Statistical Methods & Applications
https://doi.org/10.1007/s10260-022-00630-9

ORIGINAL PAPER

Double-calibration estimators accounting for under-coverage and nonresponse in socio-economic surveys

Maria Michela Dickson[1] · Giuseppe Espa[1] · Lorenzo Fattorini[2] · Flavio Santi[3]

Accepted: 19 February 2022

Abstract

Under-coverage and nonresponse problems are jointly present in most socio-economic surveys. The purpose of this paper is to propose an estimation strategy that accounts for both problems by performing a two-step calibration. The first calibration exploits a set of auxiliary variables only available for the units in the sampled population to account for nonresponse. The second calibration exploits a different set of auxiliary variables available for the whole population, to account for under-coverage. The two calibrations are then unified in a double-calibration estimator. Mean and variance of the estimator are derived up to the first order of approximation. Conditions ensuring approximate unbiasedness are derived and discussed. The strategy is empirically checked by a simulation study performed on a set of artificial populations. A case study is derived from the European Union Statistics on Income and Living Conditions survey data. The strategy proposed is flexible and suitable in most situations in which both under-coverage and nonresponse are present.

Keywords Auxiliary variables · Calibration estimators · First-order Taylor series approximation · Simulation study

✉ Maria Michela Dickson
mariamichela.dickson@unitn.it

[1] Department of Economics and Management, University of Trento, Trento, Italy

[2] Department of Economics and Statistics, University of Siena, Siena, Italy

[3] Department of Economics, University of Verona, Verona, Italy

Published online: 10 March 2022

1 Introduction

Särndal et al. (1992, p. 8) establish four requirements to select a probability sample, setting the perimeter for the definition of a sampling design under the randomization principle. One requirement is that the procedure to select the sample ensure invariably positive probabilities to enter the sample for all units in the population.

This requirement may not be suitable in some situations such as in establishment surveys, such as the Economic Census conducted by the U.S. Census Bureau, in which the population of businesses is characterized by a highly skewed distribution in the survey variables (Glasser 1962). In this case, different approaches are commonly used, essentially based on the partition of population into strata determined by several business characteristics (e.g. size), and some strata are completely censused, some are sampled, and some are neglected, based on the features of units or the ability to contact them (Sigman and Monsour 1995). As happens in establishment surveys conducted by the U.S. Bureau of Economic Analysis, very small establishments are excluded a priori from the population to be sampled, due to the costs in building and updating a sampling frame, against an expected slight gain in efficiency of the estimators (see e.g. Hidiroglou 1986; De Haan et al. 1999; Rivest 2002). These instances are known in the literature as cut-off sampling (Knaub 2008; Benedetti et al. 2010; Haziza et al. 2010a). A similar position can be seen in social surveys on households, such as the Household Finance and Consumption Survey managed by the European Central Bank, characterized by the missed observation of population units considered ineligible for the survey, i.e. dwellings that are vacant, not habitable, with non-eligible members, etc., with consequences on the estimation of living conditions and poverty rate (Nicoletti et al. 2011). In this framework it is worth distinguishing between cut-off sampling, alternatively referred to as planned under-coverage, which is often used in socio-economic surveys and unplanned under-coverage which is typical in social surveys. In the first case, auxiliary information is available for all the units in the non-covered portion of the population, whereas only population totals are available in the second case (see e.g. Lehtonen and Veijanen 2009). Owing to the aforementioned under-coverage of the whole population, the unadjusted estimator is biased in these situations. Bias is usually corrected in the literature by means of model-based techniques (see, among others, Kott 2006; Haziza et al. 2010a). Recently, a solution to under-coverage problem has been proposed by Fattorini et al. (2018) in which the properties of the resulting estimator are evaluated in relation to the sampling design while all the population characteristics are held fixed. In particular, the authors propose adopting a calibration technique in which the weights originally attributed to each sample observation are modified in such a way as to be able to estimate the population totals of a set of auxiliary variables without error. The rationale behind calibration is evident: if the calibrated weights guess the population totals of the auxiliary variables without errors, they should also be suitable for estimating the total of the survey variable, providing a relationship exists between the survey variable and the auxiliaries. Obviously, calibration is likely to perform well in terms of precision under a strong linear relationship.

Socio-economic surveys also involve unit nonresponse, the more so the higher the sensitivity of the survey variables (e.g. sexual behavior, drug consumption, etc.). However undesirable, nonresponse is a natural contingency in surveys, so the damage to estimations and inferences needs to be addressed (Groves and Peytcheva 2008). This is crucial in survey sampling theory and is extensively treated in the literature (e.g. Brick and Montaquila 2009). Extensively applied methods include post-stratification (Holt and Smith 1979), response homogeneity groups (Särndal et al. 1992), and, more recently, model-based techniques including imputation and nonresponse propensity weighting (Särndal and Lundström 2005; Haziza et al. 2010b). In particular, nonresponse propensity weighting assumes that each unit of the sampled population has a strictly positive probability to respond. A model is then used to estimate the probabilities of respondent units from the sample by connecting these probabilities to auxiliary information by means of logistic regression models (Chang and Kott 2008). In addition to this source of uncertainty, the requirement of positive response probability seems to tighten in socio-economic surveys, because some units will not respond in any situation (e.g. homeless and geographically mobile individuals and families). Alternatively, Fattorini et al. (2013) attempt a design-based solution in which population values and nonresponse are viewed as fixed characteristics. For this purpose, they once again use the calibration technique, defined in the literature as nonresponse calibration weighting by Haziza et al. (2010b). In this case, weights originally attributed to each respondent unit are modified in such a way as to be able to estimate the population totals of a set of auxiliary variables without error.

In most cases under-coverage and nonresponse problems are jointly present in socio-economic surveys. Therefore, a general indication in the treatment of both problems concerns the use of any available auxiliary information, even if some is not available to all units of the population. In this paper, we build on the availability of a set of auxiliary variables for the whole population while another set is available only for the sampled portion. In establishment surveys, for example, much financial information may be available only for businesses of adequate size, such as corporations, and may not be for small businesses excluded from the sampling, such as micro-enterprises. Moreover, owing to recent data collection developments, the additional information may derive from big data, e.g. data from internet and telephone use, social networks, online purchases, etc.

The purpose of this paper is to propose double-calibration estimators. The use of calibration in two or more steps is not new and has already been used, among others, by Folsom and Singh (2000) and Estevao and Särndal (2006). Moreover, it has been routinely adopted by National Statistical Offices for many years. Here we propose an estimation strategy that considers both under-coverage and nonresponse problems, solving them by performing double calibration. The first calibration exploits a set of auxiliary variables available only for the units in the sampled population to account for nonresponse; the second calibration exploits a different set of auxiliary variables available for the whole population, to account for under-coverage. Joining together the two calibrations, we propose a double-calibration estimator that is applicable to all cases in which both under-coverage and nonresponse problems are present.

The paper is structured as follow. In Sect. 2, some preliminaries and notations are given. Section 3 is devoted to the construction of the double-calibration estimator and in Sect. 4 some statistical properties (expectation and variance) are derived. In order to check the efficiency of the strategy, in Sect. 5 Monte Carlo simulation studies are performed to explore several scenarios. In Sect. 6, using data from the European Union Statistics on Income and Living Conditions survey and from Statistics Denmark data, a case study to estimate the total income of Danish households in 2013 is presented and discussed. Some concluding remarks are given in Sect. 7.

2 Preliminaries and notation

Denote as $U = \{u_1, ..., u_N\}$ a finite population of N units. Let y_j, with $j \in U$, the value for unit j of the survey variable Y. We aim to estimate the population total $T_Y = \sum_{j \in U} y_j$. For the whole population there exists a vector $\boldsymbol{Z}$ of M auxiliary variables whose values $\boldsymbol{z}_j = \left[z_{j1}, ..., z_{jM}\right]^t$ are known for each $j \in U$, in such a way that the vector of totals $\boldsymbol{T}_Z = \sum_{j \in U} \boldsymbol{z}_j$ is also known.

In this setting, for one of the reasons mentioned in the introduction, only a sub-population U_B of size $N_B < N$ units is sampled using a fixed-size design having first- and second-order inclusion probabilities π_j, π_{jh} for any $h > j \in U_B$. Denote by $T_{Y(B)} = \sum_{j \in U_B} y_j$ the unknown total of Y in U_B. Moreover, suppose that additional information exists in the sub-population U_B. More precisely suppose that there exists a vector $\boldsymbol{X}$ of K auxiliary variables whose values $\boldsymbol{x}_j = \left[x_{j1}, ..., x_{jM}\right]^t$ are known for each $j \in U_B$ in such a way that the vector of totals $\boldsymbol{T}_{X(B)} = \sum_{j \in U_B} \boldsymbol{x}_j$ is also known. In this setting, denote by $\boldsymbol{T}_{Z(B)} = \sum_{j \in U_B} \boldsymbol{z}_j$ the known vector of total of the $\boldsymbol{z}_j$s in the sub-population U_B.

A random sample S of $n < N_B$ units is selected from the sub-population U_B by means of the adopted sampling scheme. As often happens in practice, especially in socio-economic surveys, the sample may be affected by nonresponses, in such a way that the sample is split into two sub-samples, the sub-sample $R \subset S$ of the respondent units and the sub-sample $S - R$ of the nonrespondent units.

The set presented above shows two problems to solve: first, a correction for nonresponses is necessary, in order to estimate $T_{Y(B)}$; second, since the sample S is selected from U_B and not from U, any $T_{Y(B)}$ estimator is biased, so a correction is needed in order to estimate T_Y. We propose a calibration in two steps, developed in the following sub-sections.

3 The double-calibration estimator

3.1 First calibration: from respondent group to sampled sub-population

The first issue to deal with is the nonresponse problem in a sample. Since S is selected in U_B, in the absence of nonresponses, it would be possible to estimate $T_{Y(B)}$ by means of the well-known Horvitz–Thompson (HT) estimator

$$\hat{T}_{Y(B)} = \sum_{j \in S} \frac{y_j}{\pi_j} \tag{1}$$

and $\hat{T}_{Y(B)}$ would be an unbiased estimator for $T_{Y(B)}$ if all π_j are positive. However, owing to nonresponses, any unadjusted estimator is destined to be a biased estimator of $T_{Y(B)}$. Following results obtained in Särndal and Lundström (2005), the bias may be reduced by exploiting the $\boldsymbol{X}$-vector of auxiliary information. The resulting estimator is

$$\hat{T}_{Y(B)cal} = \hat{\boldsymbol{b}}_R^t \boldsymbol{T}_{X(B)} \tag{2}$$

where $\hat{\boldsymbol{b}}_R = \hat{\boldsymbol{A}}_R^{-1} \hat{\boldsymbol{a}}_R$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{X}$, performed on the respondent sample R, i.e. $\hat{\boldsymbol{A}}_R = \sum_{j \in R} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^t}{\pi_j}$ and $\hat{\boldsymbol{a}}_R = \sum_{j \in R} \frac{y_j \boldsymbol{x}_j}{\pi_j}$ and the unit constant is tacitly adopted as the first auxiliary variable in the vector $\boldsymbol{X}$.

The properties of $\hat{T}_{Y(B)cal}$ are derived in Fattorini et al. (2013). The population is partitioned into respondent and nonrespondent strata and the estimator is approximately unbiased if the relationship between Y and $\boldsymbol{X}$ is similar in both the strata. Practically speaking, this condition is similar to the one assumed in most model-based nonresponse treatments (for a discussion, see Haziza and Lesage 2016).

3.2 Second calibration: from sampled sub-population to the whole population

Because $\hat{T}_{Y(B)cal}$ is, at most, an approximately unbiased estimator of $T_{Y(B)}$, it is a biased estimator of T_Y. Indeed, the sampling scheme adopted to select S generates a sampling design onto U_B but not onto U, and units of $U - U_B$ cannot enter the sample. Therefore, the missed selection of some population units leads to a bias due to population under-coverage and it is necessary to correct the estimator $\hat{T}_{Y(B)cal}$.

Fattorini et al. (2018) called these schemes as pseudo designs and proposed a design-based calibration estimation based on a single auxiliary variable having a proportional relationship with the survey variable. In order to extend this approach to vectors of auxiliary variables and to more general linear relationships, the population under-coverage is handled by the calibration criterion proposed by Särndal and Lundström (2005). Specifically, if the y_js were available for each $j \in S$, the information furnished by the M auxiliary variables $\boldsymbol{Z}$, available for all the population units, could be exploited by means of the calibration estimator

$$\hat{T}_{Y(cal)} = \hat{\boldsymbol{d}}_B^t \boldsymbol{T}_Z \tag{3}$$

where $\hat{\boldsymbol{d}_B} = \hat{\boldsymbol{C}}_B^{-1}\hat{\boldsymbol{c}}_B$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{Z}$, performed on the whole sample S, i.e. $\hat{\boldsymbol{C}}_B = \sum_{j\in S}\frac{z_j z_j^t}{\pi_j}$ and $\hat{\boldsymbol{C}}_B = \sum_{j\in S}\frac{y_j z_j}{\pi_j}$.

If we suppose once again that the unit constant is adopted as the first auxiliary variable in the vector $\boldsymbol{Z}$, then the calibration estimator (3) could be rewritten as

$$\hat{T}_{Y(cal)} = \hat{T}_{Y(B)} + \hat{\boldsymbol{d}}_B^t(\boldsymbol{T}_Z - \hat{\boldsymbol{T}}_{Z(B)}) \tag{4}$$

where $\hat{\boldsymbol{T}}_{Z(B)} = \sum_{j\in S}\frac{z_j}{\pi_j}$ is the HT estimator of the totals of the z_js in the sampled sub-population U_B (see Appendix A.1 for the proof).

However, the estimator $\hat{T}_{Y(cal)}$ is only virtual, because knowing the values of the survey variable only for the respondent subset R, neither the HT estimator $\hat{T}_{Y(B)}$ nor the least-squares coefficient vector $\hat{\boldsymbol{d}_B} = \hat{\boldsymbol{C}}_B^{-1}\hat{\boldsymbol{c}}_B$ are known. Therefore, exploiting Eq. (4), a *double calibration estimator* can be constructed by using $\hat{T}_{Y(B)cal}$ instead of $\hat{T}_{Y(B)}$ and $\hat{\boldsymbol{d}_R} = \hat{\boldsymbol{C}}_R^{-1}\hat{\boldsymbol{c}}_R$, instead of $\hat{\boldsymbol{d}_B}$ where $\hat{\boldsymbol{C}}_R = \sum_{j\in R}\frac{z_j z_j^t}{\pi_j}$ and $\hat{\boldsymbol{C}}_R = \sum_{j\in R}\frac{y_j z_j}{\pi_j}$. Practically speaking, the resulting estimator of the whole population total turns out to be

$$\hat{T}_{Y(dcal)} = \hat{T}_{Y(B)cal} + \hat{\boldsymbol{d}}_R^t(\boldsymbol{T}_Z - \hat{\boldsymbol{T}}_{Z(B)}) = \hat{\boldsymbol{b}}_R^t \boldsymbol{T}_{X(B)} + \hat{\boldsymbol{d}}_R^t(\boldsymbol{T}_Z - \hat{\boldsymbol{T}}_{Z(B)}) \tag{5}$$

With the double calibration estimator, the information provided by $\boldsymbol{X}$ and $\boldsymbol{Z}$ is exploited to handle both nonresponses and population under-coverage.

4 Statistical properties of the double calibration estimator

Denote by $U_{B(R)}$ the stratum of respondent units in the sub-population U_B and by $U_{B(NR)}$ the stratum of nonrespondent units. As suggested by Fattorini et al. (2013), introduce a dummy variable as $r_j = 1$ if $j \in U_{B(R)}$ and $r_j = 0$ if $j \in U_{B(NR)}$. Therefore, using the r_js indicators $\hat{\boldsymbol{A}}_R$, $\hat{\boldsymbol{a}}_R$, $\hat{\boldsymbol{C}}_R$ and $\hat{\boldsymbol{c}}_R$ can be rewritten as $\hat{\boldsymbol{A}}_R = \sum_{j\in S}\frac{r_j \boldsymbol{x}_j \boldsymbol{x}_j^t}{\pi_j}$, $\hat{\boldsymbol{a}}_R = \sum_{j\in S}\frac{r_j y_j \boldsymbol{x}_j}{\pi_j}$,$\hat{\boldsymbol{C}}_R = \sum_{j\in S}\frac{r_j z_j z_j^t}{\pi_j}$ and $\hat{\boldsymbol{C}}_R = \sum_{j\in S}\frac{r_j y_j z_j}{\pi_j}$. Therefore, the previous matrices and vectors as well as the double calibration estimator $\hat{T}_{Y(dcal)}$ depend on the selection of the sole sample S, while nonresponses are accounted for in the r_js, which are a fixed characteristic of the population.

It is worth noting that in this perspective, $\hat{\boldsymbol{A}}_R$, $\hat{\boldsymbol{a}}_R$, $\hat{\boldsymbol{C}}_R$, $\hat{\boldsymbol{c}}_R$ and $\hat{\boldsymbol{T}}_{Z(B)}$ are HT estimators of $\boldsymbol{A}_R = \sum_{j\in U_B} r_j \boldsymbol{x}_j \boldsymbol{x}_j^t = \sum_{j\in U_{B(R)}} \boldsymbol{x}_j \boldsymbol{x}_j^t$, $\boldsymbol{a}_R = \sum_{j\in U_B} r_j y_j \boldsymbol{x}_j = \sum_{j\in U_{B(R)}} y_j \boldsymbol{x}_j$, $\boldsymbol{C}_R = \sum_{j\in U_B} r_j \boldsymbol{z}_j \boldsymbol{z}_j^t = \sum_{j\in U_{B(R)}} \boldsymbol{z}_j \boldsymbol{z}_j^t$, $\boldsymbol{c}_R = \sum_{j\in U_B} r_j y_j \boldsymbol{z}_j = \sum_{j\in U_{B(R)}} y_j \boldsymbol{z}_j$ and of $\boldsymbol{T}_{Z(B)}$, respectively. Therefore, because $\hat{T}_{Y(dcal)}$ is differentiable with respect to $\hat{\boldsymbol{A}}_R$, $\hat{\boldsymbol{a}}_R$, $\hat{\boldsymbol{C}}_R$, $\hat{\boldsymbol{c}}_R$ and $\hat{\boldsymbol{T}}_{Z(B)}$, it can be approximated up to the first term by a Taylor series around the true population counterparts $\boldsymbol{A}_R$, $\boldsymbol{a}_R$, $\boldsymbol{C}_R$, $\boldsymbol{c}_R$ and $\boldsymbol{T}_{Z(B)}$. The equation of the first-order Taylor series approximation of $\hat{T}_{Y(dcal)}$ is derived in Appendix A.2.

4.1 Approximate expectation

From the first-order Taylor series approximation of $\hat{T}_{Y(dcal)}$ it immediately follows that

$$AE(\hat{T}_{Y(dcal)}) = \boldsymbol{b}_R^t \boldsymbol{T}_{X(B)} + \boldsymbol{d}_R^t(\boldsymbol{T}_Z - \boldsymbol{T}_{Z(B)}) \tag{6}$$

where $\boldsymbol{b}_R = \boldsymbol{A}_R^{-1}\boldsymbol{a}_R$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{X}$ performed on the respondent stratum $U_{B(R)}$ and $\boldsymbol{d}_R = \boldsymbol{C}_R^{-1}\boldsymbol{c}_R$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{Z}$ performed in the same stratum. Exploiting equation (6), after some algebra shown in Appendix A.3, proves that the double calibration estimator is unbiased up to the first-order approximation if:

1. the linear relationship between Y and $\boldsymbol{X}$ is similar in the respondent and nonrespondent strata of U_B, i.e. $\boldsymbol{b}_R \approx \boldsymbol{b}_{NR}$, where $\boldsymbol{b}_{NR}$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{X}$ performed on the nonrespondent stratum $U_{B(NR)}$;
2. the linear relationship between Y and $\boldsymbol{Z}$ is similar in the respondent stratum and in the whole sub-population U_B, i.e. $\boldsymbol{d}_R \approx \boldsymbol{d}_B$, where $\boldsymbol{d}_B$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{Z}$ performed on the whole sub-population U_B;
3. the linear relationship between Y and $\boldsymbol{Z}$ is similar in the two sub-populations U_B and $U - U_B$, i.e. $\boldsymbol{d}_B \approx \boldsymbol{d}_{NB}$, where $\boldsymbol{d}_{NB}$ is the least-square coefficient vector of the regression of Y vs $\boldsymbol{Z}$ performed on the whole sub-population $U - U_B$.

It is worth noting that the approximate expectation in Eq. (6) does not depend on the design (e.g., first and second order inclusion probabilities), but only on the population characteristics. Therefore, under conditions 1–3, design-unbiasedness holds irrespective of the sampling design adopted.

4.2 Approximate variance and variance estimation

From equation (A.3) of Appendix A.2, the first-order Taylor series approximation of $\hat{T}_{Y(dcal)}$ is rewritten as a translation of an HT estimator, in the sense that

$$\hat{T}_{Y(dcal)} = cost + \sum_{j \in S} \frac{u_j}{\pi_j}$$

where

$$u_j = r_j\left(y_j \boldsymbol{x}_j^t - \boldsymbol{a}_R^t \boldsymbol{A}_R^{-1} \boldsymbol{x}_j \boldsymbol{x}_j^t\right)\boldsymbol{A}_R^{-1}\boldsymbol{T}_{X(B)} + \\ + r_j\left(y_j \boldsymbol{z}_j^t - \boldsymbol{c}_R^t \boldsymbol{C}_R^{-1} \boldsymbol{z}_j \boldsymbol{z}_j^t\right)\boldsymbol{C}_R^{-1}\left(\boldsymbol{T}_Z - \boldsymbol{T}_{Z(B)}\right) - \boldsymbol{c}_R^t \boldsymbol{C}_R^{-1} \boldsymbol{z}_j, j \in U_B$$

are the influence values (e.g. Davison and Hinkley 1997).

Therefore, the approximate variance of $\hat{T}_{Y(dcal)}$ turns out to be (e.g. Särndal et al. 1992, p. 175)

$$AV\left(\hat{T}_{Y(dcal)}\right) = \sum_{h>j\in U_B} (\pi_j\pi_h - \pi_{jh})\left(\frac{u_j}{\pi_j} - \frac{u_h}{\pi_h}\right)^2 \tag{7}$$

On the basis of Eq. (7), the well-known Sen–Yates–Grundy (SYG) variance estimator is given by

$$\hat{V}^2_{SYG} = \sum_{h>j\in S} (\pi_j\pi_h - \pi_{jh})\left(\frac{\hat{u}_j}{\pi_j} - \frac{\hat{u}_h}{\pi_h}\right)^2 \tag{8}$$

where

$$\hat{u}_j = r_j\left(y_j\boldsymbol{x}_j^t - \hat{\boldsymbol{a}}_R^t\hat{\boldsymbol{A}}_R^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^t\right)\hat{\boldsymbol{A}}_R^{-1}\boldsymbol{T}_{X(B)} + \\ + r_j\left(y_j\boldsymbol{z}_j^t - \hat{\boldsymbol{C}}_R^t\hat{\boldsymbol{C}}_R^{-1}\boldsymbol{z}_j\boldsymbol{z}_j^t\right)\hat{\boldsymbol{C}}_R^{-1}\left(\boldsymbol{T}_Z - \hat{\boldsymbol{T}}_{Z(B)}\right) - \hat{\boldsymbol{C}}_R^t\hat{\boldsymbol{C}}_R^{-1}\boldsymbol{z}_j, j \in S$$

are the empirical influence values computed for each sample unit.

5 Simulation study

Simulations were used to check the performance of the proposed estimator. We considered a population U of $N = 10000$ units and a sub-population $U_B \subset U$ of $N_B = 7500$ units. We assumed that the values z_j of an auxiliary variable Z were available for each $j \in U$ and were adopted for sample under-coverage calibration. Moreover, we assumed that the values x_j of an auxiliary variable X achieved from additional information were available for each $j \in U_B$ and were adopted in nonresponse calibration. We also assumed that the sub-population U_B was partitioned into respondent and non-respondent strata $U_{B(R)}$ and $U_{B(NR)}$, respectively. Three sizes were assumed for the respondent stratum, $N_{B(R)} = 2250; 4500; 6750$ units corresponding to response rates of 30%, 60% and 90%, respectively. Moreover, variables were generated respecting some criteria, in order to explore several scenarios, as explained below.

5.1 Unbiasedness of $\hat{T}_{Y(dcal)}$

The auxiliary variables X and Z and the survey variables Y were generated from a tri-variate normal distribution. The expectations and variances of X and Z were assumed to be equal to 1, while the expectation and variance of Y were assumed to be equal to 2 and 4, respectively. These setups assured that each variable had a coefficient of variation of 1. The correlation between X and Y was set at $\rho_{XY} = 0.3; 0.6; 0.9$; similarly, the correlation between Z and Y was set at $\rho_{ZY} = 0.3; 0.6; 0.9$, giving rise to nine scenarios. The correlation between X and Z was set at the minimum possible value ρ_{XZ} such that the resulting variance-covariance matrix is positive-definite. Once the nine variance-covariance matrices were established the 10000 values of Z and Y and the 7500 values of X were generated using the triangular square root of the variance-covariance matrix (e.g.

Johnson 2013, Sect. 4.1). Subsequently, the first $N_{B(R)}$ units of U_B were assumed to be the respondent portion of the population, ensuring in this way compliance with conditions 1.−3., i.e. the approximate unbiasedness of the double calibration estimator. Simple random sampling without replacement (SRSWOR) was the sampling scheme adopted to select samples of sizes $n = 75; 250; 500$ from U_B. If the same sampling efforts were adopted to select samples from the whole population U and in the absence of nonresponses, then the HT estimator of the total would give rise to relative root means squared errors

$$RRMSE_{SRSWOR} = \sqrt{\frac{N-n}{Nn}} CV_Y \tag{9}$$

where CV_Y is the coefficient of variation of the survey variable. Equation (9) was taken as the benchmark for the performance of the double calibration estimator.

For each combination of respondent sizes $N_{B(R)}$, correlations between X and Y, correlations between Z and Y, and sample sizes n, 10000 random samples were selected by means of SRSWOR from U_B, and the double calibration estimates $\hat{T}_i = (i = 1, ..., 10000)$ were computed using equation (5). Moreover, from each simulated sample, the variance estimates $\hat{V}_i^2 = (i = 1, ..., 10000)$ were also computed using equation (8), which under SRSWOR is reduced to

$$\hat{V}_{SYG}^2 = N_B(N_B - n)\frac{s_{\hat{u}}^2}{n} \tag{10}$$

where $s_{\hat{u}}^2$ is the sampling variance of the $\hat{u}_j$s. Once the variance estimates were computed from (10), the *RRMSE* estimates $\hat{RRMSE}_i = \frac{\hat{V}_i}{\hat{T}_i}$ were achieved together with the confidence intervals at the nominal level of 0.95, $\hat{T}_i \pm 2\hat{V}_i$. Therefore, from the resulting Monte Carlo distributions of these quantities, the expectations $E(\hat{T}_{Y(dcal)}) = \frac{1}{10000}\sum_{i=1}^{10000}\hat{T}_i$ and mean squared errors $MSE(\hat{T}_{Y(dcal)}) = \frac{1}{10000}\sum_{i=1}^{10000}(\hat{T}_i - T_Y)^2$ of the double calibration estimator were empirically derived from which the relative bias $RB = \frac{E(\hat{T}_{Y(dcal)}) - T_Y}{T_Y}$ and the relative root mean squared errors $RRMSE = \frac{\sqrt{MSE(\hat{T}_{Y(dcal)})}}{T_Y}$ were derived. The expectations of the *RRMSE* estimator $ERRMSEE = \frac{1}{10000}\sum_{i=1}^{10000}\hat{RRMSE}_i$ and the coverage of the 0.95 confidence interval $COV95 = \frac{1}{10000}\sum_{i=1}^{10000} I(\hat{T}_i - 2\hat{V}_i \le T_Y \le \hat{T}_i + 2\hat{V}_i)$ are also computed. The most relevant results of the Monte Carlo simulations are shown in Tables 1 and 2, while the remaining simulation results are shown in Tables B.1–B.7 of the Appendix B.

The simulation results suggest the following remarks. The first order approximation of relative bias and *RRMSE* are very accurate in most cases. The discrepancies between approximation and the empirical values achieved from the Monte Carlo distributions are usually smaller than one percent point and become lower with high levels of response and correlations. The theoretical findings for the bias reduction, shown in Sect. 4.1 are fully confirmed by the simulation results. The artificial populations considered in the study meet unbiasedness conditions 1.−3.

Table 1 Percentage values of *RB*, *ARRMSE*, *RRMSE*, *ERRMSEE*, *COV*95 and first order approximation of relative bias (*ARB*) achieved from a population of 10000 units, a sampled sub-population of 7500 units with 2250, 4500 and 6750 respondent units, sample sizes $n = 75; 250; 500$ selected by means of simple random sampling without replacement

$N_{B(R)}$	*ARB*	*n*	*RB*	*ARRSME*	*RRMSE*	*ERRMSEE*	*COV*95
2250	− 1.7	75	− 1.7	20.0	21.0 (11.5)	20.7	92.1
2250	− 1.7	250	− 1.8	10.8	11.1 (6.2)	11.1	94.4
2250	− 1.7	500	− 1.9	7.5	7.8 (4.4)	7.7	94.4
4500	− 1.0	75	− 1.0	14.1	14.4 (11.5)	14.4	94.2
4500	− 1.0	250	− 1.1	7.6	7.7 (6.2)	7.7	94.8
4500	− 1.0	500	− 1.1	5.3	5.4 (4.4)	5.4	95.0
6750	− 1.5	75	− 1.6	11.2	11.5 (11.5)	11.5	94.8
6750	− 1.5	250	− 1.6	6.1	6.3 (6.2)	6.2	94.4
6750	− 1.5	500	− 1.5	4.2	4.5 (4.4)	4.3	94.1

Correlations $\rho_{XY} = 0.3$ and $\rho_{ZY} = 0.3$. Values in parentheses are the *RRMSE*s of the Horvitz–Thompson estimator in the absence of nonresponse and under-coverage

Table 2 Percentage values of *RB*, *ARRMSE*, *RRMSE*, *ERRMSEE*, *COV*95 and first order approximation of relative bias (*ARB*) achieved from a population of 10000 units, a sampled sub-population of 7500 units with 2250, 4500 and 6750 respondent units, sample sizes $n = 75; 250; 500$ selected by means of simple random sampling without replacement

$N_{B(R)}$	*ARB*	*n*	*RB*	*ARRSME*	*RRMSE*	*ERRMSEE*	*COV*95
2250	− 0.6	75	− 0.6	8.3	8.7 (11.5)	8.7	95.3
2250	− 0.6	250	− 0.6	4.5	4.6 (6.2)	4.6	95.3
2250	− 0.6	500	− 0.7	3.1	3.2 (4.4)	3.2	95.1
4500	− 1.0	75	− 0.9	6.8	7.0 (11.5)	7.0	95.4
4500	− 1.0	250	− 0.9	3.7	3.8 (6.2)	3.7	95.1
4500	− 1.0	500	− 0.9	2.6	2.7 (4.4)	2.6	94.1
6750	− 1.0	75	− 1.0	6.2	6.4 (11.5)	6.4	95.7
6750	− 1.0	250	− 0.9	3.4	3.5 (6.2)	3.4	94.3
6750	− 1.0	500	− 0.9	2.3	2.5 (4.4)	2.4	93.6

Correlations $\rho_{XY} = 0.9$ and $\rho_{ZY} = 0.9$. Values in parentheses are the *RRMSE*s of the Horvitz–Thompson estimator in the absence of nonresponse and under-coverage

Indeed the empirical values of the relative bias are negligible (invariably about one percentage point) irrespective of the level of correlation of the survey variable with the auxiliaries. While the level of correlation does not affect the bias reduction, it has a relevant impact on the precision. When correlations are strong the double calibration estimator proves efficient, reaching values of *RRMSE* that are even smaller than those achieved by the HT estimator with the same sampling effort and

in the absence of nonresponse and under-coverage. Obviously precision increases with the level of response.

The *RRMSE* estimator obtained from the variance estimator (8) is approximately unbiased providing also confidence intervals with coverage near to the nominal level of 95% in most cases. Because the estimator (8) actually estimates the approximate variance, some exceptions occur when the variance approximations (and subsequently the *RRMSE*) turn out be smaller than the true values.

5.2 Robustness of $\hat{T}_{Y(dcal)}$ when conditions 1.-3. do not hold

Additional simulations were performed to achieve insights on the robustness of the proposed estimator when the approximate unbiasedness conditions 1.−3. were moderately violated in such a way that an amount of bias was invariably involved. Indeed, as stated by Särndal and Lundström, (2005, p. 98), when an estimator is biased, its bias should be the main concern, given that *"if an estimator is greatly biased, it is poor consolation that its variance is low"*. Hence, here too, if a massive bias were present it would heavily impact on *RRMSE*, deteriorating the estimator performance. To investigate this issue, the linear relationship between Y and X was assumed to be different in the respondent and nonrespondent strata of U_B, as in the following scheme:

(a) when the correlation among Y and X in the respondent stratum was equal to 0.3, the same correlation in the nonrespondent stratum was decreased or increased to 0.2 or to 0.4;
(b) when the correlation among Y and X in the respondent stratum was equal to 0.6, the same correlation in the nonrespondent stratum is decreased or increased to 0.5 or to 0.7;
(c) when the correlation among Y and X in the respondent stratum was equal to 0.9, the same correlation in the nonrespondent stratum is decreased or increased to 0.8 or to 0.95.

Similarly, the linear relationship between Y and Z was assumed to be different in the subpopulations U_B and $U - U_B$, following the scheme:

(a) when the correlation among Y and Z in the sub-population $U - U_B$ was equal to 0.3, the same correlation in the sub-population U_B was decreased or increased to 0.2 or to 0.4;
(b) when the correlation among Y and Z in the sub-population $U - U_B$ was equal to 0.6, the same correlation in the sub-population U_B was decreased or increased to 0.5 or to 0.7;
(c) when the correlation among Y and Z in the sub-population $U - U_B$ was equal to 0.9, the same correlation in the sub-population U_B was decreased or increased to 0.8 or to 0.95.

As explained, each decrease or increase in the correlation between Y and X was paired with the corresponding decrease or increase in the correlation between Y and

Z, giving rise to a total of twelve scenarios: six representing different (and weaker) relationships of Y with X and Z in the nonrespondent stratum and subpopulation U_B, respectively; the other six representing different (and stronger) relationships. As in the previous simulation experiment, expectations and variances of X and Z were assumed to be equal to 1 and expectation and variance of Y were assumed to be equal to 2 and 4, respectively. The correlation between X and Z was set at the minimum possible value ensuring a positive-definite variance-covariance matrix. Once the twelve variance-covariance were established, simulations proceeded as described above, with the same performance indices computed on the resulting Monte Carlo distributions. Some results are rset out in Tables 3 and 4, while remaining simulation results are given in Tables C.1–C.10 in the Appendix C.

The simulation results suggest the following remarks. The first order approximation of relative bias and *RRMSE* remain accurate with discrepancies usually smaller than one percent point. Even under different relationships of *Y*with X and Z, the relative bias remains moderate (invariably below 1.6 percentage point). The moderate increases in bias also entail moderate increases in *RRMSE* and approximately unbiased *RRMSE* estimation, with confidence intervals having coverages near to their nominal value. These results show a promising robustness of the estimator in the presence of moderate differences in the relationships of Y with X and Z in respondent and nonrespondent strata and sub-populations, respectively.

Table 3 Percentage values of *RB*, *ARRMSE*, *RRMSE*, *ERRMSEE*, *COV*95 and first order approximation of relative bias (*ARB*) achieved from a population of 10000 units, a sampled sub-population of 7500 units with 2250, 4500 and 6750 respondent units, sample sizes $n = 75; 250; 500$ selected by means of simple random sampling without replacement

$N_{B(R)}$	*ARB*	*n*	*RB*	*ARRSME*	*RRMSE*	*ERRMSEE*	*COV*95
2250	− 1.1	75	− 1.0	10.4	10.8 (11.6)	10.4	93.6
2250	− 1.1	250	− 1.1	5.6	5.6 (6.3)	5.5	94.7
2250	− 1.1	500	− 1.1	3.9	3.9 (4.4)	3.8	94.3
4500	− 0.4	75	− 0.2	7.2	7.2 (11.6)	7.2	95.1
4500	− 0.4	250	− 0.4	3.8	3.8 (6.3)	3.8	95.3
4500	− 0.4	500	− 0.4	2.7	2.7 (4.4)	2.6	95.0
6750	− 0.1	75	− 0.1	5.8	5.8 (11.6)	5.8	95.0
6750	− 0.1	250	− 0.1	3.1	3.1 (6.3)	3.1	95.4
6750	− 0.1	500	− 0.1	2.1	2.1 (4.4)	2.1	95.0

Correlations $\rho_{XY} = 0.8$ in U_B, $\rho_{XY} = 0.9$ in the respondent stratum, $\rho_{ZY} = 0.3$ in $U - U_B$, and $\rho_{ZY} = 0.2$ in U_B. Values in parentheses are the *RRMSE*s of the Horvitz–Thompson estimator in the absence of nonresponse and under-coverage

Table 4 Percentage values of *RB*, *ARRMSE*, *RRMSE*, *ERRMSEE*, *COV*95 and first order approximation of relative bias (*ARB*) achieved from a population of 10000 units, a sampled sub-population of 7500 units with 2250, 4500 and 6750 respondent units, sample sizes $n = 75; 250; 500$ selected by means of simple random sampling without replacement

$N_{B(R)}$	*ARB*	*n*	*RB*	*ARRSME*	*RRMSE*	*ERRMSEE*	*COV*95
2250	− 0.8	75	− 1.0	15.8	16.7 (11.6)	16.2	93.4
2250	− 0.8	250	− 1.1	8.6	8.7 (6.3)	8.6	94.7
2250	− 0.8	500	− 1.3	6.0	6.0 (4.4)	6.0	94.8
4500	− 0.1	75	− 0.2	10.0	10.1 (11.6)	10.0	94.7
4500	− 0.1	250	− 0.1	5.4	5.3 (6.3)	5.4	95.3
4500	− 0.1	500	0.0	3.7	3.7 (4.4)	3.7	95.2
6750	0.1	75	0.2	7.0	7.1 (11.56)	7.0	94.6
6750	0.1	250	0.2	3.8	3.8 (6.3)	3.7	95.0
6750	0.1	500	0.2	2.6	2.6 (4.4)	2.6	95.6

Correlations $\rho_{XY} = 0.2$ in U_B, $\rho_{XY} = 0.3$ in the respondent stratum, $\rho_{ZY} = 0.9$ in $U - U_B$, and $\rho_{ZY} = 0.8$ in U_B. Values in parentheses are the *RRMSE*s of the Horvitz–Thompson estimator in the absence of nonresponse and under-coverage

6 An application to the European Union Statistics on Income and Living Conditions survey

National statistical institutes periodically collect data on living conditions through household surveys. Information contents concern several aspects of living conditions, such as, among others, features and expenses incurred to manage the dwelling, material deprivation and welfare indicators, individual and household incomes. The European Union Statistics on Income and Living Conditions survey was created from the previous experience of the European Community Household Panel (ECHP). The survey was launched in 2003 in seven countries (Belgium, Denmark, Greece, Ireland, Luxembourg, Austria and Norway), and was extended to all the 28-EU member countries, plus Switzerland, Norway, Iceland, FYROM and Serbia. It is conducted yearly and gathers information about European households. Some rules on how to conduct the survey are established by Eurostat, such as, among others, the frequency and the period to which questions must refer, and the aggregation level of some longitudinal and cross-sectional estimates. Other aspects of the survey are set independently by each country, such as, for instance, the sampling design and the sample size, leading to several discrepancies between countries (see, among others, Goedemé 2013; Lohmann 2011).

Moreover, the population coverage of surveys like these is incomplete. Individuals who do not live in households, as well as the homeless, the physically or mentally unable, geographically mobile and displaced individuals are not always represented in national-level data. It is estimated that worldwide some 300 to 350 million people may be missing from survey sampling frames, at least 45% omitted altogether by design, or because they are likely to be undercounted (Carr-Hill 2013).

The European Union Statistics on Income and Living Conditions survey, which involves approximately 300,000 households across Europe, is no exception and is affected by under-coverage, and the samples selected are affected by nonresponses. We propose an example of the use of the double-calibration estimator in the 2013 wave of the European Union Statistics on Income and Living Conditions survey in Denmark (hereafter DK-SILC). Data on respondents are freely available from the Eurostat website, while the further information required was taken from Statistics Denmark.

The reference population U consists of households residing in Denmark, except for those habitually living in a foreign country or cohabitations as orphanages, religious institutes, etc. On the Statistics Denmark website, the household population size in 2013 was equal to 2891119 units. The DK-SILC survey is based on a simple random sampling without replacement design, so that inclusion probabilities are equal for all units in the population. The sampling unit is the individual person and the household is defined as the household in which the selected person is member. This is because a household in Denmark is defined as comprising one or more individuals. Households eligible for DK-SILC are those in which the sampling unit is a person aged 16 or over, living alone or together in private dwellings and through marriage, parentage, affinity or other relationships. Hence, the eligible population U_B of Danish households is equal to 2416597, leading to an under-coverage rate of 0.16%.

The 2013 DK-SILC survey was featured a nonresponse rate of about 63%. In fact, the respondent number was equal to 5419, against a sample of 14702 households. Micro-data about respondents include a great deal of information, grouped into four sections: Household Register (D), Personal Register (R), Household Data (H) and Personal Data (P). Variables collected concern items, most of them qualitative. To implement the present case study, we use quantitative variables (in euro) with reference to the previous survey year (2012), contained in the H-section. Specifically, the tax on income and social contributions (HY140G) is used as the X variable to correct for nonresponse, while the total housing cost (HH070) is used as the Z variable to correct for under-coverage. The variable Y to be estimated is total household disposable income (HY020). Sample data suggest that both auxiliary variables are slightly correlated with the variable to be estimated (0.38 for X and Y; 0.17 for Z and Y, in the respondent group), revealing an unfavorable situation, worse than all those presented in Sect. 5. However, from simulation results, the weak relationships between the survey and the auxiliary variable should deteriorate precision but, fortunately, bias reduction should not deteriorate. The estimated total household disposable income is equal to 125739.17 million euros, equivalent to an average household disposable income on U is 43491.52 euros. Since the sampling design is SRSWOR, the variance estimate is computed as in (10) and the *RRMSE* estimate is 0.05.

The results obtained need to be understood as an illustration and do not claim to be official estimates. Clearly, the quality of the results relies on the quality of the available data. Howsoever, results are in line with those disseminated by Statistics Denmark. In fact, the average disposable income for all households (population U)

in 2012 is 329803 Danish krones, corresponding to approximately 44221 euros (at the average exchange rate in 2013).

7 Final remarks

The proposed double-calibration estimator can be adopted in socio-economic surveys to jointly account for nonresponse and under-coverage, adopting a two-step calibration. The first calibration, performed to reduce nonresponse bias, requires a set of auxiliary variables whose totals are known for the sampled sub-populations and whose values are known for the respondent units in the sample. The second calibration, performed to reduce the bias generated by the cut-off sampling, requires a further set of auxiliary variables whose totals are known for the whole populations and whose values are known for all the units in the sample. In this setting, no frame is necessary for the non-sampled sub-population. If the relationships of the survey variable with the two sets of auxiliaries are approximately similar in sampled and non-sampled sub-populations as well as in respondent and nonrespondent strata (conditions 1.−3.), the proposed estimator proves to be effective for reducing bias, and is also efficient for high-quality auxiliary variables correlated with the variable of interest. Interestingly, bias remains negligible and precision remains satisfactory including after moderate changes in the relationship of the variable of interest with the auxiliary variables in the respondent and nonrespondent strata and subpopulations. Socio-economic surveys may benefit from the application of the double-calibration estimator. It leads to results very close to those disseminated by national institutes of statistics and typically achieved by integrating several data sources, with far less effort in terms of data collection and integration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10260-022-00630-9.

References

Benedetti R, Bee M, Espa G (2010) A framework for cut-off sampling in business survey design. J Off Stat 26(4):651

Brick JM, Montaquila JM (2009) Nonresponse and weighting. In: Handbook of statistics, volume 29, pages 163–185. Elsevier

Carr-Hill R (2013) Missing millions and measuring development progress. World Dev 46:30–44

Chang T, Kott PS (2008) Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika 95(3):555–571

Davison AC, Hinkley DV (1997) Bootstrap methods and their application (vol. 1)

De Haan J, Opperdoes E, Schut CM (1999) Item selection in the consumer price index: cut-off versus probability sampling. Surv Methodol 25:31–42
Estevao VM, Särndal C-E (2006) Survey estimates by calibration on complex auxiliary information. Int Stat Rev 74(2):127–147
Fattorini L, Franceschi S, Maffei D (2013) Design-based treatment of unit nonresponse in environmental surveys using calibration weighting. Biomet J 55(6):925–943
Fattorini L, Gregoire TG, Trentini S (2018) The use of calibration weighting for variance estimation under systematic sampling: applications to forest cover assessment. J Agric Biol Environ Stat: 1–16
Folsom RE, Singh AC (2000) The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In: Proceedings of the American Statistical Association, Survey Research Methods Section, volume 598603
Glasser G (1962) On the complete coverage of large units in a statistical study. In: Revue de l'Institut International de Statistique, pages 28–32
Goedemé T (2013) How much confidence can we have in eu-silc? Complex sample designs and the standard error of the Europe 2020 poverty indicators. Soc Indicat Res 110(1):89–110
Groves RM, Peytcheva E (2008) The impact of nonresponse rates on nonresponse bias: a meta-analysis. Publ Opin Quart 72(2):167–189
Haziza D, Lesage É (2016) A discussion of weighting procedures for unit nonresponse. J Off Stat 32(1):129
Haziza D, Chauvet G, Deville J-C (2010) Sampling and estimation in the presence of cut-off sampling. Aust New Zeal J Stat 52(3):303–319
Haziza D, Thompson KJ, Yung W (2010) The effect of nonresponse adjustments on variance estimation. Surv Methodol 36(1):35–43
Hidiroglou MA (1986) The construction of a self-representing stratum of large units in survey design. Am Stat 40(1):27–31
Holt D, Smith TF (1979) Post stratification. J R Stat Soc Ser A (Gen) 142(1):33–46
Johnson ME (2013) Multivariate statistical simulation: A guide to selecting and generating continuous multivariate distributions. John Wiley & Sons
Knaub Jr JR (2008) Cutoff vs. design-based sampling and inference for establishment surveys. InterStat
Kott PS (2006) Using calibration weighting to adjust for nonresponse and coverage errors. Surv Methodol 32(2):133
Lehtonen R, Veijanen A (2009) Design-based methods of estimation for domains and small areas. In: Handbook of statistics, volume 29, pages 219–249. Elsevier
Lohmann H (2011) Comparability of eu-silc survey and register data: the relationship among employment, earnings and poverty. J Eur Soc Policy 21(1):37–54
Nicoletti C, Peracchi F, Foliano F (2011) Estimating income poverty in the presence of missing data and measurement error. J Bus Econ Stat 29(1):61–72
Rivest L-P (2002) A generalization of the lavallée and hidiroglou algorithm for stratification in business surveys. Surv Methodol 28(2):191–198
Särndal C-E, Lundström S (2005) Estimation in surveys with nonresponse. John Wiley & Sons
Särndal C-E, Swensson B, Wretman J (1992) Model assisted survey sampling
Sigman RS, Monsour NJ (1995) Selecting samples from list frames of businesses. Bus Surv Methods 295:133