Adjustments for Survey Imputed Datasets to Achieve First and Second–order Properties

Damião N. da Silva* Li-Chun Zhang †

Abstract

For practical convenience reasons, survey datasets are often disseminated to both internal and external secondary users with imputed values in place of the missing observations that occur during the data collection and processing stages. However, the released dataset generally leads to incorrect analyses if standard complete-data methods are applied directly without taking into account the imputation models. To alleviate this problem the survey statistician may, firstly, obtain some key results in any manner that is considered appropriate, and, secondly, calibrate the disseminated data so that these results can be reproduced by relevant standard complete-data procedures. In this paper, we discuss an approach that allows us to control both the first- and second-order properties of the imputed data and can be applied to complex surveys. We illustrate the implementation of the proposed approach with a numerical example.

Key Words: Complex surveys, Missing data, Survey nonresponse, Reverse calibration, Secondary data analysis, Hot deck imputation

1. Introduction

Survey organizations face frequently the problem of missing data due to nonresponse or inconsistencies of particular item values. This problem affects the quality of the survey data, generates an incomplete and more difficult data structure to analyze and also reduces the sample sizes originally intended. An option for data dissemination for secondary use is to complete the dataset by imputation, that is by applying a set of procedures to replacing "suitably chosen" or estimated values for the missing or inconsistent information. The use of imputation is a practical alternative to carrying out additional work to recontact nonrespondents that is beyond the survey budget and time constraints, especially because recontacts can not always be successful.

In this paper, we address the imputation perspective pointed by Sedransk (1985, p. 446) to allowing secondary data analysts to perform correct statistical analyses by applying simple tools available on standard statistical software. We propose an extension for the reverse calibration approach (Chambers and Ren 2004), originally formulated in the context of handling outliers in survey data. Their goal is to create an imputed dataset that could allow secondary users to reproduce robust estimates of totals by applying simple estimation methods in the absence of outliers. It should be stressed that the robust estimates are obtained by the data producer using appropriate statistical methods of estimation in the presence of the outliers. Imputation is used as a device to produce a clean and complete dataset in which the outlying observations are replaced by properly chosen values.

It is worth noting that, while much attention has traditionally been given to the public data users, the need of internal reuses has grown in the recent years. Many statistical offices are currently developing a corporate-wide data warehouse architecture, whereby the statistical data across the different surveys and register sources are brought to a central

*Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Campus Universitário s/n, Natal, RN, Brazil 59078-970

†Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton, Hampshire SO17 1BJ

database for future reuses. The ability for such internal secondary users to reproduce the appropriate estimates and the associated uncertainty measures by using standard complete-data procedures is just as important.

The reverse calibration approach of Chambers and Ren offers a feasible way of adjusting an imputed dataset to satisfy first order properties of the data. As far as point estimation of totals and means are concerned, the calibrated data allow the imputation–based estimates to reproduce chosen valid targets. The reverse calibration approach can be related to other calibrated imputation methods in Beaumont (2005), Favre et al. (2005), and Chauvet et al. (2011). All these methods use first order constraints and variance estimation has to be undertaken by specific methods for imputed data.

There are basically two types of variance estimation methods based on imputed data under the frequentist framework of inference. Kim and Rao (2009) outline a unified linearization approach to survey data after imputation of item nonresponse. The other class of methods are based on replication procedures, such as the Jackkinfe (Rao and Shao 1992) or the Bootstrap (Shao and Sitter 1996). These methods are developed to be used with a single imputed dataset. However, while these techniques are feasible for the data producer, they are not easy or even applicable by the secondary data analyst. This may be due to the additional complexities inherent to the replication procedure, need of specific software, or information about the assumed response process and imputation models that are not available in the disseminated data.

However, as it is of a common survey interest to address variance estimation of the point estimates, control of the second moment of the imputed data comes into play. This additional requirement has, under the perspective of facilitating secondary data analyses, the same motivation as the multiple imputation technique proposed by Rubin (1978, 1987), where the secondary data analyst can by means of a simple formula combine estimates from multiply imputed datasets in order to obtain measures of precision for the target estimates. We notice that Bjørnstad (2007) provides a frequentist modification of the multiple imputation variance formula, for several specific combinations of imputation method and nonresponse model.

The extension to the Chambers and Ren approach that we propose in this article operates by adding a second calibration condition so that the variances in the imputed dataset will also conform to given benchmarks. As a result, the calibrated imputed data will generally offer the secondary data user the possibility of computing valid point and variance estimates with simple full–sample formulas.

Our approach shares a similar spirit with Lanke (1983), in the sense the imputed values are modified so that complete–data formula for the imputed data gives desirable point and variance estimators. An important difference, however, is that the latter approach addresses only the case of simple random sampling with a missing completely at random response mechanism. Our method can be applied to complex surveys by means of the usual stratified ultimate cluster approximation (Skinner 1989, chap. 2) to variance estimation and, at least in principle, allows any nonresponse mechanism to be incorporated.

The paper is organized as follows. Section 2 describes the reverse calibration approach and the proposed extension of interest here. A more detailed discussion for this method under hot–deck imputation is given in Section 3 and the implementation of the method is illustrated in Section 4. Finally, in Section 5, we give an overview of the presented methodology and outline some other aspects to be investigated in further research.

2. Proposed approach

Consider a finite population of N units indexed by $U = \{1, 2, ..., N\}$ and let y_i be the value of a survey variable y for the i–th unit, $i \in U$. Some population characteristics of this variable that may be of interest are the total $t_y = \sum_{i \in U} y_i$, the mean $\bar{y}_U = \sum_{i \in U} y_i/N$ and the variance $S_y^2 = \sum_{i \in U}(y_i - \bar{y}_U)^2/(N-1)$. Suppose that the information about y is initially intended to be collected on a probability sample A of size n from U. We assume that A is selected according to a sampling design $p(\cdot)$, which does not depend on the variable y and yields positive inclusion probabilities $\pi_i = \Pr\{i \in A\}$ and $\pi_{ik} = \Pr\{i \in A, k \in A\}$. We also assume that it is available a known set of survey weights $\{w_i : i \in A\}$ that, under full observation of the values $\{y_i : i \in A\}$, could be applied to yield valid estimates of the population parameters. One possible specification for those weights is $w_i = \pi_i^{-1}$, which yields the weighted estimator of the population total t_y

$$\widehat{t}_y = \sum_{i \in A} w_i y_i = \sum_{i \in A} \pi_i^{-1} y_i. \tag{1}$$

This estimator is design unbiased in the sense that $E_p(\widehat{t}_y) = t_y$, where $E_p(\cdot)$ denotes expectation with respect to the sampling design.

Suppose that, after carrying out the survey, r $(r < n)$ units responded to the item y and $m = n - r$ values were missing. Defining R_i as a response indicator having the value 1, if y_i is observed, and the value 0, if the y_i is unobserved, the observed and missing parts of the sample A are respectively $A_r = \{i \in A : R_i = 1\}$ and $A_m = \{i \in A : R_i = 0\}$. Consider the use of imputation to create the complete dataset $D = \{(w_i, y_i^*) : i \in A\}$, where w_i are the design weights and y_i^* is either the observed value of y_i, if $i \in A_r$, or an imputed value for y_i, when $i \in A_m$.

However, standard full-sample formula, when applied to the imputed dataset D for point and variance estimation, generally do not produce valid results. For instance, in the case of the population total t_y, the analogous imputed estimator to (1) is

$$\widehat{t}_{yI} = \sum_{i \in A} w_i y_i^* \tag{2}$$

with the associated variance estimator

$$\widehat{V}_F(\widehat{t}_{yI}) = \sum_{i \in A} \sum_{k \in A} \Omega_{ik} w_i y_i^* w_k y_k^*, \tag{3}$$

where $\Omega_{ik} = (\pi_{ik} - \pi_i \pi_k)/\pi_{ik}$ and the subscript F in the variance estimator is to denote that the formula used is based on the formula that would be used under full response. The reason that (2) and (3) do not yield valid inferences is because the sampling weights w_i and variance coefficients Ω_{ik} correspond to the situation where the y_i are fully observed.

2.1 Estimation of population totals

We review here the main idea of the reverse calibration approach of Chambers and Ren (2004) in the context of estimating the population total t_y. However, instead of their setting where imputation is used to handling outliers in data, we focus here on the treatment of nonresponse. Suppose a working dataset is completed by taking $\{(w_i, R_i y_i + (1 - R_i)\widetilde{y}_i) : i \in A\}$, where $\widetilde{y}_i$ is an initial imputed value for y_i, if $R_i = 0$.

Suppose it is possible to obtain an unbiased or consistent estimate of t_y, namely $\widehat{t}_{y0}$. For instance, this correct estimate may correspond in the present context to the reweighted

estimator $\widehat{t}_{ywr} = \sum_{i \in A_r} w_{ir} y_i$, where $w_{ir} = w_i / \widehat{\phi}_i$ and $\widehat{\phi}_i$ is an estimate of the response probability of the i–th unit, $i \in A_r$. Hence, the aim of reverse calibration is to create the complete dataset $\{(w_i, y_i^*) : i \in A\}$ where the values y_i^* are chosen as close as possible to the initial values $R_i y_i + (1 - R_i)\widetilde{y}_i$ so that the completed sample imputed estimator

$$\widehat{t}_{yI} \equiv \sum_{i \in A} w_i y_i^* = \widehat{t}_{y0}. \tag{4}$$

The practical appeal of property (4) is that it allows a simple standard survey total estimator to reproduce the correct estimate $\widehat{t}_{y0}$ based on the calibrated dataset $\{(w_i, y_i^*) : i \in A\}$, regardless of how complicated $\widehat{t}_{y0}$ is. The terminology *reverse calibration* reflects the facts that the imputed variable y_i^* plays the role of the auxiliary variables in the traditional calibration setting of Deville and Särndal (1992) and the roles of y_i^* and w_i are interchanged (Chambers and Ren 2004, p. 3337).

In order to find an expression for the calibrated values y_i^*, Chambers and Ren (2004) first restrict $y_i^* = y_i$ if $i \in A_r$, so that the required task to satisfy condition (4) is equivalent to the task of finding values $\{y_j^* : j \in A_m\}$ such that

$$\sum_{j \in A_m} w_j y_j^* = \widehat{t}_{y0} - \sum_{i \in A_r} w_i y_i \equiv \widehat{t}_{ym}. \tag{5}$$

Then, they propose to obtain the y_j^* as the values that minimize the distance function

$$d_1(y^*, \widetilde{y}) = \sum_{j \in A_m} (y_j^* - \widetilde{y}_j)^2 / 2 q_j \widetilde{y}_j \tag{6}$$

subjected to condition (5), where $q_j > 0$ are known constants to be specified and is assumed that $\widetilde{y}_j > 0$ for all $j \in A_m$. The resulting solution to this optimization problem is given by

$$y_j^* = \widetilde{y}_j \left[1 + q_j w_j \frac{\widehat{t}_{ym} - \sum_{j \in A_m} w_j \widetilde{y}_j}{\sum_{j \in A_m} q_j w_j^2 \widetilde{y}_j} \right], \ j \in A_m, \tag{7}$$

which can be seen easily to satisfy (5) and, as a consequence, allow the imputed estimator to satisfy (4). In (7), many possible choices can be taken for the q_j. For instance, one possible choice is simply $q_j = 1$. Another choice is $q_j = w_j^{-1}$, which calibrates the initial imputed $\widetilde{y}_j$ values through the ratio adjustment $\widehat{t}_{ym} / \sum_{i \in A_m} w_i \widetilde{y}_j$.

2.2 Variance estimation

One motivation to extend the reverse calibration approach is that condition (4), a "first–moment consistency" property for the imputed dataset, allows only reproduction of the point estimates. Suppose that, in addition to (4), it is desirable that the imputed dataset also enables simple full–sample variance estimation procedures to reproduce a valid target variance for $\widehat{t}_{yI}$, say $\widehat{v}_{y0}$. This is what we will term as a "second–moment consistency" of the imputed dataset. We assume that $\widehat{v}_{y0}$ does not depend on the calibrated data, but it may or may not depend on the initial imputed values $\widetilde{y}_j$, $j \in A_m$. For example, variance estimates under missing–at-random assumptions can be derived without using the imputed values by considering the available information at the respondents of item y. Approaches, such as those in Rao and Shao (1992) and Shao and Sitter (1996), are examples of variance estimators that also take into account the imputed values.

When it comes to the full-sample variance estimator, one could choose the imputed variance $\widehat{V}_F(\widehat{t}_{yI})$ given in (3). In simple sampling designs, the use of this variance estimator is indeed feasible for our purposes, where

$$\widehat{V}_F(\widehat{t}_{yI}) = \frac{N^2(1-n/N)}{n(n-1)} \sum_{i\in A}(y_i^* - \bar{y}^*)^2,$$

and $\bar{y}^*$ is the mean of the imputed data. However, for complex designs, (3) generally does not have a simple computing expression so it will require the release of the matrix of Ω_{ik}, which is impractical. Hence, we consider instead the simpler full–sample variance estimator of $\widehat{t}_{yI}$, as if the sample had been selected with replacement, which is given by

$$\widehat{V}_F(\widehat{t}_{yI}) = \frac{n}{n-1} \sum_{i\in A}(u_i^* - \bar{u}^*)^2, \tag{8}$$

where $u_i^* = w_i y_i^*$, $i \in A$, and $\bar{u}^* = n^{-1}\sum_A u_i^* = n^{-1}\widehat{t}_{yI}$. We assume here an unstratified design for the sake of simplicity. The extension of (8) to stratified multistage sampling is based on considering the stratified version of (8).

The extended reverse calibration we propose in this article is based then on finding the values y_i^* as close as possible to the the values $R_i y_i + (1-R_i)\widetilde{y}_i$ so that $\sum_{i\in A} w_i y_i^* = \widehat{t}_{y0}$ and $\widehat{V}_F(\widehat{t}_{yI}) = \widehat{v}_{y0}$ are simultaneously satisfied. These requirements impose the two calibration conditions

$$\begin{aligned} \sum_{j\in A_m} w_j y_j^* &= \widehat{t}_{y0} - \sum_{i\in A_r} w_i y_i \equiv \widehat{t}_{ym} \\ \sum_{j\in A_m} w_j^2 {y_j^*}^2 &= \frac{n-1}{n}\widehat{v}_{y0} - \left(\sum_{i\in A_r} w_i^2 y_i^2 - \frac{\widehat{t}_{y0}^2}{n}\right) \equiv \widehat{t}_{yym}, \end{aligned} \tag{9}$$

where we assume that $\widehat{t}_{y0}$ and $\widehat{v}_{y0}$ are such that $\widehat{t}_{yym}$ is strictly positive. Again, keeping the restriction that $y_i^* = y_i$ if $i \in A_r$, then the y_j^* ($j \in A_m$) could be found by minimizing the distance function $d_1(y^*, \widetilde{y})$, in (6), subjected to (9). However, instead of d_1, we use a more general expression

$$d_2(y^*, \widetilde{y}) = \sum_{j\in A_m} (y_j^* - \widetilde{y}_j)^2/2Q_j \tag{10}$$

where $Q_j > 0$ does not depend on y_j^* but can depend on w_j and $\widetilde{y}_j$ in any chosen way. By applying the Lagrange multipliers method, the solution can be seen to be given by

$$y_j^* \equiv y_j^*(\lambda_1, \lambda_2) = \frac{\widetilde{y}_j + \lambda_1 w_j Q_j}{1 - 2\lambda_2 w_j^2 Q_j}, \;\; j \in A_m, \tag{11}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$ solves

$$\begin{aligned} f_1 \equiv f_1(\boldsymbol{\lambda}) &= \sum_{j\in A_m} w_j y_j^* - \widehat{t}_{ym} = 0 \\ f_2 \equiv f_2(\boldsymbol{\lambda}) &= \sum_{j\in A_m} w_j^2 {y_j^*}^2 - \widehat{t}_{yym} = 0. \end{aligned} \tag{12}$$

Because equations (12) cannot in general be solved analytically, a numerical approximation for $\boldsymbol{\lambda}$ must be sought. A simple method to obtain this approximation is the Newton–Raphson algorithm.

Next, we give some remarks to highlight the usefulness of the extended approach.

Remark 1 The proposed reverse calibration approach defined by (11) is feasible whenever the system of nonlinear equations (12) can be solved. The solution of these equations is to be undertaken "in office" and, hence, it may be seen as an additional stage of the preparation of the data before dissemination. The approach offers a useful procedure for secondary data analysts to obtain valid point and variance estimation in the presence of nonresponse and imputation. The usual complications associated with the correct treatment of nonresponse and the use of imputation are embedded into the computation of the targets $\widehat{t}_{y0}$ and $\widehat{v}_{y0}$. But these are tasks for survey analysts possessing resources and more information to guarantee that those estimates are indeed valid. Under this perspective, the secondary data user has only to apply the simple standard estimation procedures (2) and (8) on the single–imputed data to reproduce the valid targets.

Remark 2 Confidence intervals for the population total of t_y or smooth functions $g(t_y)$ can be easily constructed by applying standard software on the calibrated data. Conditions to make these inferential procedure valid are the usual assumptions (Haziza 2009, p. 235) of asymptotic normality and unbiasedness, or asymptotic unbiasedness, for the estimator that generated the target $\widehat{t}_{y0}$ and also the assumption of consistency of the estimator by which $\widehat{v}_{y0}$ is based upon. Then, by construction, the standard confidence interval based on the calibrated data

$$\widehat{t}_{yI} \ \pm \ z_{\alpha/2}\{\widehat{V}_F(\widehat{t}_{yI})\}^{1/2},$$

where $z_{\alpha/2}$ is the $1-\alpha$ quantile of the standard Normal distribution, has approximately $100(1-\alpha)\%$ coverage for t_y. This interval has wider scope than the one that could be obtained by applying the methods proposed by Lanke (1983) and discussed by Sedransk (1985), where the presented formulas relate simple random sampling under a uniform response mechanism. The interval also enables the secondary data analyst to use a simple formula to make inferences not only for the population total t_y, but also inferences for smooth functions $g(t_y)$ using

$$g(\widehat{t}_{yI}) \ \pm \ z_{\alpha/2}|g'(\widehat{t}_{yI})|\{\widehat{V}_F(\widehat{t}_{yI})\}^{1/2},$$

where $g'(\cdot)$ denotes the first derivative of $g(\cdot)$.

Remark 3 Stochastic or random imputation are commonly used procedures to compensate for item nonresponse surveys. A drawback is that the total variance of the imputed estimator will increase due to a non–zero imputation variance component. And it causes some extra difficulty for variance estimation. The application of the proposed reverse calibration approach can eliminate the extra imputation variance, while allowing the secondary user to obtain the valid variance estimate using the simple full-sample formula (8). An illustration will be given for hot-deck imputation in Section 3, with comparisons to the alternative approaches in the literature.

Remark 4 The response sample may contain representative outliers (Chambers and Ren 2004). Suppose that the benchmark targets $\widehat{t}_{y0}$ and $\widehat{v}_{y0}$ are based on some appropriate robust estimation methods which curtail the contributions of the outliers. The reverse calibration conditions (9) may need to be adjusted accordingly. Let A_o denote the identified outliers, where $A_o \subset A_r$, and replace (9) with

$$\sum_{j\in A_m} w_j y_j^* + \sum_{j\in A_o} w_j y_j^* = \widehat{t}_{y0} - \sum_{i\in A_r\setminus A_o} w_i y_i$$

$$\sum_{j\in A_m} w_j^2 {y_j^*}^2 + \sum_{j\in A_o} w_j^2 {y_j^*}^2 = \frac{n-1}{n}\widehat{v}_{y0} - \left(\sum_{i\in A_r\setminus A_o} w_i^2 y_i^2 - \frac{\widehat{t}_{y0}^2}{n}\right),$$

i.e. leading to imputation of both the outliers and the nonrespondents. The first constraint was proposed by Chambers and Ren (2004). The second one is included under the our approach to facilitate full-sample variance formula based on the imputed dataset. We notice that the choice of Q_j in d_2 may differ for the units in A_m and A_o.

2.3 Preserving variability

It is sometimes argued that, for certain data analyses, the imputed values should preserve the variance of the observed item values of y (Sedransk 1985, p. 448). This property is connected with the estimation of the finite population variance S_y^2 using the imputed data. One way to attain this property by the proposed approach is as follows. Let

$$\widehat{t}_{yyI} = \sum_{i \in A} w_i y_i^{*2}$$

be the completed sample imputed estimator of the population total of y^2, that is $t_{yy} = \sum_{i \in U} y_i^2$. The weighted sample variance of the imputed data can be expressed as

$$\widehat{s}_{yI}^2 \equiv \frac{\sum_{i \in A} w_i (y_i^* - \bar{y}_I)^2}{\sum_{i \in A} w_i - 1} = \frac{\widehat{t}_{yyI}}{\widehat{N} - 1} - \frac{\widehat{t}_{yI}^2}{\widehat{N}(\widehat{N} - 1)},$$

where $\bar{y}_I = \sum_{i \in A} w_i y_i^* / \sum_{i \in A} w_i$ is the weighted sample mean of the imputed dataset and $\widehat{N} = \sum_{i \in A} w_i$. A second–moment consistency property for the imputed dataset could be phrased so that $\widehat{s}_{yI}^2$ reproduces a given $\widehat{s}_{y0}^2$. This could be taken, for instance, as the reweighted estimator $s_{ywr}^2 = \sum_{i \in A_r} w_{ir} (y_i - \widehat{t}_{ywr} / \sum_{i \in A_r} w_{ir})^2 / [\sum_{i \in A_r} w_{ir} - 1]$. The second–moment property can be attained by minimizing the distance function $d_2(y^*, \widetilde{y})$ constrained to $\widehat{t}_{yI} = \widehat{t}_{y0}$ and $\widehat{s}_{yI}^2 = \widehat{s}_{y0}^2$. In this case, the optimal y_j^* ($j \in A_m$) values are

$$y_j^* = \left[\frac{1}{Q_j} - \frac{2\lambda_2}{\widehat{N} - 1} \right]^{-1} \left[\frac{\widetilde{y}_j}{Q_j} + \lambda_1 - \frac{2\lambda_2 \widehat{t}_{y0}}{\widehat{N}(\widehat{N} - 1)} \right], \; j \in A_m, \tag{13}$$

and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$ as the solution to (12) using $\widehat{t}_{ym} = \widehat{t}_{y0} - \sum_{A_r} w_i y_i$ and

$$\widehat{t}_{yym} = (\widehat{N} - 1)\widehat{s}_{y0}^2 + \sum_{i \in A_r} w_i y_i^2 - \frac{\widehat{t}_{y0}^2}{\widehat{N}}. \tag{14}$$

Again, since the solution can not be usually expressed analytically an one could apply Newton–Raphson algorithm.

It is important to notice that $\widehat{t}_{yym}$ in (14) is generally different from $\widehat{t}_{yym}$ in (12). It follows that it is generally not possible to calibrate the data to satisfy both $\widehat{s}_{yI}^2 = \widehat{s}_{y0}^2$ and $\widehat{V}_F(\widehat{t}_{yI}) = \widehat{v}_{y0}$, so that a decision must be taken on the estimation targets to be reproduced. Another possibility is to disseminate two calibrated variables according to each estimation purpose. Either way, it serves as a reminder that genuinely all-purpose *single* imputation is impossible. Only the true data can do that.

One possible extension for this preservation of variability property is the estimation of covariances and correlation coefficients between survey variables. This can be achieved by adding constraints relative to the cross–product totals. Some earlier work with this goal is Cohen (2003). A recent reference, taking into account constraints in the imputed data, is Gelein et al. (2014).

3. Hot–deck imputation

To focus on the main idea, consider simply the estimation of the population total t_y based on a subsample A_r of r respondents of a simple random sample without replacement A of size n. Assume that the nonresponse process follows a uniform response mechanism so that missing observations are imputed by hot–deck imputation, where the imputed values $\widetilde{y}_j$, $j \in A_m$, are chosen by $m = n - r$ independent random draws from the set $\{y_i : i \in A_r\}$ with replacement. Based on the data $\{(w_i, R_i y_i + (1 - R_i)\widetilde{y}_i) \ \ i \in A\}$, the resulting hot–deck imputed estimator of t_y is

$$\widehat{t}_{yI,HD} \equiv \sum_{i \in A} w_i \widetilde{y}_j = \frac{r}{n} N \bar{y}_r + \frac{m}{n} N \bar{\widetilde{y}}_m, \tag{15}$$

where $\bar{y}_r = r^{-1} \sum_{i \in A_r} y_i$ is the mean of respondents and $\bar{\widetilde{y}}_m = m^{-1} \sum_{j \in A_m} \widetilde{y}_j$ is the mean of the imputed observations for the nonrespondents.

The unconditional expectation of this estimator with respect to the joint distribution of the sampling (p), response (R) and imputation (I) mechanisms is given by

$$E[\widehat{t}_{yI,HD}] = E_{pR}[E_I(\widehat{t}_{yI,HD} \mid A, A_r)]. \tag{16}$$

Because the imputation expectation $E_I[\bar{\widetilde{y}}_m \mid A, A_r] = \bar{y}_r$, it follows that $E_I[\widehat{t}_{yI,HD} \mid A, A_r] = N\bar{y}_r$ and, as $\bar{y}_r$ is pR–unbiased for the population mean of y, $\widehat{t}_{yI,HD}$ is unbiased for t_y. For the variance of (15), we note first that the imputation variance $V_I[\bar{\widetilde{y}}_m \mid A, A_r] = (r-1)s^2_{yr}/(mr)$, where $s^2_{yr} = \sum_{i \in A_r}(y_i - \bar{y}_r)^2/(r-1)$ is the variance of the respondents. Then, the unconditional variance of $\widehat{t}_{yI,HD}$ is

$$\begin{aligned} V(\widehat{t}_{yI,HD}) &= V_{pR}[E_I(\widehat{t}_{yI,HD} \mid A, A_r)] + E_{pR}[V_I(\widehat{t}_{yI,HD} \mid A, A_r)] \\ &= V_{pR}[N\bar{y}_r] + E_{pR}\left[N^2 \frac{m^2}{n^2} \frac{r-1}{r} \frac{s^2_{yr}}{m}\right]. \end{aligned} \tag{17}$$

The second component in (17) is an inflation term that is due to the positive hot–deck imputation variance, which is a main drawback of hot-deck and any other unconstrained random imputation methods. One can modify the hot-deck imputation in a way that avoids the imputation variance. However, unconstrained hot-deck or not, the secondary user will need extra effort to apply special variance estimation method based on the imputed data, as explained below.

To start with, both Bayesian and non-Bayesian methods can be used to estimate the variance of unconstrained hot-deck imputed sample mean. Under the approximate Bayesian bootstrap (Rubin and Schenker, 1986), one needs to resample the full data set including the nonresponse indicators, so that the number of missing observations as well as the observed y-values will vary form one bootstrap replicate sample to another. Notice that, Kim (2002) demonstrated that the corresponding variance estimator has a non–negligible negative bias in moderate sample sizes, which could be minimized upon a modification in the algorithm underlying the approximate Bayesian bootstrap method. For a frequentist approach, Rao and Shao (1992) propose a jackknife method. For each jackknife replicate sample where an observed value is deleted, all the imputed values need to be adjusted by an amount that is equal to the difference between the jackknife replicate response sample mean and the initial response sample mean, to reflect the fact the donor set if changed by jackknife.

Next, the data provider can apply a constrained hot-deck imputation that does not result in extra imputation variance, by modifying the m imputed values so that their mean is equal $\bar{y}_r$. One method by Chauvet et al. (2011) uses the idea of balanced sampling and employs

a special algorithm to select the residuals to adjust the imputed data in order to satisfy the benchmark mean constraint. It is possible to incorporate additional auxiliary variables available for the full sample.

A simpler method to eliminate the imputation variance in (17), in this situation without auxiliary variables, is the adjusted random imputation method of Chen et al. (2000). It uses the imputed values $\widetilde{y}_j = \bar{y}_r + (\bar{z}^*_j - \bar{z}^*_m)$ for $j \in A_m$, where $\bar{z}^*_j$ is an independent random draw from $\{y_i : i \in A_r\}$ with replacement, and $\bar{z}^*_m = m^{-1}\sum_{j\in A_m} \bar{z}^*_j$. Hence, by construction, $E_I[\bar{\widetilde{y}}_m \,|\, A, A_r] = \bar{y}_r$ and $V_I[\bar{\widetilde{y}}_m \,|\, A, A_r] = 0$.

The variance of the imputed estimator (15), provided it is constrained to the observed response sample mean in one way or another, is given by

$$V(\widehat{t}_{yI,HD}) = V_{pR}[N\bar{y}_r] = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + N^2 E_p\left\{E_R\left[\left(\frac{1}{r} - \frac{1}{n}\right)s_y^2 \,|\, A\right]\right\}, \quad (18)$$

where $E_R[\cdot \,|\, A]$ denotes the expectation under the conditional distribution of the observed number of respondents, r, given the original sample A. Since $E_R[s^2_{yr} \,|\, A, r] = s_y^2$ and $E_{pR}[s^2_{yr}] = S_y^2$, an unbiased estimator of $V(\widehat{t}_{yI,HD})$ is

$$\widehat{V}_{unb} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s^2_{yr} + N^2\left(\frac{1}{r} - \frac{1}{n}\right)s^2_{yr} = N^2\left(\frac{1}{r} - \frac{1}{N}\right)s^2_{yr}. \quad (19)$$

The sample variance of the imputed values is given by

$$s^2_{yI} = \frac{1}{n-1}\left[\sum_{i\in A_r}(y_i - \bar{y}_I)^2 + \sum_{j\in A_m}(\widetilde{y}_j - \bar{y}_I)^2\right]$$

where $\bar{y}_I = (r\bar{y}_r + m\bar{\widetilde{y}}_m)/n = \bar{y}_r$, and $E_I(s^2_{yI} \,|\, A, A_r) = (r-1)s^2_{yr}/r \doteq s^2_{yr}$ when r is large. So the full–sample SRS variance estimator, which uses n^{-1} instead of r^{-1}, is expected to under–estimate the variance. An approximate unbiased estimator of (18) can be obtained on replacing s^2_{yr} by s^2_{yI} in (19). However, without explicitly deriving (19) first, one would not be able to clarify the relationship even in this special case. Thus, Chen et al. (2000) suggest the jackknife variance estimator, which requires adjusting the jackknife imputed values so that the jackknife variance estimator is algebraically equal to the variance estimator for mean imputation.

In short, whether the hot–deck imputation is unconstrained or constrained, the secondary user is left to deal with the problem of variance estimation. The reverse calibration approach is helpful since, on the one hand, it provides a simple means for the data provider to obtain constrained random imputation data without the extra imputation variance, and, on the other hand, it allows the secondary user to obtain valid variance estimate using simple full–sample formula, without the need to apply a special variance estimation method tailored for the missing data problem.

Put the estimation targets $\widehat{t}_{y0} = N\bar{y}_r$ and $\widehat{v}_{y0} = \widehat{V}_{unb}$, where $\widehat{V}_{unb}$ is given in (19). Let $\widetilde{y}_j$, $j \in A_m$, denote the initial imputed values obtained by unconstrained hot–deck imputation. Let the calibrated imputed values be given by (11) using $w_j = N/n$, $Q_j = 1$, and (λ_1, λ_2) as the solution of (12) with $\widehat{t}_{ym} = \widehat{t}_{y0} - \sum_{i\in A_r} w_i y_i = N\bar{y}_r(1 - r/n)$ and $\widehat{t}_{yym} = (n-1)\widehat{v}_{y0}/n - (\sum_{i\in A_r} w_i^2 y_i^2 - \widehat{t}^2_{y0}/n)$. Hence, by applying the original full–sample weights $w_i = N/n$ to the calibrated data y^*_i, $i \in A$, we have by construction that

$$\widehat{t}_{yI} \equiv \sum_{i\in A} w_i y^*_i = N\bar{y}_r = N\bar{y}^*$$

where $\bar{y}^*$ is simply the imputed sample mean, and, taking with $u_i^* = w_i y_i^*$,

$$\widehat{V}_F(\widehat{t}_{yI}) = \frac{n}{n-1}\sum_{i\in A}(u_i^* - \bar{u}^*)^2 = N^2\frac{s^{*\,2}}{n} = \widehat{V}_{unb}$$

where $s^{*\,2}$ is simply the sample variance of the imputed dataset. Evidently, the calibration condition based on the first target $\widehat{t}_{y0} = N\bar{y}_r$ eliminates the imputation variance of the unconstrained hot–deck imputation, and the second constraint ensures that the valid variance estimate $\widehat{v}_{y0} = \widehat{V}_{unb}$ is reproduced for $\widehat{t}_{yI}$. The full-sample formulas allows the secondary user to treat the imputed sample as if it were a simple random sample without nonresponse, which is particularly convenient.

4. Numerical illustration

In this section, we illustrate the implementation of the reverse calibration approach for mean and hot–deck imputation. We consider an artificial finite population of $N = 300$

Table 1: Observed dataset

i	w_i	y_{i1}	y_{i2}
1	10	$*$	$*$
2	10	$*$	5.92
3	10	95.97	5.58
4	10	101.76	13.16
5	10	101.01	12.50
6	10	99.32	$*$
7	10	$*$	$*$
8	10	91.90	$*$
9	10	$*$	11.29
10	10	100.24	4.95
11	10	$*$	6.26
12	10	103.73	12.04
13	10	93.75	9.62
14	10	$*$	11.43
15	10	97.32	12.57
16	10	102.41	6.77
17	10	111.20	10.39
18	10	95.61	$*$
19	10	$*$	19.21
20	10	97.42	13.84
21	10	$*$	4.45
22	10	$*$	$*$
23	10	$*$	10.46
24	10	$*$	5.41
25	10	$*$	7.36
26	10	108.51	6.61
27	10	$*$	5.90
28	10	$*$	4.29
29	10	$*$	12.77
30	10	$*$	8.10

elements with two items y_1 and y_2. The values of these variables were generated from the $N(100, 5^2)$ and $G(shape = 10/3, scale = 3)$ distributions, rounded to two decimal places. The dataset on Table 1 corresponds to a nonreplacement simple random sample of $n = 30$ observations from this population. The columns of this table gives the identifier, sampling weight and observed values of y_1 and y_2 for the elements in the sample. The responding values on these two variables were generated by independent Bernoulli sampling with response probabilities 0.5 and 0.7, respectively. Taking into account the sampling design and the uniform response mechanism, valid targets for the imputed estimators and its variances are based on the reweighted estimates $\widehat{t}_{y0} = N\bar{y}_r$ and $\widehat{v}_{y0} = N^2\left(r^{-1} - N^{-1}\right)s^2_{yr}$, where $\bar{y}_r$ and s^2_{yr} are the sample mean and variance of the respondent elements for an item y. The values of these estimates are given in Table 2.

Table 2: Unbiased estimates of the population totals of y_1 and y_2 and respective unbiased variance estimates

Item	$\widehat{t}_{y0}$	$\widehat{v}_{y0}$
y_1	30003.21	177370.99
y_2	2761.00	49903.76

Table 3 gives a working dataset created to demonstrate the computing steps of the reverse calibration approach for both variables. In this table, R_{i1} and R_{i2} are the item response indicators; column $\widetilde{y}_{i1}$ is the imputed variable for y_1 obtained by replacing the missing values of that variable by the mean of the item responding units $\bar{y}_{1r} = 100.011$ plus a zero mean random noise. The addition of this random noise is because mean imputation has no variation among the imputed data, making then impossible to calibrate those values to satisfy both the first and second order constraints; column $\widetilde{y}_{i2}$ corresponds to entries in y_{i2} with the missing values imputed by hot–deck imputation. The values were selected by independent draws from the item respondent values; and, finally, the entries y^*_{i1} and y^*_{i2} are the reverse calibrated values obtained by (11) and (12) using as preliminary imputed values the modified mean imputed variable $\widetilde{y}_{i1}$ and the hot–deck imputed $\widetilde{y}_{i2}$, respectively.

The computing process to apply the reverse calibration method on the initial imputed variables $\widetilde{y}_{i1}$ and $\widetilde{y}_{i2}$ is as follows. Given the estimates in Table 2, the values of the calibration constraints are by (9)

$$\widehat{t}_{ym} = \widehat{t}_{y0} - \sum_{i\in A_r} w_i y_i = \begin{cases} 16{,}001.71, & \text{for } y_1, \\ 552.20, & \text{for } y_2, \end{cases}$$

$$\widehat{t}_{yym} = \frac{n-1}{n}\widehat{v}_{y0} - \left(\sum_{i\in A_r} w_i^2 y_i^2 - \frac{\widehat{t}^2_{y0}}{n}\right) = \begin{cases} 16{,}137{,}263.23, & \text{for } y_1, \\ 65{,}791.93, & \text{for } y_2. \end{cases}$$

Hence, the calibrated values y^*_{i1} and y^*_{i2} in Table 3 can be computed by running separately the reverse calibration procedure, defined by (11) and (12), with $Q_j = 1$ and the corresponding values of $\widehat{t}_{ym}$ and $\widehat{t}_{yym}$ above, for each case. Note the reverse calibration procedure changes, by construction, only the initial imputed values and not any of the observed values. The dataset to be disseminated could be formatted as in the simple layout of Table 4, with a single set of weights and a single imputed variable for each item. The secondary user needs just to use the simple expressions (2) and (8) to obtain the reverse calibration estimates of the population total and its variance. These estimates follow easily by noting that the sample total and variance of the values $\{w_i y^*_{i1} : i \in A\}$ and $\{w_i y^*_{i2} : i \in A\}$ are

Table 3: Working dataset for calibrating the imputed data

i	w_i	R_{i1}	y_{i1}	$\widetilde{y}_{i1}$	y^*_{i1}	R_{i2}	y_{i2}	$\widetilde{y}_{i2}$	y^*_{i2}
1	10	0	$*$	96.25	95.14775	0	$*$	13.16	14.63372
2	10	0	$*$	102.98	107.79921	1	5.92	5.92	5.92000
3	10	1	95.97	95.97	95.97000	1	5.58	5.58	5.58000
4	10	1	101.76	101.76	101.76000	1	13.16	13.16	13.16000
5	10	1	101.01	101.01	101.01000	1	12.50	12.50	12.50000
6	10	1	99.32	99.32	99.32000	0	$*$	12.50	13.62223
7	10	0	$*$	99.34	100.95652	0	$*$	4.95	2.05137
8	10	1	91.90	91.90	91.90000	0	$*$	10.39	10.38852
9	10	0	$*$	96.07	94.80937	1	11.29	11.29	11.29000
10	10	1	100.24	100.24	100.24000	1	4.95	4.95	4.95000
11	10	0	$*$	100.41	102.96797	1	6.26	6.26	6.26000
12	10	1	103.73	103.73	103.73000	1	12.04	12.04	12.04000
13	10	1	93.75	93.75	93.75000	1	9.62	9.62	9.62000
14	10	0	$*$	93.46	89.90294	1	11.43	11.43	11.43000
15	10	1	97.32	97.32	97.32000	1	12.57	12.57	12.57000
16	10	1	102.41	102.41	102.41000	1	6.77	6.77	6.77000
17	10	1	111.20	111.20	111.20000	1	10.39	10.39	10.39000
18	10	1	95.61	95.61	95.61000	0	$*$	11.29	11.76782
19	10	0	$*$	102.98	107.79921	1	19.21	19.21	19.21000
20	10	1	97.42	97.42	97.42000	1	13.84	13.84	13.84000
21	10	0	$*$	93.91	90.74887	1	4.45	4.45	4.45000
22	10	0	$*$	90.53	84.39494	0	$*$	5.41	2.75635
23	10	0	$*$	107.17	115.67583	1	10.46	10.46	10.46000
24	10	0	$*$	91.48	86.18081	1	5.41	5.41	5.41000
25	10	0	$*$	103.27	108.34437	1	7.36	7.36	7.36000
26	10	1	108.51	108.51	108.51000	1	6.61	6.61	6.61000
27	10	0	$*$	105.20	111.97250	1	5.90	5.90	5.90000
28	10	0	$*$	98.76	99.86620	1	4.29	4.29	4.29000
29	10	0	$*$	96.66	95.91849	1	12.77	12.77	12.77000
30	10	0	$*$	102.92	107.68642	1	8.10	8.10	8.10000

30003.21 and 5912.366 for y_1 and 2761 and 1663.459 for y_2. Then,

$$\widehat{t}_{y_{1I}} = \sum_{i \in A} w_i y^*_{i1} = 30003.21, \quad \widehat{t}_{y_{2I}} = \sum_{i \in A} w_i y^*_{i2} = 2761,$$

$$\widehat{V}_F(\widehat{t}_{y_{1I}}) = n[\text{sample variance of } (w_i y^*_{i1})] = 30(5912.366) = 177371 \quad \text{and}$$

$$\widehat{V}_F(\widehat{t}_{y_{2I}}) = n[\text{sample variance of } (w_i y^*_{i2})] = 30(1663.459) = 49903.76.$$

Hence, the standard expressions for the imputed estimator and its variance, apart from rounding error, reproduce the targets $\widehat{t}_{y0}$ and $\widehat{v}_{y0}$ for both cases. Note the estimates are produced without making use of the item response indicators.

A nicer looking output for the calibrated dataset in Table 4 is obtained by rounding the calibrated variables y^*_{i1} and y^*_{i2} to have the same number of decimal places as the original values. Of course, the rounded calibrated data may no longer match exactly the calibration constraints. For example, by rounding y^*_{i1} and y^*_{i2} to two decimal places, the resulting estimates are $\widehat{t}_{y_{1I}} = 30003.21$, $\widehat{t}_{y_{2I}} = 2761$, $\widehat{V}_F(\widehat{t}_{y_{1I}}) = 177370.4$ and $\widehat{V}_F(\widehat{t}_{y_{2I}}) = 49905.08$. A more refined rounding algorithm may be used if it is deemed necessary.

Table 4: Calibrated dataset for dissemination

i	w_i	R_{i1}	y^*_{i1}	R_{i2}	y^*_{i2}
1	10	0	95.14775	0	14.63372
2	10	0	107.79921	1	5.92000
3	10	1	95.97000	1	5.58000
4	10	1	101.76000	1	13.16000
5	10	1	101.01000	1	12.50000
6	10	1	99.32000	0	13.62223
7	10	0	100.95652	0	2.05137
8	10	1	91.90000	0	10.38852
9	10	0	94.80937	1	11.29000
10	10	1	100.24000	1	4.95000
11	10	0	102.96797	1	6.26000
12	10	1	103.73000	1	12.04000
13	10	1	93.75000	1	9.62000
14	10	0	89.90294	1	11.43000
15	10	1	97.32000	1	12.57000
16	10	1	102.41000	1	6.77000
17	10	1	111.20000	1	10.39000
18	10	1	95.61000	0	11.76782
19	10	0	107.79921	1	19.21000
20	10	1	97.42000	1	13.84000
21	10	0	90.74887	1	4.45000
22	10	0	84.39494	0	2.75635
23	10	0	115.67583	1	10.46000
24	10	0	86.18081	1	5.41000
25	10	0	108.34437	1	7.36000
26	10	1	108.51000	1	6.61000
27	10	0	111.97250	1	5.90000
28	10	0	99.86620	1	4.29000
29	10	0	95.91849	1	12.77000
30	10	0	107.68642	1	8.10000

5. Discussion

The proposed reverse calibration approach gives a way of calibrating a single dataset in the form $\{(w_i, y^*_i) : i \in A\}$ so that valid estimates of the population total and its variance estimator can be recovered by a weighted imputed estimator and a simple complete–sample variance formula. The approach is feasible whenever the system of equations (9) can be solved. The advantage of this approach is the possibility of reproduction of correct estimation of the parameter of interest under sampling and nonresponse using simple statistical standard formulas. This may benefit not only secondary data users that are external to the survey organization, but also users inside the organization that do not have the capability or information required to obtain the correct target estimates.

For this research, we have already developed an algorithm to allow the calibration method to be applied in a stratified multistage survey. For the future investigations, it is of interest to extend the proposed approach to a multivariate setting and also to parameter estimation in pre–specified population domains.

REFERENCES

Beaumont, J.-F. (2005), "Calibrated imputation in surveys under a quasi-model-assisted approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 445–458.

Bjørnstad, J. F. (2007), "Non-Bayesian Multiple Imputation," *Journal of Official Statistics*, 23, 433–452.

Chambers, R. L. and Ren, R. (2004), "Outlier Robust Imputation of Survey Data," in *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, pp. 3336–3344.

Chauvet, G., Deville, J.-C., and Haziza, D. (2011), "On balanced random imputation in surveys," *Biometrika*, 98, 459–471.

Chen, J., Rao, J. N. K., and Sitter, R. R. (2000), "Efficient Random Imputation for Missing Data in Complex Surveys," *Statistica Sinica*, 10, 1153–1169.

Cohen, M. P. (2003), "Imputation Allowing Standard Variance Formulas," *Journal of Data Science*, 1, 275–292.

Deville, J.-C. and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Favre, A.-C., Matei, A., and Tillé, Y. (2005), "Calibrated random imputation for qualitative data," *Journal of Statistical Planning and Inference*, 128, 411 – 425.

Gelein, B., Haziza, D., and Causeur, D. (2014), "Preserving relationships between variables with {MIVQUE} based imputation for missing survey data," *Journal of Multivariate Analysis*, 131, 197 – 208.

Haziza, D. (2009), "Imputation and Inference in the Presence of Missing Data," in *Handbook of Statistics, Volume 29A, Sample surveys: Design, methods and applications*, eds. Pfeffermann, D. and Rao, C. R., Amsterdam, The Netherlands: Elsevier, pp. 215–246.

Kim, J. K. (2002), "A note on approximate Bayesian bootstrap imputation," *Biometrika*, 89, 470–477.

Kim, J. K. and Rao, J. N. K. (2009), "A unified approach to linearization variance estimation from survey data after imputation for item nonresponse," *Biometrika*, 96, 917–932.

Lanke, J. (1983), "Hot deck imputation techniques that permit standard methods for assessing precision of estimates," *Statistical Review*, 21, (Essays in Honour of Tore E. Dalenius), 105–110.

Rao, J. N. K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," in *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 20–28.

— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Sedransk, J. (1985), "The Objectives and Practice of Imputation," in *Proceedings of the First Annual Research Conference*, U.S. Bureau of the Census, pp. 445–452.

Shao, J. and Sitter, R. R. (1996), "Bootstrap for Imputed Survey Data," *Journal of the American Statistical Association*, 91, 1278–1288.

Skinner, C. J. (1989), "Introduction to Part A," in *Analysis of Complex Surveys*, eds. Skinner, C. J., Holt, D., and Smith, T. M. F., Chichester: Wiley, pp. 23–58.