_{y,C}$, in general. There are some exceptions to this rule; that is, there exist vectors of auxiliary variables, $\mathbf{x}_k$, that lead to the same calibration estimator regardless of $F(\cdot)$. These estimators include the Hàjek estimator (Hàjek, 1971), the ratio estimator and the post-stratified estimator. They are presented below:

(i) When the population size $N$ is known, it is natural to require the estimated population size $\sum_{k\in S} \tilde{w}_k$ to match the true size $N$. To that end, it suffices to set $\mathbf{x}_k = 1$ and $q_k = q$ for all $k$ so that $\sum_{k\in U} \mathbf{x}_k = N$. The calibrated weight $\tilde{w}_k$ in (4.5) reduces to $\tilde{w}_k = d_k(N/\hat{N}_\pi)$, where $\hat{N}_\pi = \sum_{k\in S} d_k$ , and the calibration estimator (4.6) reduces to the Hàjek estimator

$$\hat{t}_{y,C} = N\frac{\hat{t}_{y,\pi}}{\hat{N}_\pi}.$$

(ii) Suppose that a quantitative variable $x_k$ is available for all $k \in S$ and the population total of the $x$-values is known. Setting $q_k = x_k^1$ in (4.5) leads to $\tilde{w}_k = d_k(t_x/\hat{t}_{x,\pi})$ and the calibration estimator (4.6) reduces to the ratio estimator

$$\hat{t}_{y,C} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} t_x.$$

(iii) Suppose that a categorical variable (e.g., gender, age group, etc.) with $J$ categories is available for all the sample units and that the vector of population counts $\mathbf{N} = (N_1, \ldots, N_j, \ldots, N_J)^\top$ is known, where $N_j$ denotes the number of units in the population belonging to the $j$th category. Let $\mathbf{x}_k = (x_{1k}, \ldots, x_{Jk})^\top$ be the $J$-vector attached to unit $k$ such that $x_{jk} = 1$ if unit $k$ belongs to the $j$th category and $x_{jk} = 0$, otherwise. Setting $q_k$ in (4.5) to $q_j$ if unit $k$ belongs to the $j$th category leads to the post-stratified weights

$$\tilde{w}_k = d_k \frac{N_j}{\hat{N}_{j,\pi}} \quad \text{if unit } k \text{ belongs to the } j\text{th category,} \tag{4.7}$$

where $\hat{N}_{j,\pi} = \sum_{k\in S\cap U_j} d_k$ represents the estimated number of individuals belonging to the $j$th category and $U_j$ is the set of population units belonging to the $j$th category. The calibration estimator (4.6) reduces to the post-stratified estimator

$$\hat{t}_{y,C} = \sum_{j=1}^{J} \frac{N_j}{\hat{N}_{j,\pi}} \hat{t}_{jy,\pi},$$

where $\hat{t}_{jy,\pi} = \sum_{k\in S\cap U_j} d_k y_k$. Let $n_j$ be the size of $S \cap U_j$, $j = 1, \ldots, J$. The $n_j$'s are random variables, which implies that some post-strata could be very small

or empty; that is, $n_j = 0$ for some $j$. In such cases, it is recommended to use some form of collapsing in order to reduce the number of post-strata; see, for example, Fuller (1966).

4.1.2 *The linear method.* This method is based on the generalized chi-square distance function

$$G(\tilde{w}_k/d_k) = \frac{1}{2}\left(\frac{\tilde{w}_k}{d_k} - 1\right)^2.$$

The resulting calibrated weights are given by

$$\begin{aligned}\tilde{w}_k &= d_k(1 + q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k) \\ &= d_k\left\{1 + q_k(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})^\top \left(\sum_{k\in S} d_k \mathbf{x}_k q_k \mathbf{x}_k^\top\right)^{-1} \mathbf{x}_k\right\}.\end{aligned}$$

With the linear method, there always exists a solution $\hat{\boldsymbol{\lambda}}$ that satisfies (4.4). However, some weights may be negative. Negative weights tend to occur when (i) the size $J$ of the vector $\mathbf{x}$ is large, which translates into a large number of calibration constraints and/or (ii) when $\hat{t}_{x_j,\pi}$ is far from $t_{x_j}$ for some $j = 1, \ldots, J$. Although a few negative weights may have little impact on the estimates, their presence on the data file is usually deemed unacceptable by data users because a negative weight is hard to interpret. In some instances, negative weights could lead to negative estimates of totals or means of intrinsically nonnegative variables.

The resulting calibration estimator is

$$\hat{t}_{y,C} = \sum_{k\in S} \tilde{w}_k y_k = \hat{t}_{y,\pi} + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})^\top \hat{\mathbf{B}}, \tag{4.8}$$

where

$$\hat{\mathbf{B}} = \left(\sum_{k\in S} d_k \mathbf{x}_k q_k \mathbf{x}_k^\top\right)^{-1} \sum_{k\in S} d_k \mathbf{x}_k q_k y_k.$$

The estimator (4.8) is the generalized regression estimator, which has been widely studied in the context of model-assisted estimation; see Särndal et al. (1992), Chapter 6 and Breidt and Opsomer (2016) for discussions of the generalized regression estimator.

4.1.3 *The exponential method.* This method is based on the Kullback–Leibler information distance

$$G(\tilde{w}_k/d_k) = \frac{\tilde{w}_k}{d_k}\log\frac{\tilde{w}_k}{d_k} - \frac{\tilde{w}_k}{d_k} + 1.$$

The resulting calibrated weights are given by

$$\tilde{w}_k = d_k \exp(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k). \tag{4.9}$$

TABLE 1
*Population counts in the case of two categorical variables*

| $x_1 \backslash x_2$ | 1 | 2 | ... | $J_2$ | Margin |
|---|---|---|---|---|---|
| 1 | $N_{11}$ | $N_{12}$ | ... | $N_{1J_2}$ | $N_{1\bullet}$ |
| 2 | $N_{21}$ | $N_{22}$ | ... | $N_{2J_2}$ | $N_{2\bullet}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $J_1$ | $N_{J_1 1}$ | $N_{J_1 2}$ | ... | $N_{J_1 J_2}$ | $N_{J_1\bullet}$ |
| Margin | $N_{\bullet 1}$ | $N_{\bullet 2}$ | ... | $N_{\bullet J_2}$ | |

Although the calibrated weights (4.9) are always positive, some may be extreme, which can cause the resulting estimator to be unstable especially if the calibration variables are poorly related to the $y$-variable. Also, a solution to (4.4) may not exist.

One special case of the exponential method arises in the context of two categorical variables (e.g., gender and age group) with $J_1$ and $J_2$ distinct categories, respectively. Conceptually, the population $U$ is partitioned into $J_1 \times J_2$ cells, $U_{j_1 j_2}$, of size $N_{j_1 j_2}$, $j_1 = 1, \ldots, J_1$; $j_2 = 1, \ldots, J_2$. If the cell counts $N_{j_1 j_2}$ are known, then we can construct a vector $\mathbf{x}_k$ of size $J_1 \times J_2$ consisting of $J_1 \times J_2 - 1$ entries equal to zero and a single entry equal to 1, which identifies the cell to which the unit $k$ belongs. This situation is a special case of post-stratification discussed in Section 4.1.1. Often, the individual cell counts are unknown, in which case post-stratification is not feasible. However, the marginal counts $N_{j_1\bullet} = \sum_{j_2=1}^{J_2} N_{j_1 j_2}$ and $N_{\bullet j_2} = \sum_{j_1=1}^{J_1} N_{j_1 j_2}$ are available and can be incorporated in the estimation procedure. Let $\hat{N}_{j_1 j_2,\pi} = \sum_{k\in S_{j_1 j_2}} d_k$ be the estimated counts in cell $(j_1, j_2)$ based on the basic weighting system, where $S_{j_1 j_2} = S \cap U_{j_1 j_2}$; see Tables 1 and 2.

It is virtually certain that the estimated marginal counts based on the basic design weights will not

TABLE 2
*Estimated counts in the case of two categorical variables based on the basic weighting system*

| $x_1 \backslash x_2$ | 1 | 2 | ... | $J_2$ | Margin |
|---|---|---|---|---|---|
| 1 | $\hat{N}_{11,\pi}$ | $\hat{N}_{12,\pi}$ | ... | $\hat{N}_{1J_2,\pi}$ | $\hat{N}_{1\bullet,\pi}$ |
| 2 | $\hat{N}_{21,\pi}$ | $\hat{N}_{22,\pi}$ | ... | $\hat{N}_{2J_2,\pi}$ | $\hat{N}_{2\bullet,\pi}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $J_1$ | $\hat{N}_{J_1 1,\pi}$ | $\hat{N}_{J_1 2,\pi}$ | ... | $\hat{N}_{J_1 J_2,\pi}$ | $\hat{N}_{J_1\bullet,\pi}$ |
| Margin | $\hat{N}_{\bullet 1,\pi}$ | $\hat{N}_{\bullet 2,\pi}$ | ... | $\hat{N}_{\bullet J_2,\pi}$ | |

match the known population marginal counts; that is, $\hat{N}_{j_1\bullet,\pi} \neq N_{j_1\bullet}$ and $\hat{N}_{\bullet j_2,\pi} \neq N_{\bullet j_2}$. Let $\delta_{j_1\bullet k}$ be the row indicator for unit $k$ such that

$$\delta_{j_1\bullet k} = \begin{cases} 1, & \text{if } k \text{ belongs to row } j_1, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, define the column indicator $\delta_{\bullet j_2 k}$ as

$$\delta_{\bullet j_2 k} = \begin{cases} 1, & \text{if } k \text{ belongs to column } j_2, \\ 0, & \text{otherwise.} \end{cases}$$

We define $\mathbf{x}_k = (\delta_{1\bullet k}, \ldots, \delta_{J_1\bullet k}, \delta_{\bullet 1k}, \ldots, \delta_{\bullet J_2 k})^\top$ of size $J_1 + J_2$. The vector of known marginal population counts is given by

$$\mathbf{t}_\mathbf{x} = \sum_{k\in U} \mathbf{x}_k = (N_{1\bullet}, \ldots, N_{J_1\bullet}, N_{\bullet 1}, \ldots, N_{\bullet J_2})^\top.$$

Under this set-up, the calibration weights (4.9) reduces to the raking ratio weights and are identical to those generated by the iterative proportional fitting procedure of Deming and Stephan (1940). The calibration estimator

$$\hat{t}_{y,C} = \sum_{k\in S} d_k \exp(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k) y_k$$

reduces to the well-known raking ratio estimator. When there are more than two categorical calibration variables, the resulting estimator is often referred to as the generalized raking estimator; for example, Deville et al. (1993).

4.1.4 *The empirical likelihood method.* This method is based on the inverse Kullback–Leibler information distance

$$G(\tilde{w}_k/d_k) = \frac{\tilde{w}_k}{d_k} \log \frac{\tilde{w}_k}{d_k} - \frac{\tilde{w}_k}{d_k} + 1.$$

The resulting calibrated weights are given by

$$\tilde{w}_k = \frac{d_k}{1 - q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k}. \tag{4.10}$$

Although the calibrated weights (4.10) are always positive, some may be extreme when $\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k$ is close to $q_k^{-1}$. Also, a solution to (4.4) may not exist. The weight (4.10) is often encountered in the context of empirical likelihood estimation; see, for example, Wu and Rao (2006).

4.1.5 *The truncated linear and the logit methods.* The truncated linear and the logit methods can be used in order to ensure that the calibration adjustment factors lie between pre-specified lower and upper bounds. The use of these methods may help avoiding the occurrence of negative or extreme weights. The truncated linear method is based on the following distance function:

$$G(\tilde{w}_k/d_k) = \begin{cases} \dfrac{1}{2}\left(\dfrac{\tilde{w}_k}{d_k} - 1\right)^2, & L < \dfrac{\tilde{w}_k}{d_k} < M, \\ \infty, & \text{otherwise,} \end{cases}$$

for constants $L$ and $M$, representing the lower and upper bounds, respectively. The corresponding calibration function is given by

$$F(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k) = \begin{cases} 1 + q_k \boldsymbol{\lambda}^\top \mathbf{x}_k, & (L-1) \leq q_k \boldsymbol{\lambda}^\top \mathbf{x}_k \leq (M-1), \\ M, & q_k \boldsymbol{\lambda}^\top \mathbf{x}_k > (M-1), \\ L, & q_k \boldsymbol{\lambda}^\top \mathbf{x}_k < (L-1). \end{cases}$$

The logit method is based on the following distance function:

$$G(\tilde{w}_k/d_k) = \begin{cases} \left(a_k \log \dfrac{a_k}{1-L} + b_k \log \dfrac{b_k}{M-1}\right)\dfrac{d_k}{A}, & \\ \qquad L < \dfrac{\tilde{w}_k}{d_k} < M, & \\ \infty, \quad \text{otherwise,} & \end{cases}$$

where

$$a_k = \frac{\tilde{w}_k}{d_k} - L, \quad b_k = M - \frac{\tilde{w}_k}{d_k},$$
$$A = \frac{M-L}{(1-L)(M-1)}.$$

The corresponding calibration function is given by

$$F(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k) = \frac{L(M-1) + M(1-L)\exp(A q_k \boldsymbol{\lambda}^\top \mathbf{x}_k)}{M - 1 + (1-L)\exp(A q_k \boldsymbol{\lambda}^\top \mathbf{x}_k)}. \tag{4.11}$$

Examples of intervals $[L, M]$ used in practice include $[1/2, 2]$, $[1/3, 3]$, $[1/4, 4]$. Imposing bounds restricts the solution space, and hence lowers the possibility of the existence of a solution to the calibration equations. Therefore, the tighter the interval the less of a chance that a solution exists. If an interval is too small, the weights will have a tendency to amass at the bounds if a solution exists. The larger the sample size $n$, the

more restrictive the chosen interval can be. An examination of the distribution of $\tilde{w}_k/d_k$ is needed in order to identify the most appropriate interval. Folsom and Singh (2000) proposed an extension of (4.11) to allow for nonuniform bounds $(L, M)$ for different subgroups of weights.

4.1.6 *Choice of the distance function*. The calibration function is usually chosen so that the distribution of the calibration factors $\tilde{w}_k/d_k$ is "cosmetically attractive", which we characterize as a distribution with no negative or extreme weights and that does not amass at the bounds. In practice, we suggest following these steps for choosing the calibration function: Start with the linear method and check whether or not there are negative calibration factors. If any, use the exponential method or the empirical likelihood method that ensure positive weights. Check whether or not the selected method has generated some extreme weights. If so, use any method that permits to bound the weights such as the truncated linear function or the logit function in order to ensure that the calibration adjustment factors lie between pre-specified lower and upper bounds.

### 4.2 Choice of Auxiliary Variables

At the estimation stage, there are two conflicting requirements. On the one hand, survey managers may be tempted to ensure consistency between survey estimates and known totals for a large vector $\mathbf{x}$ of auxiliary variables. On the other hand, survey statisticians strive for efficient point estimators. If the number of auxiliary variables is large, the resulting calibrated weights are likely to be highly dispersed, leading to unstable estimates for characteristics of interest poorly related to the calibration variables. To overcome this problem, Silva and Skinner (1997) suggested performing variable selection procedures to identify the calibration variables that are related to the characteristics of interest (or at least, a subgroup). They proposed a forward selection procedure that consists of starting with a noncalibrated estimator and adding the calibration variable that leads to a calibrated estimator with the smallest estimated mean square error. This procedure is repeated until the mean square error starts to increase, at which point the subset of the calibration variables leading to the smallest estimated mean square error, is selected. An alternative is ridge calibration, whereby a cost is assigned to each calibration constraint. On the one hand, assigning a large cost to a given constraint means that it is desired to stay close to the calibration constraint. In the extreme case, one can assign an infinite cost to a given constraint if one wishes to exactly satisfy that constraint. On the other hand, with a small cost, one is willing to tolerate a potentially large discrepancy between the survey estimate and the corresponding population total. In the extreme case, one can assign a cost value of zero in order to discard a constraint. Ridge calibration can also be used to satisfy pre-specified lower and upper bounds on the weights. The reader is referred to Chambers (1996), Beaumont and Bocci (2008) and Breidt and Opsomer (2016) for a discussion of ridge calibration.

### 4.3 Calibration for Two-Stage Sampling

In the context of two-stage sampling, some complexity arises because two types of sampling units are considered: the primary sampling unit (p.s.u.) such as a household or a business and the secondary sampling unit (s.s.u.) such as an individual or an employee. This structure generates two levels of auxiliary information. For instance, in a survey of individuals, we may have auxiliary information at the household level such as the number of individual in the household, the number of individuals in each age group and the owner/renter status. We may also have auxiliary information at the individual level such as gender, age group of each individual in the household and profession of each individual. The reader is referred to Estevao and Särndal (2006) for an excellent discussion of calibration procedures in the context of two-stage/phase sampling. In the presence of two-level information, it seems desirable to simultaneously calibrate on known totals both at the p.s.u. and s.s.u. levels.

Many surveys using two-stage sampling serve a double objective: (i) provide estimates of totals for the population of s.s.u.'s; and (ii) provide estimates of totals for the population of p.s.u.'s. Examples of characteristics of interest at the s.s.u. level include *individual income* and *job status*. We denote these variables by $y_{(e)}$. Examples of characteristics of interest at the p.s.u. level include *household income* and *household spending*. We denote these variables by $y_{(c)}$. The interest lies in estimating two types of totals: $t_{y_{(e)}} = \sum_{i=1}^{M} \sum_{k=1}^{N_i} y_{(e)k}$ and $t_{y_{(c)}} = \sum_{i=1}^{M} y_{(c)i}$, where $N_i$ is the size of the $i$th p.s.u., $i = 1, \ldots, M$.

The design weight of the $k$th s.s.u. in the $i$th p.s.u. is given by

$$d_k = \pi_i^{-1} \pi_{k|i}^{-1} \tag{4.12}$$

if the $k$th s.s.u. belongs to the $i$th p.s.u.;

see Section 2.1. Expansion types estimators of $t_{y_{(e)}}$ and $t_{y_{(c)}}$ are given by

$$\hat{t}_{y_{(e)},\pi} = \sum_{i\in S_1}\sum_{k\in S_i} d_k y_{(e)k}, \quad \hat{t}_{y_{(c)},\pi} = \sum_{i\in S_1} \pi_i^{-1} y_{(c)i},$$

respectively.

Let $\mathbf{x}_{1i}$ be a vector of p.s.u.-level auxiliary variables of size $J_1$ such that $\mathbf{x}_{1i}$ is observed for all $i \in S_1$ and $\sum_{i=1}^{M} \mathbf{x}_{1i}$ is known. Let $\mathbf{x}_{2k}$ be a vector of s.s.u.-level auxiliary variables of size $J_2$ such that $\mathbf{x}_{2k}$ is observed for all $k \in \bigcup_{i=1}^{M} S_i$ and $\sum_{i=1}^{M}\sum_{k=1}^{N_i} \mathbf{x}_{2k}$ is known. Let $\tilde{w}_i$ be the calibrated weight for the $i$th p.s.u. and $\tilde{w}_k$ be the calibrated weight for the $k$th s.s.u.

Mimicking (4.12), we impose the following relationship between $\tilde{w}_i$ and $\tilde{w}_k$:

$$\tilde{w}_k = \tilde{w}_i \pi_{k|i}^{-1} \quad \text{if the } k\text{th element belongs to the } i\text{th cluster.} \tag{4.13}$$

When a convenient relationship such as (4.13) is imposed, we have integrated weighting, a term coined by Lemaître and Dufour (1987). Estevao and Särndal (2006) discuss alternatives to (4.13). The effect of integrated weighting on the efficiency of calibration estimators was discussed in Steel and Clark (2007).

The calibration constraints are given by

$$\sum_{i\in S_1} \tilde{w}_i \mathbf{x}_{1i} = \sum_{i=1}^{M} \mathbf{x}_{1i}, \tag{4.14}$$

$$\sum_{i\in S_1}\sum_{k\in S_i} \tilde{w}_k \mathbf{x}_{2k} = \sum_{i=1}^{M}\sum_{k=1}^{N_i} \mathbf{x}_{2k}. \tag{4.15}$$

Thanks to (4.13), the constraints (4.14) and (4.15) can be compactly written as

$$\sum_{i\in S_1} \tilde{w}_i \begin{pmatrix} \mathbf{x}_{1i} \\ \hat{\mathbf{x}}_{2i} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{M} \mathbf{x}_{1i} \\ \sum_{i=1}^{M}\sum_{k=1}^{N_i} \mathbf{x}_{2i} \end{pmatrix}, \tag{4.16}$$

where

$$\hat{\mathbf{x}}_{2i} = \sum_{k\in S_i} \frac{\mathbf{x}_{2k}}{\pi_{k|i}}.$$

From (4.16), the calibration procedure can be realized in a single step. The goal is to obtain a set of calibrated weights $\tilde{w}_i$ as close as possible to the design weights $\pi_i^{-1}$ subject to (4.16). Any calibration method presented in Section 4.1 can be used. Calibration estimators of $t_{y_{(e)}}$ and $t_{y_{(c)}}$ are given by

$$\hat{t}_{y_{(e)},C} = \sum_{i\in S_1}\sum_{k\in S_i} \tilde{w}_k y_{(e)k}, \quad \hat{t}_{y_{(c)},C} = \sum_{i\in S_1} \tilde{w}_i y_{(c)i}.$$

REMARK 4.1. If all the elements in a selected p.s.u. are selected, then $\pi_{k|i} = 1$ and the weight of all the s.s.u.'s belonging to the p.s.u. are all equal to the weight of that p.s.u. This is a desirable property.

REMARK 4.2. Nonintegrated weighting consists of performing two independent calibration without imposing a relationship such as (4.13). That is, starting with the weights $\pi_i^{-1}$ as the initial weights, compute the calibration weights $\tilde{w}_i$ subject to (4.14). In a second independent calibration, starting with the weights $d_k$, compute the calibration weights $\tilde{w}_k$ subject to (4.15). A consequence of nonintegrated calibration is the potential inconsistency between estimates at the element level obtained using the sets of weights: $\tilde{w}_i \pi_{k|i}^{-1}$ and $\tilde{w}_k$. That is,

$$\sum_{i\in S_1}\sum_{k\in S_i} \tilde{w}_i \pi_{k|i}^{-1} \mathbf{x}_{2k} \neq \sum_{i\in S_1}\sum_{k\in S_i} \tilde{w}_k \mathbf{x}_{2k} = \sum_{i=1}^{M}\sum_{k=1}^{N_i} \mathbf{x}_{2k},$$

which is somehow embarrassing.

### 4.4 Calibration for Two-Phase Sampling Designs

As in two-stage sampling, there are two levels of auxiliary information in two-phase sampling designs. The notation in this section follows closely the one in Estevao and Särndal (2006). We distinguish between two types of auxiliary variables: (i) a vector of auxiliary variables $\mathbf{x}_k^*$ of size $J^*$ and (ii) a vector of auxiliary variables $\mathbf{x}_k^o$ of size $J^o$. The availability of the auxiliary information is as follows: (a) For every $k \in S_1$, we observe $\mathbf{x}_k^o$. (b) For every $k \in S_2$, we observe both $\mathbf{x}_k^*$ and $\mathbf{x}_k^o$. (c) The vector of population totals, $\sum_{k\in U} \mathbf{x}_k^*$, is known but the vector of population totals $\sum_{k\in U} \mathbf{x}_k^o$ is unknown. Available to us is the vector of estimated totals $\sum_{k\in S_1} \pi_{1k}^{-1} \mathbf{x}_k^o$.

A vector of auxiliary variables, $\mathbf{x}_k$, of size $J = J^* + J^o$, is constructed by stacking the auxiliary information to obtain

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}.$$

There exist at least three alternatives to obtain the calibrated weights $\tilde{w}_k$. Here, we focus on the so-called single-step calibration; see Estevao and Särndal (2006) for alternative procedures. Starting with the initial weights $\tilde{d}_k = 1/\{\pi_{1k}\pi_{2k}(\mathbf{I}_1)\}$ (see Section 3.4), compute the calibrated weights $\tilde{w}_k$ for $k \in S_2$ subject to the calibration constraints

$$\sum_{k\in S_2} \tilde{w}_k \mathbf{x}_k = \begin{pmatrix} \sum_{k\in U} \mathbf{x}_k^* \\ \sum_{k\in S_1} \pi_{1k}^{-1} \mathbf{x}_k^o \end{pmatrix}.$$

Any calibration method covered in Section 4.1 can be used. The resulting calibration estimator of $t_y$ is given by

$$\hat{t}_{y,C} = \sum_{k \in S_2} \tilde{w}_k y_k.$$

The material presented in this section provides a platform for single-step weighting often used for treating unit nonresponse; see Section 5.

### 4.5 Model Calibration

We argued that the conventional calibration estimators (see Section 4.1) are efficient when the $y$-variable is linearly related to $\mathbf{x}$ and the relationship is strong. The procedures presented previously are thus efficient when a linear regression model fits the data well. If the $y$-variable is not continuous, a linear regression model may not be appropriate and the resulting calibration may be inefficient. To cope with this problem, Wu and Sitter (2001) introduced the concept of model calibration, which may improve the efficiency of estimators when the relationship between $y$ and $\mathbf{x}$ is not well described by a linear regression model. Suppose that the relationship between $y$ and $\mathbf{x}$ can be modelled by

$$E(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\beta}), \quad V(y_k|\mathbf{x}_k) = \sigma^2 c_k,$$

where $\mu(\cdot, \boldsymbol{\beta})$ is a known function, $\boldsymbol{\beta}$ is a vector of unknown coefficients, $\sigma^2$ is an unknown parameter and $c_k$ is a known factor attached to unit $k$. The above model includes two important special cases: (i) linear regression models and (ii) generalized linear and other nonlinear regression models. Let $\mu(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$ denote the fitted values $k = 1, \ldots, N$, where $\hat{\boldsymbol{\beta}}$ is a design-consistent estimator of $\boldsymbol{\beta}$. Wu and Sitter (2001) proposed the following model calibration estimator of $t_y$:

$$\hat{t}_{y,\mathrm{MC}} = \sum_{k \in S} \tilde{w}_k y_k,$$

where the calibrated weights $\tilde{w}_k$ are as close as possible to the design weights $d_k$ subject to

$$\sum_{k \in S} \tilde{w}_k = N, \qquad \sum_{k \in S} \tilde{w}_k \mu(\mathbf{x}_k, \hat{\boldsymbol{\beta}}) = \sum_{k \in U} \mu(\mathbf{x}_k, \hat{\boldsymbol{\beta}}).$$

Once again, any calibration method $F(\cdot)$ described in Section 4.1 can be used. The estimator $\hat{t}_{y,\mathrm{MC}}$ is design-consistent for $t_y$ irrespective of whether the model holds.

Wu (2003) showed that the method of Wu and Sitter (2001) is optimal among a class of calibration estimators in the sense of minimising the expected asymptotic design variance under a superpopulation model and the sampling design.

While conventional calibration methods require the auxiliary information to be recorded for the sample units only and population quantities imported from external sources, model calibration requires $\mathbf{x}_k$ to be recorded for all $k \in U$, which is often referred to as complete auxiliary information. This is a fairly restrictive requirement because most often, the values of $\mathbf{x}_k$ are not available individually for the nonsample units. Montanari and Ranalli (2005) extended the model calibration procedure of Wu and Sitter (2001) to handle nonparametric models.

## 5. WEIGHTING SYSTEM ADJUSTED FOR UNIT NONRESPONSE

The main issue with nonresponse is the bias introduced when the respondents are different from the nonrespondents with respect to the survey variables. An additional component of variance is also added due to the observed sample size, $n_r$, that is smaller than the initially planned sample size, $n$. The key to reducing both nonresponse bias and variance is to use nonresponse weighting methods that take advantage of auxiliary information available for both respondents and nonrespondents.

### 5.1 The Double Expansion Estimator

Unit nonresponse is often viewed as a second phase of sampling (e.g., Särndal and Swensson, 1987) with the main difference that the response mechanism is unknown. Let $r_k$, $k \in S$, be response indicators such that $r_k = 1$, if unit $k$ responds, and $r_k = 0$, otherwise, and let $\mathbf{R}_S$ be the vector containing $r_k$ in its $k$th element. Note that, in Section 2.2, we used the notation $\mathbf{I}_2$ instead of $\mathbf{R}_S$ when we introduced two-phase sampling. We use a different notation in this section to emphasize that nonresponse is a special second phase of sampling. The set of respondents, denoted by $S_r$, is the subset of $S$ which contains all the units $k \in S$ such that $r_k = 1$. The response mechanism is the distribution of $\mathbf{R}_S$ given $\mathbf{I}$. By analogy with sampling, we may assume that the true unknown response mechanism depends only on a certain vector of variables $\tilde{\mathbf{v}}_k$, $k \in S$. We denote by $\tilde{\mathbf{V}}_S$, the matrix containing $\tilde{\mathbf{v}}_k^\top$ in its $k$th row. The response mechanism can thus be denoted by $\mathcal{F}(\mathbf{R}_S|\mathbf{I}, \tilde{\mathbf{V}}_S)$ and the response probability by $p_k = P(r_k = 1|\mathbf{I}, \tilde{\mathbf{V}}_S)$. If the response probabilities $p_k$ were known and greater than 0 for all $k \in S$, the double expansion estimator

could be used. Using the above notation, it is written as

$$\hat{t}_{y,\mathrm{DE}} = \sum_{k\in S_r} \tilde{d}_k y_k = \sum_{k\in S} r_k \tilde{d}_k y_k,$$

where $\tilde{d}_k = d_k a_k$ and $a_k = 1/p_k$ is a nonresponse weight adjustment for unit $k$. The double expansion estimator is conditionally unbiased in the sense that $E(\hat{t}_{y,\mathrm{DE}} - \hat{t}_{y,\pi}|\mathbf{I}, \boldsymbol{\Omega}) = 0$, where $\boldsymbol{\Omega}$ is here any quantity involved in the inference but the vectors $\mathbf{I}$ and $\mathbf{R}_S$. As a result, the double expansion estimator is also unconditionally unbiased for $t_y$. This follows from the unbiasedness property of the simple expansion estimator $\hat{t}_{y,\pi}$.

### 5.2 The Empirical Double Expansion Estimator

The response probabilities $p_k$ are never known in practice. This problem has often been addressed in the literature by determining a model for the response indicators $r_k$, called a response model, and then obtaining the estimated probabilities $\hat{p}_k$ using the model (e.g., Särndal and Swensson, 1987; Ekholm and Laaksonen, 1991). This approach leads to the empirical double expansion estimator

$$\hat{t}_{y,\mathrm{EDE}} = \sum_{k\in S_r} \hat{\tilde{d}}_k y_k = \sum_{k\in S} r_k \hat{\tilde{d}}_k y_k,$$

where $\hat{\tilde{d}}_k = d_k \hat{a}_k$ and $\hat{a}_k = 1/\hat{p}_k$ is the nonresponse weight adjustment for unit $k$. The empirical double expansion estimator is also known as the propensity score adjusted estimator. Beaumont (2005a) and Kim and Kim (2007) showed that estimating response probabilities often leads to a smaller nonresponse variance than using the true response probabilities; that is, often we have $V(\hat{t}_{y,\mathrm{EDE}}|\mathbf{I}, \boldsymbol{\Omega}) \leq V(\hat{t}_{y,\mathrm{DE}}|\mathbf{I}, \boldsymbol{\Omega})$. As shown below, the estimation of response probabilities may lead to an implicit calibration of weights, which stabilizes the estimators of totals.

The modelling task normally involves two main steps. The first step consists of selecting a vector of explanatory variables, $\mathbf{v}_k$, that are predictive of the response indicator $r_k$. The goal of that step is to find $\tilde{\mathbf{v}}_k$ or at least a good proxy for $\tilde{\mathbf{v}}_k$. The use of explanatory variables that are highly predictive of response tends to yield some small estimated response probabilities, and thus large nonresponse weight adjustments $\hat{a}_k$. This may lead to an unstable empirical double expansion estimator. Little and Vartivarian (2005) and Beaumont (2005b) argued that $\mathbf{v}_k$ should contain explanatory variables that are related to both the response indicator and the variables of interest. Explanatory variables that are related only to the response indicator but not to any variable of interest should not be used in the estimation of response probabilities as they do not contribute to reducing the nonresponse bias of the empirical double expansion estimator and may increase substantially its nonresponse variance, $V(\hat{t}_{y,\mathrm{EDE}}|\mathbf{I}, \boldsymbol{\Omega})$, especially when there are some relatively large nonresponse weight adjustments. The second modelling step consists of determining a suitable model for the relationship between the response indicator $r_k$ and the selected explanatory variables, $\mathbf{v}_k$. In addition, the response indicators are typically assumed to be mutually independent. The latter assumption may be violated in practice in the context of two-stage sampling designs because sample units within the same cluster (e.g., household) may not respond independently of one another. The reader is referred to Skinner and D'Arrigo (2011) and Kim et al. (2016) for a discussion of estimation procedures accounting for the possible intra-cluster correlation. The next two subsections are devoted to the modelling and estimation of response probabilities $p_k$ under this independence assumption.

### 5.3 Estimation of Response Probabilities Under a Missing at Random Response Model

Let us start by assuming that $p_k = P(r_k = 1|\mathbf{I}, \tilde{\mathbf{V}}_S) = P(r_k = 1|I_k = 1, \mathbf{v}_k)$ and that $\mathbf{v}_k$, $k \in S$, do not contain any missing values. In that case, the missing $y$ values are said to be missing at random (see Rubin, 1976). A simple parametric response model is the logistic regression model

$$p_k = \rho(\mathbf{v}_k; \boldsymbol{\alpha}) = \frac{\exp(\mathbf{v}_k^\top \boldsymbol{\alpha})}{1 + \exp(\mathbf{v}_k^\top \boldsymbol{\alpha})},$$

$k \in S$, where $\rho$ is the logistic function and $\boldsymbol{\alpha}$ is a vector of unknown model parameters. We focus on the logistic function in this paper because it is widely used in practice. There are of course other functions $\rho$ that may be suitable for the modelling of response probabilities and that can be found in standard textbooks on categorical data. The estimator of the response probability $p_k$ is denoted by $\hat{p}_k = \rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}$. There are many possible estimators of $\boldsymbol{\alpha}$. The most natural one is obtained through maximum likelihood estimation. The maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$ satisfies the equation

$$\sum_{k\in S} \big(r_k - \rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}})\big)\mathbf{v}_k = \mathbf{0}.$$

Iannacchione et al. (1991) suggested another estimating equation. Their estimator $\hat{\boldsymbol{\alpha}}$ satisfies the equation

$$\sum_{k\in S} d_k \frac{(r_k - \rho(\mathbf{v}_k;\hat{\boldsymbol{\alpha}}))}{\rho(\mathbf{v}_k;\hat{\boldsymbol{\alpha}})}\mathbf{v}_k = \mathbf{0}.$$

This estimating equation is unbiased for $\boldsymbol{\alpha}$ provided that the design weight $d_k$ is independent of the response indicator $r_k$ after conditioning on $\mathbf{v}_k$. This implies that the design weight should be considered as a potential explanatory variable. The estimating equation suggested by Iannacchione et al. (1991) has a nice calibration property since it can be easily shown that the resulting weight adjustment $\hat{a}_k = 1/\hat{p}_k$ satisfies the calibration equation

$$\sum_{k\in S_r} (d_k\hat{a}_k)\mathbf{v}_k = \sum_{k\in S} d_k\mathbf{v}_k.$$

This is a calibration constraint with benchmark totals estimated from the full sample $S$. For the $\mathbf{v}$ variables, the nonresponse error completely disappears. We may thus expect a low nonresponse bias and variance for $y$ variables that are highly correlated with $\mathbf{v}$.

We might also considered the following unweighted version of the estimating equation proposed by Iannacchione et al. (1991):

$$\sum_{k\in S} \frac{(r_k - \rho(\mathbf{v}_k;\hat{\boldsymbol{\alpha}}))}{\rho(\mathbf{v}_k;\hat{\boldsymbol{\alpha}})}\mathbf{v}_k = \mathbf{0}.$$

The resulting weight adjustments satisfy the following calibration equation:

$$\sum_{k\in S_r} (d_k\hat{a}_k)(\pi_k\mathbf{v}_k) = \sum_{k\in S} d_k(\pi_k\mathbf{v}_k).$$

Should the design weights $d_k$ be used when estimating $\boldsymbol{\alpha}$? It seems to really depend on whether it is preferable to use the calibration variables $\mathbf{v}_k$ or the calibration variables $\pi_k\mathbf{v}_k$. Although there is no definite answer to this question, calibration on $\mathbf{v}_k$ may be more meaningful in general unless the variables in $\mathbf{v}_k$ have already been multiplied by the design weight. This suggests that using a weighted estimating equation may be preferable in many cases.

There are two main practical issues associated with the use of a parametric model, such as the logistic model. First, the logistic function $\rho$ may not be appropriate for describing the relationship between the response indicator and the explanatory variables and there may be missing interactions in the model that were not detected during model selection. In other words, the model may fail, even though careful model selection and validation have been done, and we may want to find a procedure that is robust to model misspecifications. Second, the logistic model tends to yield some estimated response probabilities, $\hat{p}_k$, that are relatively small resulting in very large weight adjustments $\hat{a}_k$ and potentially unstable estimates. A possible solution to these issues is obtained through the creation of classes homogeneous with respect to the propensity to respond using the so-called score method (e.g., Little, 1986, Eltinge and Yansaneh, 1997; or Haziza and Beaumont, 2007). It can be implemented using the following steps: (i) Obtain estimated response probabilities, $\hat{p}_k^{\mathrm{LR}}$, $k \in S$, from a logistic regression. (ii) Order the sample from the lowest estimated response probability computed in step (i) to the largest. (iii) Form a certain number of classes homogeneous with respect to $\hat{p}_k^{\mathrm{LR}}$, $k \in S$. Classes of equal size can be formed or a clustering algorithm can be used. The number of classes should be as small as possible but large enough to capture most of the variability of $\hat{p}_k^{\mathrm{LR}}$, $k \in S$. (iv) Compute the final estimated response probability, $\hat{p}_k$, for a unit $k$ in some homogeneous class $c$ as the weighted (or unweighted) response rate within class $c$. Forming homogeneous classes using the above procedure provides some robustness to model misspecifications and is less prone to extreme weight adjustments than directly using $\hat{p}_k^{\mathrm{LR}}$. If the creation of classes does not remove all the extreme weight adjustments, then weight trimming or collapsing classes are possible solutions.

The above score method is one method of forming homogeneous classes. There are other methods such as the CHi square Automatic Interaction Detection (CHAID) algorithm developed by Kass (1980) and regression trees (Phipps and Toth, 2012). In stratified business surveys, classes are sometimes taken to be the strata for simplicity and because there may be no other explanatory variable available. Da Silva and Opsomer (2006, 2009) described a nonparametric regression method using kernel smoothing as an alternative to forming classes. Nonparametric methods are expected to provide some robustness if the form of $\rho(\cdot)$ is misspecified and protect (to some extent) against the noninclusion of predictors accounting for curvature or interactions.

### 5.4 Estimation of Response Probabilities Under a Missing Not at Random Response Model

We have assumed so far that the explanatory variables were observed for all sample units. There are cases where it might be more reasonable to assume that

the variables explaining response are indeed some of the survey variables, which are subject to missing values. This is a case where data are said to be missing not at random. An example would be a survey on income where people refusing to answer the survey are those who have a low or large income. Generalized calibration can be used in this context (e.g., Deville, 1998, Sautory, 2003; and Kott, 2006). To understand the idea, suppose that the vector of explanatory variables is $\mathbf{v}_k^\top = (\mathbf{v}_{1k}^\top \mathbf{v}_{2k}^\top)$, where $\mathbf{v}_{1k}$ is a vector of explanatory variables observed for all sample units and $\mathbf{v}_{2k}$ contain survey variables observed only for the respondents. The above logistic function can again be used to model the response probability $p_k = \rho(\mathbf{v}_k; \boldsymbol{\alpha})$ but the estimation of $\boldsymbol{\alpha}$ is more difficult because of missing values in $\mathbf{v}_k$. The solution obtained through generalized calibration makes use of a vector of instrumental variables, observed for all sample units, $\mathbf{v}_{2k}^I$, as a substitute for $\mathbf{v}_{2k}$. For efficiency considerations, the instrumental variables should be associated as much as possible with their corresponding variables in $\mathbf{v}_{2k}$. For validity considerations, they should also be independent of the response indicator after conditioning on $\mathbf{v}_k$. Generalized calibration is based on an estimating equation very similar to the one proposed in Iannacchione et al. (1991). The resulting estimator of $\boldsymbol{\alpha}$ satisfies the equation:

$$\sum_{k\in S} d_k \frac{(r_k - \rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}}))}{\rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}})} \mathbf{v}_k^I = \mathbf{0},$$

where $\mathbf{v}_k^{I\top} = (\mathbf{v}_{1k}^\top \mathbf{v}_{2k}^{I\top})$. An equivalent expression is

$$\sum_{k\in S_r} \frac{d_k}{\rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}})} \mathbf{v}_k^I - \sum_{k\in S} d_k \mathbf{v}_k^I = \mathbf{0}.$$

It is interesting to note that this equation does not use the values of $\mathbf{v}_{2k}$, $k \in S - S_r$, and that the resulting weight adjustments, $\hat{a}_k = [\rho(\mathbf{v}_k; \hat{\boldsymbol{\alpha}})]^{-1}$, satisfy the calibration equation:

$$\sum_{k\in S_r} (d_k \hat{a}_k) \mathbf{v}_k^I = \sum_{k\in S} d_k \mathbf{v}_k^I.$$

The explanatory variables in the response model are $\mathbf{v}_k$ and contain variables subject to missing values. The instrumental variables $\mathbf{v}_{2k}^I$ are also calibration variables like $\mathbf{v}_{1k}$. Chang and Kott (2008) considered the extension to the case where there are more calibration variables than model parameters to be estimated (i.e., the dimension of $\mathbf{v}_k^I$ is larger than the dimension of $\mathbf{v}_k$). Note that Deville (1998, 2002), Estevao and Särndal (2000), Kott (2006) and Chang and Kott (2008) defined $\mathbf{v}_k$ as the vector of instrumental variables. We prefer to call $\mathbf{v}_k$ the vector of explanatory variables of the response model as it better reflects its role. Our definition of instrumental variable is different than the previous authors but is in line with the econometric literature where instrumental variables have been used for decades (e.g., Kmenta, 1971).

In practice, generalized calibration and methods that handle data that are not missing at random are rarely used. The additional complexity and lack of software may be part of the explanation. More importantly, it seems almost impossible to validate the logistic function through model diagnostics because the explanatory variables are not observed for the nonrespondents. A score method cannot be used to obtain some robustness to model failures because the response probabilities cannot be estimated for the nonrespondents. Generalized calibration may thus yield extreme weight adjustments with the logistic function. Also, appropriate instrumental variables may not be easy to find in practice. All these reasons may explain why generalized calibration have not been used more often so far. This does not mean that generalized calibration should not be used. There are cases where it may be reasonable to assume that response depends directly on some of the survey variables and failure to account for that dependence would lead to significant nonresponse bias. This may happen in surveys collecting information on sensitive topics.

### 5.5 Nonresponse Weighting Through Calibration

We have seen earlier that nonresponse weight adjustment methods often lead to an implicit calibration. We may then wonder why not consider directly calibrating as in Section 4.4, without modelling the response indicator $r_k$. This calibration approach was advocated in Lundström and Särndal (1999), Folsom and Singh (2000) and Särndal and Lundström (2005). It is certainly a valid point of view. However, we prefer a response model approach because it allows us to justify the choice of calibration variables and the form of the weight adjustment necessary to cope with the nonresponse bias. A large number of model diagnostics are available in the literature to come up with a suitable weight adjustment. Of course, nothing precludes an additional calibration step after nonresponse weight adjustment to further increase the efficiency of the estimators, like in two-phase sampling.

## 6. WEIGHT TRIMMING AND WEIGHT SMOOTHING

As we argued throughout the paper, estimators may suffer from inefficiency when the weights are highly dispersed and exhibit a poor correlation with the characteristics of interest. Early references on the topic include Rao (1966) and Basu (1971). This issue is common in multipurpose surveys. In such cases, the final weights may undergo further modification through the use of weight trimming or weight smoothing methods in order to improve the efficiency of point estimators.

### 6.1 Weight Trimming

Units exhibiting a large weight are deemed harmful because they lead to highly dispersed weights, which can potentially cause the estimators to be unstable. The rationale behind weight trimming is to modify the survey weights so that the resulting estimators (called trimmed estimators) have a lower mean square error than that of commonly used estimators (e.g., expansion estimators, calibration estimators, etc.). This is generally achieved at the expense of introducing a bias.

A number of weight trimming methods have been considered in the literature. All rely on a specification of a threshold, $w_0$ above which a unit is identified as problematic. The idea is to force the weights above this threshold to this value. Let $w_k$ and $w_{k,\mathrm{t}}$ be the weights associated with unit $k$ before and after trimming, respectively. We have

$$w_{k,\mathrm{t}} = \begin{cases} w_0, & \text{if } w_k \geq w_0, \\ \gamma w_k, & \text{if } w_k < w_0, \end{cases} \tag{6.1}$$

where $\gamma$ is a rescaling factor that ensures that

$$\sum_{k\in S} w_{k,\mathrm{t}} = \sum_{k\in S} w_k.$$

That is, the excess weight is redistributed among the units below the threshold so that the sum of the weights after trimming is still a valuable estimate of the population size. The resulting trimmed estimator of $t_y$ is

$$\hat{t}_{y,\mathrm{trim}}(w_0) = \sum_{k\in S} w_{k,\mathrm{t}} y_k.$$

The choice of the threshold $w_0$ is very important because a bad choice may potentially lead to a trimmed estimator with a mean square error larger than that of the untrimmed estimator. The choice of $w_0$ has been discussed by Potter (1990); see also Chen et al. (2016) for a description of weight trimming procedures. If the trimmed weights $w_{k,\mathrm{t}}$ are modified through some form of calibration, some calibrated weights may be larger than the threshold $w_0$. An alternative to weight trimming followed by calibration consists of minimizing (4.2) subject to (4.1) and additional constraints on the weights.

### 6.2 Weight Smoothing

Weight smoothing was introduced by Beaumont (2008). The choice of a threshold for trimming weights may prove difficult in multipurpose surveys, since an appropriate threshold for one characteristic of interest may be inappropriate for another. Unlike weight trimming procedures, weight smoothing does not require the specification of a threshold. Gains in efficiency are achieved at the expense of introducing a model for the survey weights.

Our discussion focusses on the design weights but weight smoothing can naturally be applied to other types of weights (e.g., weights adjusted for nonresponse and final weights) or to nonresponse adjustment factors and calibration adjustment factors. Weight smoothing starts by postulating a model linking the design weights to the characteristics of interest:

$$d_k = f(\mathbf{y}_k; \boldsymbol{\gamma}) + \varepsilon_k, \tag{6.2}$$

where $\mathbf{y}_k = (y_{1k}, \ldots, y_{pk})^\top$ and $\boldsymbol{\gamma}$ is a $p$-vector of unknown coefficients. Model (6.2) suggests that a portion of the weight $d_k$ is explained by (some of) the characteristics of interest, while the other portion corresponds to random noise. The smoothed weights are given by $\hat{d}_k = f(\mathbf{y}_k; \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is a suitable estimator (e.g., the generalized least square estimator) of $\boldsymbol{\gamma}$. The rationale behind weight smoothing is to "get rid" of the portion of the weight $d_k$ corresponding to random noise while preserving the portion which is related to the characteristics of interest. The resulting smoothed expansion estimator is $\hat{t}_{y,\mathrm{smooth}} = \sum_{k\in S} \hat{d}_k y_k$ and is unbiased (or at least approximately unbiased) for $t_y$ in the sense that $E(\hat{t}_{y,\mathrm{smooth}}|\boldsymbol{\Omega}) = t_y$, where $\boldsymbol{\Omega}$ is used to denote any quantity involved in the inference except the vectors $\mathbf{I}$ and $\mathbf{Z}$. The properties of the smoothed estimator depend on the validity of the postulated model. Classical model selection and validation techniques can be used to determine an appropriate model. If (6.2) is a linear regression model, then $V(\hat{t}_{y,\mathrm{smooth}}|\boldsymbol{\Omega}) \leq V(\hat{t}_{y,\pi}|\boldsymbol{\Omega})$. That is, the smoothed estimator $\hat{t}_{y,\mathrm{smooth}}$ is more efficient than the expansion estimator. The reader is referred to Beaumont (2008) for more details on the theoretical properties of smoothed estimators.

## 7. DISCUSSION

Recently, the "culture" of a unique weighting system has been criticized by a number of authors, including Salgado et al. (2012) and Gelman (2007). In large-scale surveys, there is often a rich source of auxiliary variables, which could be used to produce tailor-made estimates for every parameter of interest. Such estimation procedures have the potential to be more efficient than a procedure based on a unique weighting system.

For specific purposes, we agree that tailor-made estimation procedures can be useful. This is customary in small area estimation (Rao and Molina, 2015), where it is usually not always possible to obtain estimates using a single set of weights. There are other instances for which using specialized estimation procedures may be desirable in terms of efficiency. It occurs, for example, when the $y$-variable contains a large number of zeroes (Kalberg, 2000) or when one is in the presence of a nonlinear relationship (Firth and Benett, 1998). Also, for the variable selection procedure of Silva and Skinner (1997) discussed in Section 4.2, the set of selected calibration variables is specific to a particular $y$-variable and the same set applied to another variable may lead to an unstable estimate. For most surveys, the amount of work required for developing specialized estimation procedures for the production of every estimate of interest seems to be a considerable task. For these reasons, a single set of weights seems to be necessary to obtain the vast majority of the estimates in a reasonable amount of time. In addition, a unique weighting system permits two intuitively appealing properties to be satisfied: (i) the internal consistency property and (ii) the external consistency property. These are discussed below.

Let $\{\tilde{w}_k; k \in S\}$ be the unique weighting system consisting of final weights. Assume we have $G$ characteristics of interest, $y_1, \ldots, y_G$, and that $y_G = \sum_{g=1}^{G-1} y_g$. For instance, $y_1$ and $y_2$ may represent the profit and costs of a business, respectively. Its turnover, $y_3$, can then be expressed as $y_3 = y_1 + y_2$. If we are interested in estimating the population totals $t_{y_g} = \sum_{k\in U} y_{gk}$, $g = 1, \ldots, G$, we have the relationship:

$$t_{y_G} = \sum_{k\in U} y_{Gk} = \sum_{k\in U} \sum_{g=1}^{G-1} y_{gk} = \sum_{g=1}^{G-1} t_{y_g}.$$

This property is often referred to as internal consistency. It seems desirable that this property also holds at the estimate level. With a single weighting system $\{\tilde{w}_k; k \in S\}$ applied to $y_1, \ldots, y_G$, we obtain $G$ estimates $\hat{t}_{y_g} = \sum_{k\in S} \tilde{w}_k y_{gk}$ satisfying the relationship

$$\hat{t}_{y_G} = \sum_{g=1}^{G-1} \hat{t}_{y_g}.$$

In this case, we say that the system of estimates $\{\hat{t}_{y_1}, \ldots, \hat{t}_{y_g}, \ldots, \hat{t}_{y_G}\}$ is internally consistent. Had different weighting systems been used for different variables, the internal consistency would have been violated.

We now turn to domain estimation. In virtually all surveys, estimates are needed for domains. A domain is any subpopulation for which we require an estimate. For instance, we may be interested in average income by gender or by age group. We assume that the population $U$ of size $N$ is divided into $D$ mutually disjoint domains of interest, $U_1, \ldots, U_d, \ldots, U_D$ of size $N_1, \ldots, N_d, \ldots, N_D$, respectively. We have $\bigcup_{d=1}^{D} U_d = U$ and $\sum_{d=1}^{D} N_d = N$. Let $S_d = S \cap U_d$ denote the set of sample units falling in domain $d$. Let $t_d = \sum_{k\in U_d} y_k$ be the total of the $y$-values in the domain $d$. Estimating a domain total is straightforward as $t_d$ can be expressed as $t_d = \sum_{k\in U} \delta_k y_k = \sum_{k\in U} \breve{y}_k$, where $\delta_k$ is a domain indicator attached to unit $k$ such that $\delta_k = 1$ if unit $k$ belongs to the domain $d$ and $\delta_k = 0$, otherwise, and $\breve{y}_k = \delta_k y_k$. Therefore, estimating $t_d$ is equivalent to estimating the population total of the $\breve{y}$-values. If $t_y$ denotes the overall population total, we have the relationship:

$$t_y = \sum_{k\in U} y_k = \sum_{d=1}^{D} t_d.$$

That is, the sum over the domains equals the overall population total. This property is often referred to as external consistency. Once again, it seems desirable that this property also holds at the estimate level. With a single weighting system $\{\tilde{w}_k; k \in S\}$ applied to to the $\breve{y}$-values, we obtain $D$ estimates, which are naturally consistent.

In the context of domain estimation, it is possible to develop domain-specific weighting systems. Hidiroglou and Patak (2004) showed empirically that domain-specific estimators are often more efficient than those based on a single weighting system. But with domain-specific estimators, the external consistency property is no longer satisfied, in general.

In conclusion, we believe that a single weighting system is generally effective in practice, achieving internal and external consistency and a manageable workload for the survey statisticians.

## ACKNOWLEDGMENTS

The authors are grateful to Chris Skinner and Jon Wakefield for their invitation to contribute to this special issue. They also thank an Associate Editor and two referees for their useful comments. The work of the first author was supported by a grant from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

## REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 203–242. Holt, Rinehart and Winston, Toronto.

BEAUMONT, J.-F. (2005a). Calibrated imputation in surveys under a quasi-model-assisted approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 445–458. MR2155348

BEAUMONT, J.-F. (2005b). On the use of data collection process information for the treatment of unit nonreponse through weight adjustment. *Surv. Methodol.* **31** 227–231.

BEAUMONT, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95** 539–553. MR2443174

BEAUMONT, J.-F. and BOCCI, C. (2008). Another look at ridge calibration. *Metron—International Journal of Statistics* **LXIV** 5–20.

BEAUMONT, J.-F. and HAZIZA, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Surv. Methodol.* **42** 319–323.

BREIDT, F. J. and OPSOMER, J. D. (2016). Model-assisted survey estimation with modern prediction techniques. *Statist. Sci.* **32** 190–205.

BREWER, K. R. W. (2002). *Combined Survey Sampling Inference*: *Weighing Basu's Elephants*. Arnold, London.

CHAMBERS, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Off. Stat.* **12** 3–32.

CHANG, T. and KOTT, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95** 555–571. MR2443175

CHEN, Q., ELLIOT, M. R., HAZIZA, D., YANG, Y., GOSH, M., LITTLE, R., SEDRANSK, J. and THOMPSON, M. (2016). Weights and estimation of a survey population mean: A review. *Statist. Sci.* To appear.

CLOUTIER, E. and LANGLET, É. (2014). Aboriginal Peoples Survey, 2012: Concepts and methods guide. Statistics Canada, Catalogue No. 89-653-X—No. 002, ISBN 978-1-100-22738-2.

DA SILVA, D. N. and OPSOMER, J. D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canad. J. Statist.* **34** 563–579. MR2347046

DA SILVA, D. N. and OPSOMER, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Surv. Methodol.* **35** 165–176.

DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11** 427–444. MR0003527

DEVILLE, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Actes du Colloque* de la Société Statistique du, Canada, Sherbrooke, Canada.

DEVILLE, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, *Insee*.

DEVILLE, J.-C. and LAVALLÉE, P. (2006). Indirect sampling: The foundation of the generalized weight share method. *Surv. Methodol.* **32** 165–176.

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

DEVILLE, J.-C., SÄRNDAL, C.-E. and SAUTORY, O. (1993). Generalized raking procedure in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020.

EKHOLM, A. and LAAKSONEN, S. (1991). Weighting via response modeling in the Finnish household budget survey. *J. Off. Stat.* **7** 325–337.

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. *Surv. Methodol.* **23** 33–40.

ESTEVAO, V. M. and SÄRNDAL, C.-E. (2000). A functional form approach to calibration. *J. Off. Stat.* **16** 379–399.

ESTEVAO, V. M. and SÄRNDAL, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Int. Stat. Rev.* **74** 127–147.

FIRTH, B. and BENNETT, K.-E. (1998). Robust models in probability sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 3–21. MR1625672

FOLSOM, R. E. and SINGH, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *Proceedings of the Survey Research Methods Section* 598–603. Amer. Statist. Assoc., Alexandria, VA.

FULLER, W. A. (1966). Estimation employing post strata. *J. Amer. Statist. Assoc.* **61** 1172–1183. MR0208781

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

HÁJEK, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by Basu, D. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 236. Holt, Rinehart, and Winston, Toronto.

HAZIZA, D. and BEAUMONT, J.-F. (2007). On the construction of imputation classes in surveys. *Int. Stat. Rev.* **75** 25–43.

HAZIZA, D. and LESAGE, E. (2016). A discussion of weighting procedures for unit nonresponse. *J. Off. Stat.* **32** 129–145.

HIDIROGLOU, M. A. and PATAK, Z. (2004). Domain estimation using linear regression. *Surv. Methodol.* **30** 67–78.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

IANNACCHIONE, V. G., MILNE, J. G. and FOLSOM, R. E. (1991). Response probability weight adjustments using logistic regression. In *Proceedings of the Survey Research Methods Section*, 637–642. Amer. Statist. Assoc., Alexandria, VA.

KALBERG, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *J. Off. Stat.* **16** 229–241.

KALTON, G. and FLORES-CERVANTES, I. (2003). Weighting methods. *J. Off. Stat.* **19** 81–97.

KASS, V. G. (1980). An exploratory technique for investigating large quantities of categorical data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **29** 119–127.

KIM, J. K. and KIM, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.* **35** 501–514. MR2381396

KIM, J. K., KWON, Y. and PAIK, M. H. C. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika* **103** 461–473. MR3509899

KIM, J. K. and PARK, M. (2010). Calibration estimation in survey sampling. *Int. Stat. Rev.* **78** 21–39.

KISH, L. (1992). Weighting for unequal $P_i$. *J. Off. Stat.* **8** 183–200.

KMENTA, J. (1971). *Elements of Econometrics*. MacMillan, New York.

KOTT, P. S. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Surv. Methodol.* **32** 133–142.

KOTT, P. S. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics* 29*B*, *Sample Surveys*: *Inference and Analysis*. Elsevier, Oxford.

LAVALLÉE, P. (2007). *Indirect Sampling*. Springer, New York. MR2344875

LAVALLÉE, P. and BEAUMONT, J.-F. (2016). Weighting: Principles and practicalities. In *The SAGE Handbook of Survey Methodology* (C. Wolf, D. Joye, T. W. Smith and Y.-C. Fu, eds.) 460–476, Chapter 30. Sage Publications, Thousand Oaks, CA.

LEMAÎTRE, G. and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Surv. Methodol.* **13** 199–207.

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Int. Stat. Rev.* **54** 139–157.

LITTLE, R. J. A. and VARTIVARIAN, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Surv. Methodol.* **31** 161–168.

LUNDSTRÖM, S. and SÄRNDAL, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *J. Off. Stat.* **15** 305–327.

MONTANARI, G. E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *Int. Stat. Rev.* **55** 191–202.

MONTANARI, G. E. and RANALLI, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *J. Amer. Statist. Assoc.* **100** 1429–1442. MR2236453

NARAIN, R. D. (1951). On sampling without replacement with varying probabilities. *J. Indian Soc. Agricultural Statist.* **3** 169–174. MR0045354

PHIPPS, P. and TOTH, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Ann. Appl. Stat.* **6** 772–794. MR2976491

POTTER, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Section on Survey Research Methods* 225–230. Amer. Statist. Assoc., Alexandria, VA.

RAO, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.* **6** 173–180. MR0201035

RAO, J. N. K. (1966). Alternative estimators in sampling for multiple characteristics. *Sankhyā Ser. A* **28** 47–60. MR0258176

RAO, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Stat.* **10** 153–165.

RAO, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Surv. Methodol.* **31** 117–138.

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

RAO, J. N. K., HIDIROGLOU, M., YUNG, W. and KOVACEVIC, M. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *J. Indian Soc. Agricultural Statist.* **64** 129–135, 325. MR2789774

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

SALGADO, D., PÉREZ-ARRIERO, C., HERRADOR, M. and ARBUÉS, I. (2012). Letter to the Editor. *J. Off. Stat.* **28** 471–476.

SAMPFORD, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54** 499–513. MR0223051

SÄRNDAL, C.-E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33** 99–119.

SÄRNDAL, C.-E. and LUNDSTRÖM, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

SÄRNDAL, C.-E. and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Int. Stat. Rev.* **55** 279–294. MR0963144

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SAUTORY, O. (2003). CALMAR2: A new version of the CALMAR calibration adjustment program. In Proceedings of Statistics Canada's Symposium 2003. Available at http://www.statcan.ca/english/freepub/11-522-XIE/2003001/session13/sautory.pdf.

SCOTT, A. and SMITH, T. M. F. (1969). A note on estimating secondary caracteristics in multivariate surveys. *Sankya A* **31** 497–498.

SILVA, P. L. N. and SKINNER, C. (1997). Variable selection for regression estimation in finite populations. *Surv. Methodol.* **23** 23–32.

SKINNER, C. J. and D'ARRIGO, J. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika* **98** 953–966. MR2860335

SKINNER, C. J. and WAKEFIELD, J. (2017). Introduction to the design and analysis of comlex survey data. *Statist. Sci.* **32** 165–175.

STEEL, D. G. and CLARK, R. G. (2007). Person-level and household-level regression estimation in household surveys. *Surv. Methodol.* **33** 51–60.

TILLÉ, Y. (2017). Probability sampling designs: New methods and guidelines. *Statist. Sci.* **32** 176–189.

WU, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika* **90** 937–951. MR2024768

WU, C. and LU, W. W. (2016). Calibration weighting methods for complex surveys. *Int. Stat. Rev.* **84** 79–98. MR3491280

WU, C. and RAO, J. N. K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canad. J. Statist.* **34** 359–375. MR2328549

WU, C. and SITTER, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96** 185–193. MR1952731