A calibrated imputation method for secondary data analysis of survey data

Da Silva, Damião N and Li-Chun Zhang

Da Silva, D.N., Zhang, L-C. A calibrated imputation method for secondary data analysis of survey data. **Scandinavian Journal of Statistics**. 2019; 1– 17.
https://doi.org/10.1111/sjos.12435

Received ; | Revised ; | Accepted

DOI: xxx/xxxx

ARTICLE TYPE

A Calibrated Imputation Method for Secondary Data Analysis of Survey Data.

Damião Nóbrega Da Silva*[1] | Li-Chun Zhang[2,3,4]

[1] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil
[2] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, Hampshire, UK
[3] Statistisk sentralbyrå, Oslo, Norway
[4] University of Oslo, Oslo, Norway

Correspondence
*Damião Nóbrega Da Silva, Departamento de Estatística, Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte, Natal, RN, 59078-970 Brazil. Email: damiao@ccet.ufrn.br

Funding Information

This research was supported by the Brazilian Research Council (CNPq), Grant Number: 211518/2013-1

Summary

In practical survey sampling, missing data are unavoidable due to nonresponse, rejected observations by editing, disclosure control or outlier suppression. We propose a calibrated imputation approach so that valid point and variance estimates of the population (or domain) totals can be computed by the secondary users using simple complete-sample formulae. This is especially helpful for variance estimation, which generally require additional information and tools that are unavailable to the secondary users. Our approach is natural for continuous variables, where the estimation may be either based on reweighting or imputation, including possibly their outlier-robust extensions. We also propose a multivariate procedure to accommodate the estimation of the covariance matrix between estimated population totals, which facilitates variance estimation of the ratios or differences among the estimated totals. We illustrate the proposed approach using simulation data in supplementary materials that are available online.

KEYWORDS:
analysis of incomplete data, item nonresponse, missing data, variance estimation

1 | INTRODUCTION

In the preparation of survey data for use by secondary analysts, some or all of the sample units are usually assigned estimation weights that can be applied to all the survey variables. In addition to these weights, imputed values may be needed for the units that are subjected to item missingness. It is often possible to choose the imputed values for each survey variable so that, together with the observed and retained values of this variable, the corresponding population total can be estimated by weighting as if

[0] **Abbreviations:** ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

the sample were completely observed. However, applying standard complete-sample variance estimator formulae to the same imputed sample would generally cause bias (see, e.g., Wolter, 2007, pp. 419, 421).

Variance estimation in the presence of imputed data needs to appropriately account for the underlying data generation mechanism. Some common techniques are Fay's reverse framework (e.g. Fay, 1991, 1992; Shao & Steel, 1999; Kim & Rao, 2009), two-phase sampling (Särndal, 1992; Deville & Särndal, 1994b) and replication methods (Rao & Shao, 1992; Rao, 1996; Shao & Sitter, 1996; Chen, Rao, & Sitter, 2000). To choose and apply any of these methods may be difficult for secondary users who are non-specialists, even impossible if the relevant information about the sampling design and data processing is lacking.

The needs for easier secondary analyses ensuring that different users of the imputed data could obtain the same results by simple estimation methods have long been recognized (Kalton & Kasprzyk, 1982). Early work under this estimation perspective for the imputed data was addressed by Lanke (1983), Sedransk (1985) and Kim (2001a). Some later works considered the use of constrained or calibrated imputation for point estimation in different situations (Chambers & Ren, 2004; Beaumont, 2005; Chauvet, Deville, & Haziza, 2011; Gelein, Haziza, & Causeur, 2014). Multiple imputation (Rubin, 1978a, 1987b, 1996c; Rubin & Schenker, 1986) and fractional imputation (Kim & Fuller, 2004; Fuller & Kim, 2005) are two methods based on multiply imputed values. For instance, provided the multiple imputation procedure is proper or the congeniality condition (Meng, 1994) holds, one can compute valid point and variance estimates by combining the results obtained from applying standard complete-sample formulae to each imputed sample.

In this paper, we propose a calibrated imputation approach that allows for valid point and variance estimation of the population and domain totals (or means), by applying simple complete-sample formulae to the imputed sample. Although this accommodates a more limited scope than multiple or fractional imputation, secondary users can achieve the intended analyses based on a single imputed dataset using standard software. Moreover, we provide a multivariate procedure for the estimation of a vector of totals (or means) and the associated covariance matrix, using simple complete-sample formulae. This allows one to estimate the variance of ratios or differences of the estimated totals. Finally, the proposed approach has also benefits to the data producer, such as avoiding the dissemination of multiply imputed datasets, the freedom to choose a suitable inference outlook and apply different missing data treatments from one variable to another.

The rest of the paper is organized as follows. The proposed approach is outlined in Section 2. We explain in Section 3 how it can be applied in some common situations of reweighting and imputation-based estimation, as well as domain estimation and estimation under stratified multistage sampling. The multivariate calibrated imputation procedure is described in Section 4. Some concluding remarks and future research topics are given in Section 5.

2 | CALIBRATION OF A SINGLE VARIABLE

2.1 | Estimation Setup

Consider a finite population of N units denoted by $U = \{1, 2, ..., N\}$ and let $Y_N = (y_1, \dots, y_N)$, where y_k is the value of a survey variable y for the kth unit, $k \in U$. Let A be a sample from U selected by a probability sampling design $p(\boldsymbol{I}_N)$, where $\boldsymbol{I}_N = (i_1, \dots, i_N)$ and $i_k = 1$ if the kth unit is selected to the sample A and $i_k = 0$ otherwise, and let $\boldsymbol{R}_N = (r_1, \dots, r_N)$, where $r_k = 1$ if y_k is observed and $r_k = 0$ if the y_k is unobserved or rejected during data processing ($k \in U$), $A_r = \{k : k \in A, r_k = 1\}$ be the set of units for which the observations are to be preserved and $A_m = \{k : k \in A, r_k = 0\}$ be the set for which imputation is needed. We assume the missing information of the variable y for the units in A_m is filled in by imputation and denote the corresponding imputed values by $\{y_k^* : k \in A_m\}$. We assume, in addition, that the imputed dataset will be accompanied by a set of survey weights $\{w_k : k \in A\}$, as for instance, the inverse of the inclusion probabilities (Horvitz & Thompson, 1952), or weights suitably calibrated for auxiliary population totals (Deville & Särndal, 1992a).

Suppose it is of interest to estimate the population total of the variable y, that is $t_y = \sum_{k \in U} y_k$. When it comes to complete-sample estimation of t_y using the imputed data $\{(w_k, y_k^*) : k \in A\}$, where y_k^* is the value for unit k in the imputed full sample A with $y_k^* = y_k$ for $k \in A_r$, a natural and simple choice for the imputed estimator is

$$\hat{t}_{yI} = \sum_{k \in A} w_k y_k^*. \tag{1}$$

Statistical properties of $\hat{t}_{yI}$ are studied by adopting an *inference approach* for the imputed data, which is usually specified by postulating explicitly a model for the distribution of the response indicators or a superpopulation model for the values of the variables of interest in the population (Haziza, 2009, pp. 222-223). The properties of $\hat{t}_{yI}$ are then evaluated with respect to the joint distribution of the sampling design and the assumed model, allowing the unconditional variance of the imputed estimator to be decomposed into variance components which, when estimated, lead to the estimated variance of $\hat{t}_{yI}$.

Here we consider instead the estimation of the variance of the imputed estimator $\hat{t}_{yI}$ by means of the complete-sample estimator

$$\hat{v}_F(\hat{t}_{yI}) \equiv \frac{n}{n-1}\sum_{k\in A}(u_k^* - \bar{u}^*)^2, \tag{2}$$

where $u_k^* = w_k y_k^*$ ($k \in A$) and $\bar{u}^* = \sum_{k\in A} u_k^*/n = \hat{t}_{yI}/n$, which amounts to the with-replacement *pps sampling* variance formula and, hence, may be computed more easily by secondary users using standard software. For example, when $w_k = N/n$ then $\hat{t}_{yI} = N\bar{y}_I$ and $Var(\hat{t}_{yI}) = N^2 s_{yI}^2/n$, where $\bar{y}_I$ and s_{yI}^2 are the sample mean and variance of the imputed variable.

Clearly, naive application of estimators (1) and (2) would lead to incorrect inference generally. In order for these estimators to yield valid estimates, the imputed values need to be created in a controlled manner, as it will be discussed in the next section.

2.2 | The calibrated imputation approach

The main goal of the following calibration method for the imputed data is to provide imputed values y_k^* so that the complete-sample estimators (1) and (2) satisfy

$$\hat{t}_{yI} \equiv \sum_{k\in A} w_k y_k^* = \hat{t}_{y0}, \quad \hat{v}_F(\hat{t}_{yI}) \equiv \frac{n}{n-1}\sum_{k\in A}\left(w_k y_k^* - \hat{t}_{yI}/n\right)^2 = \hat{v}_{y0}, \tag{3}$$

where $\hat{t}_{y0}$ and $\hat{v}_{y0}$ are valid *target* estimates for the population total and its corresponding variance estimate. The method requires the data producer to choose and calculate such targets for the variable specified, as well as to calibrate the imputed values to attain the conditions in (3). These targets should incorporate all the aspects of the sampling design, response mechanism and inference approach for the imputed estimator. However, as a benefit of the calibration method, the suitability of these target estimates is a matter of concern only for the data producer and not for the secondary users, who are no longer exposed to the theoretical and computational complications involved.

The calibration method can be described by the following two-step algorithm.

Calibration algorithm:

Step 1 (Imputation): Using a standard imputation procedure, obtain a set of initial imputed values $\{\tilde{y}_k : k \in A_m\}$. For each $\tilde{y}_k$, obtain a corresponding adjusted imputed value $\hat{y}_k$ so that

$$\sum_{k\in A_m} w_k \hat{y}_k = \hat{t}_{y0} - \sum_{k\in A_r} w_k y_k. \tag{4}$$

Step 2 (Calibration): For each $k \in A_r$, set $u_k^* = w_k y_k$. For $k \in A_m$, obtain an imputed value u_k^* by a minimal adjustment to $\hat{u}_k = w_k \hat{y}_k$, where $\hat{y}_k$ is computed in Step 1, so that

$$\sum_{k \in A_m} u_k^* = \hat{t}_{y0} - \sum_{k \in A_r} w_k y_k \tag{5}$$

and

$$\sum_{k \in A_m} u_k^{*2} = \frac{n-1}{n} \hat{v}_{y0} + \frac{1}{n} \hat{t}_{y0}^2 - \sum_{k \in A_r} w_k^2 y_k^2, \tag{6}$$

where $\hat{t}_{y0}$ and $\hat{v}_{y0}$ are the targets in (3). Take $y_k^* = u_k^* / w_k$ for $k \in A$.

The algorithm initiates in Step 1 by choosing an imputation scheme to provide preliminary imputed values $\hat{y}_k$, for $k \in A_m$, such that applying (1) with these values yields $\hat{t}_{y0}$. Provided the initial imputed values $\tilde{y}_k$ $(k \in A_m)$ already yields $\hat{t}_{y0}$ by (1), one can simply take $\hat{y}_k = \tilde{y}_k$, for $k \in A_m$. An example is given in Section 3.1. Otherwise, the $\tilde{y}_k$ values need to be adjusted. One simple ratio adjustment of the initial imputed values is

$$\hat{y}_k = \frac{\left(\hat{t}_{y0} - \sum_{\ell \in A_r} w_\ell y_\ell\right)}{\sum_{\ell \in A_m} w_\ell \tilde{y}_\ell} \tilde{y}_k \quad (k \in A_m), \tag{7}$$

which is a special case of the *reverse calibration* approach of Chambers & Ren (2004), originally proposed for the estimation of t_y in the presence of survey outliers. Then, in Step 2, the calibration of the imputed values is made. Optimal imputed values that are calibrated to (5) and (6) could be computed in closed-form by applying Theorem 1 below. The proof of this theorem is shown in the Appendix.

Theorem 1. Consider initial values $\hat{a}_k$ and $d_k > 0$ for all k in a non-null set $D \subset A$. Suppose $\sum_{k \in D} d_k \hat{a}_k = t_1$ for some fixed constant t_1 and $\sum_{k \in D} d_k (\hat{a}_k - t_1/t_0)^2 > 0$, where $t_0 = \sum_{k \in D} d_k > 0$. Let $t_2 > t_1^2/t_0$ be a fixed constant . Then, the adjusted a_k values that minimize $\Delta = \sum_{k \in D} d_k (a_k - \hat{a}_k)^2$ subjected to the constraints

$$\sum_{k \in D} d_k a_k = t_1, \quad \sum_{k \in D} d_k a_k^2 = t_2, \tag{8}$$

are given by

$$a_k = t_1/t_0 + \beta(\hat{a}_k - t_1/t_0), \tag{9}$$

where

$$\beta = \left(\frac{t_2 - t_1^2/t_0}{\hat{t}_2 - t_1^2/t_0} \right)^{1/2}$$

and $\hat{t}_2 = \sum_{k \in D} d_k \hat{a}_k^2$.

The optimal calibrated imputed values y_k^* of Step 2 are obtained as follows. First, we take the values of the calibration conditions t_1 and t_2 of (8) as the right-hand sides of (5) and (6), namely

$$t_1 = \hat{t}_{y0} - \sum_{k \in A_r} w_k y_k \tag{10}$$

and

$$t_2 = (n-1)\hat{v}_{y0}/n + \hat{t}_{y0}^2/n - \sum_{k \in A_r} w_k^2 y_k^2.$$

Then, we set $D = A_m$, $d_k = 1$ and $\hat{a}_k = \hat{u}_k = w_k \hat{y}_k$ for all $k \in A_m$, where the $\hat{y}_k$ $(k \in A_m)$ are obtained in Step 1. Thus, it follows from (9) that the u_k^* values of Step 2 are

$$u_k^* \equiv a_k = t_1/m + \hat{\beta}(\hat{u}_k - t_1/m) \ \ (k \in A_m), \tag{11}$$

where t_1 is defined in (10) and

$$\hat{\beta} = \left\{ \frac{(n-1)\hat{v}_{y0}/n - \sum_{k \in A_r}(u_k - \hat{t}_{y0}/n)^2 - m\left(\hat{t}_{y0}/n - t_1/m\right)^2}{\sum_{k \in A_m}(\hat{u}_k - t_1/m)^2} \right\}^{1/2}.$$

The resulting calibrated imputed variable is

$$y_k^* = \begin{cases} y_k, & k \in A_r, \\ u_k^*/w_k, & k \in A_m. \end{cases} \tag{12}$$

Remark 1. The calibrated imputation method in (12) does not modify the observed values for units in the respondent set (A_r). The values that are actually modified are the calibrated $y_k^* = u_k^*/w_k$ values ($k \in A_m$), where the u_k^* values minimize the squared distance to the imputed values $\hat{u}_k = w_k \hat{y}_k$ ($k \in A_m$) obtained in Step 1, that is, $\Delta = \sum_{k \in A_m}(u_k^* - \hat{u}_k)^2$. The resulting u_k^* values are obtained analytically as the "best" linear predictor of u_k based on the $\hat{u}_k$ ($k \in A_m$), where the slope $\hat{\beta}$ of the regression line, given in (11), dictates how the empirical variance of the u_k^* relates to that of the $\hat{u}_k$ ($k \in A_m$). In practice, unless the $\hat{y}_k$ values are created to have greater empirical variance over A_m than A_r, one may expect $\hat{\beta} > 1$. This is because the formula (2) is ostensibly aimed at a variance of the order n^{-1}, whereas the target $\hat{v}_{y0}$ is generally aimed at a variance of the order r^{-1}, where r is the size of A_r. Thus, in order for the two to be equal to each other, the imputed y_k^* values will need to have greater variation over A_m than the observed y_k over A_r.

Remark 2. Given the set of missing units A_m, the application of Theorem 1 to obtain the optimal solution (11) requires that

$$\hat{v}_{y0} > \frac{n}{n-1}\left\{ \sum_{k \in A_r}(u_k - \hat{t}_{y0}/n)^2 + m\left(t_1/m - \hat{t}_{y0}/n\right)^2 \right\} \tag{13}$$

and

$$\sum_{k \in A_m} (\hat{u}_k - t_1/m)^2 > 0. \tag{14}$$

Comparing (13) to (2), it is readily seen that, for the solution to the optimization problem in Step 2 to exist, the target estimate $\hat{v}_{y0}$ needs to be larger than the full-sample variance estimate (2) that would have been obtained had the missing values been imputed by the common value t_1/m. The second condition (14) demands that the sampling weights and the imputation scheme are such that the $\hat{u}_k = w_k \hat{y}_k$ values are different from t_1/m for at least one $k \in A_m$. This is not the case when mean imputation is used at Step 1 to fill in the missing values of an equal probability sample. In such a situation, the proposed approach could still be applied by adding some initial zero-mean noise to each mean imputed value. The calibration constraints ensure that this added variability will not affect the variance of the imputed estimator.

3 | SOME APPLICATIONS

We explain below how the two-step approach and Theorem 1 proposed in Section 2 can be applied in some general situations, which comprise reweighting and imputation-based estimation, as well as domain estimation and estimation under stratified multistage sampling.

3.1 | Ratio imputation

Suppose that, in addition to the survey variable y, there is an auxiliary variable x which is not affected by nonresponse. Assume a population ratio model ξ of the pairs $\{(x_k, y_k) : k \in U\}$, under which

$$E_\xi(y_k \mid x_k) = \beta_0 x_k, \quad Var_\xi(y_k \mid x_k) = \sigma^2 x_k,$$

for some unknown parameters β_0 and σ^2. By ratio imputation under the model ξ, the missing y_k values are imputed as

$$\tilde{y}_k = \hat{\beta}_{0r} x_k \quad (k \in A_m),$$

where $\hat{\beta}_{0r} = \sum_{k\in A_r} w_k y_k / \sum_{k\in A_r} w_k x_k$, and $w_k = 1/\pi_k$, and π_k is the sample inclusion probability, for $k \in A$. The resulting imputed estimator of the population total t_y is

$$\hat{t}_{y0} = \sum_{k\in A_r} w_k y_k + \sum_{k\in A_m} w_k \tilde{y}_k = \hat{\beta}_{0r}\hat{t}_x,$$

where $\hat{t}_x = \sum_{k\in A} w_k x_k$ is the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) of the population total $t_x = \sum_{k\in U} x_k$. Mean imputation is a special case of ratio imputation with $x_k = 1$ for all $k \in A$, by which the imputed estimator $\hat{t}_{y0}$ reduces to $\hat{t}_{y0} = N\bar{y}_r$. Under the conditions of Theorem 1 of Kim & Rao (2009), a design and model consistent estimator of the variance of $\hat{t}_{y0}$ can be expressed as

$$\hat{v}_{y0} = \hat{v}_1 + \hat{v}_2, \tag{15}$$

where

$$\hat{v}_1 = \sum_{k\in A}\sum_{\ell\in A} \frac{(\pi_{k\ell} - \pi_k\pi_\ell)}{\pi_{k\ell}} w_k\hat{\eta}_k w_\ell\hat{\eta}_\ell, \quad \hat{v}_2 = \left(\frac{\hat{t}_x}{\hat{t}_{xr}}\right)^2 \sum_{k\in A_r} w_k \hat{e}_k^2,$$
$$\hat{\eta}_k = \hat{\beta}_{0r}x_k + \frac{\hat{t}_x}{\hat{t}_{xr}} r_k \hat{e}_k \quad (k \in A)$$

$\hat{t}_{xr} = \sum_{k\in A_r} w_k x_k$ and $\hat{e}_k = y_k - \hat{\beta}_{0r}x_k$.

However, to compute $\hat{v}_1$, the secondary user needs to have access to the matrix of the second-order inclusion probabilities $\{\pi_{k\ell} : k\ell \in A\}$, which are almost never disseminated together with the imputed sample. The proposed approach avoids this complication. To calibrate the ratio imputed values $\tilde{y}_k = \hat{\beta}_{0r}x_k$ $(k \in A_m)$, we notice that $\hat{y}_k = \tilde{y}_k$ already satisfies Step 1, since $t_1 = \hat{\beta}_{0r}\hat{t}_{xm}$ in Theorem 1. For Step 2, by (12) and (11), the calibrated imputed values are

$$y_k^* = u_k^*/w_k, \tag{16}$$

where $u_k^* = w_k y_k$ $(k \in A_r)$, and for $k \in A_m$,

$$u_k^* = \frac{\hat{\beta}_{0r}\hat{t}_{xm}}{m} + \hat{\beta}\hat{\beta}_{0r}\left(w_k x_k - \frac{\hat{t}_{xm}}{m}\right) \quad (k \in A_m)$$

and

$$\hat{\beta}^2 = \frac{\frac{(n-1)}{n}(\hat{v}_1 + \hat{v}_2) - \sum\limits_{k\in A_r}\left(w_k y_k - \frac{\hat{\beta}_{0r}\hat{t}_x}{n}\right)^2 - m\hat{\beta}_{0r}^2\left(\frac{\hat{t}_x}{n} - \frac{\hat{t}_{xm}}{m}\right)^2}{\hat{\beta}_{0r}^2 \sum\limits_{k\in A_m}\left(w_k x_k - \frac{\hat{t}_{xm}}{m}\right)^2}.$$

In the case of mean imputation and simple random sampling without replacement, (15) reduces to

$$\hat{v}_{y0} = N^2\left(\frac{1}{r} - \frac{1}{N}\right)s_{yr}^2, \tag{17}$$

where $s_{yr}^2 = \sum_{k\in A_r}(y_k - \bar{y}_r)^2/(r-1)$ and $\bar{y}_r$ is the observed respondent mean.

3.2 | Domain estimation

As a realistic setting for domain total estimation, in addition to the population total, consider a domain population partition $U = U_1 \cup \cdots \cup U_D$. Let the population total of domain U_d be

$$t_{dy} = \sum_{k\in U_d} y_k = \sum_{k\in U} \delta_{kd} y_k,$$

where the domain indicator δ_{kd}, $\delta_{kd} = 1$ if $k \in U_d$ and $\delta_{kd} = 0$ otherwise, is observed for all units in the sample A ($d = 1, \dots, D$). Let $\hat{t}_{dy}$ be the target domain total estimator and $\hat{v}_{dy}$ its variance estimate. Domain estimation can be handled by separate calibration for each domain by the producer and application of the domain complete-data formulae by the secondary users, yielding $\hat{t}_{dyI} = \hat{t}_{dy}$ and $\hat{v}_F(\hat{t}_{dyI}) = \hat{v}_{dy}$, as explained in Section 2.

However, one is still interested in estimating the population total, in addition to the domain totals. Directly applying the complete-sample formula (1) to the domain-calibrated imputed sample would correctly estimate the population total. One can combine the domain variance estimates, as if the sampling were stratified by the domains. However, the resulting variance estimate is incorrect even when the domain total estimators are independent of each other, due to the additional term

$$v_b = \frac{n}{n-1}\sum_{d=1}^{D} n_d(\hat{t}_{dyI}/n_d - \hat{t}_{y0}/n)^2 = \frac{n^2}{n-1}V_n(\hat{t}_{dyI}/n_d),$$

where $V_n(\hat{t}_{dyI}/n_d)$ is the variance of $\hat{t}_{dyI}/n_d$ with respect to the empirical sample domain distribution function $(n_1/n, \dots, n_D/n)$, since

$$V_n(\hat{t}_{dyI}/n_d) = \sum_{d=1}^{D}\frac{n_d}{n}(\hat{t}_{dyI}/n_d - \hat{t}_{y0}/n)^2 \quad \text{and} \quad \hat{t}_{y0}/n = E_n(\hat{t}_{dyI}/n_d) = \sum_{d=1}^{D}\frac{n_d}{n}(\hat{t}_{dyI}/n_d).$$

We propose to introduce a *domain estimation effect factor*, denoted by γ, and use

$$\hat{v}_F(\hat{t}_{yI}) = \gamma^2\frac{n}{n-1}\sum_{k\in A}(w_k y_k^* - \hat{t}_{y0}/n)^2 = \hat{v}_{y0}. \tag{18}$$

The factor γ can be calculated after domain-calibrated imputation, and disseminated together with imputed sample.

In the separate domain calibration above, $\hat{v}_F(\hat{t}_{dyI})$ is built on the squared errors around $\hat{t}_{dyI}/n_d$. Consider using another complete-sample formula $\hat{v}_F(\hat{t}_{dyI})$, built around $\hat{t}_{yI}/n$ instead, where

$$\hat{v}_F(\hat{t}_{dyI}) = \frac{n_d}{n_d - 1}\sum_{k\in A}\delta_{kd}(w_k y_{kd}^* - \hat{t}_{y0}/n)^2.$$

We need to extend the calibration constraints as follows:

$$\begin{cases} \hat{t}_{dyI} = \sum_{k\in A} \delta_{kd} w_k y_k^* = \hat{t}_{dy} & \text{for } d = 1, ..., D \\ \hat{v}_F(\hat{t}_{dyI}) = \frac{n_d}{n_d - 1} \sum_{k\in A} \delta_{kd} (w_k y_{kd}^* - \hat{t}_{y0}/n)^2 = \hat{v}_{dy} & \text{for } d = 1, ..., D \\ \hat{v}_F(\hat{t}_{yI}) = \gamma^2 \frac{n}{n-1} \sum_{k\in A} (w_k y_k^* - \hat{t}_{y0}/n)^2 = \hat{v}_{y0}. \end{cases} \tag{19}$$

In other words, we use δ_{kd} to identify the relevant observations for domain estimation, including the special case of $U_d = U$ and $\delta_{kd} \equiv 1$, and use $\hat{t}_{y0}/n$ in all the ultimate variance estimators, including domain variance estimation. We refer to (19) as the centred domain calibration approach.

Minimum adjustments of $\{\hat{y}_k; k \in A_m\}$ from Step 2 of the proposed approach can be achieved by Theorem 1 as well. To focus the idea, suppose negligible $1/n$ and $1/n_d$. Let $\{u_k^*; k \in A_{md}\}$ be the calibrated imputations in domain d, given by

$$u_k^* = t_{1d}/m + \beta_d(\hat{u}_k - t_{1d}/m),$$

where $t_{1d} = \hat{t}_{dy} - \sum_{k\in A_{rd}} w_k y_k$ is the constrained total of $u_k^* = w_k y_k^*$ in A_{md}. However, instead of choosing β_d such that

$$\beta_d^2 \sum_{k\in A_{md}} \left(\hat{u}_k - \frac{t_{1d}}{m}\right)^2 = \hat{v}_{dy} - \sum_{k\in A_{rd}} \left(u_k - \frac{\hat{t}_{dy}}{n_d}\right)^2 - m_d\left(\frac{\hat{t}_{dy}}{n_d} - \frac{t_{1d}}{m}\right)^2,$$

as under separate domain calibration, we should now choose β_d such that

$$\beta_d^2 \sum_{k\in A_{md}} \left(\hat{u}_k - \frac{t_{1d}}{m}\right)^2 = \hat{v}_{dy} - \sum_{k\in A_{rd}} \left(u_k - \frac{\hat{t}_{y0}}{n}\right)^2 - m_d\left(\frac{\hat{t}_{y0}}{n} - \frac{t_{1d}}{m}\right)^2.$$

This allows us to estimate the domain variance $\hat{v}_{dy}$ as in (19). The domain estimation effect factor γ can be calculated afterwards to satisfy (19). The conditions for the existence of solution are formally the same as discussed in Section 2.2. Provided domain-specific calibration, it is feasible as long as $\hat{t}_{y0}/n$ does not differ too much from $\hat{t}_{dy}/n_d$ in the different domains.

In practice one may be interested in multiple sets of (overlapping) domains. For example, a user may want to have estimates by region as well as estimates by industry. Insofar as the need is known in advance, the producer can apply the approach above to the ‘atomic domains’, which arise from crossing region and industry. In addition to the separate atomic-domain calibrated sample, one can supply a domain estimation factor for the population total, a set of domain estimation factors for each of the regions, and another set of factors for each industry.

3.3 | Stratified Multistage Sampling

Let the population U be partitioned into H strata of n primary sampling units (PSUs), where a sample A_h of n_h PSUs is selected separately within the hth stratum ($h = 1, \dots, H$; $n_1 + \cdots + n_H = n$). From each PSU in A_h, additional stages of sampling are undertaken until the selection of the ultimate sampling units (USUs). Let w_i, y_i and r_i be, respectively, the weight, the y-value and the response indicator for the ith USU. Let A_{hk} be the set of USUs in the kth selected PSU of the hth stratum, where $A_{rhk} = \{i : i \in A_{hk}, r_i = 1\}$ and $A_{mhk} = \{i : i \in A_{hk}, r_i = 0\}$.

By setting $y_i^* = y_i$ if $r_i = 1$ and letting y_i^* be the calibrated imputation value if $r_i = 0$, the imputed estimate of the population total t_y can be written as $\hat{t}_{yI} = \sum_{h=1}^{H} \hat{t}_{yIh}$, where $\hat{t}_{yIh} = \sum_{k \in A_h} u_{hk}^*$ and $u_{hk}^* = \sum_{i \in A_{hk}} w_i y_i^* = \sum_{i \in A_{rhk}} w_i y_i + \sum_{i \in A_{mhk}} w_i y_i^*$. For calibrated imputation that enables (3), we can apply Theorem 1 and the 2-step approach directly at the level of USUs, ignoring the clustering structure of the multistage sampling.

Survey data analysis softwares (such as STATA, R, SAS) commonly use the stratified ultimate variance formula for variance estimation. It is therefore convenient if the secondary user can simply input the imputed sample, and let the software carry on as usual. Thus, as another possibility of full-sample variance estimator, we consider

$$\hat{v}_F(\hat{t}_{yI}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k \in A_h} (u_{hk}^* - \bar{u}_h^*)^2,$$

where $\bar{u}_h^* = \sum_{k \in A_h} u_{hk}^* / n_h = \hat{t}_{yIh} / n_h$. This choice fits naturally with the standard approach of ultimate-cluster variance estimation under stratified multistage sampling (e.g. Skinner, 1989, Section 2.13).

Given $\hat{t}_{y0h}$ for the population total in the hth stratum and its associated variances $\hat{v}_{y0h} = \hat{v}(\hat{t}_{y0h})$ ($h = 1, \dots, H$), consider the problem of finding the values y_j^* starting with $\tilde{y}_j$, $j \in \cup_{k \in A_h} A_{mhk}$, so that

$$\begin{aligned} &\hat{t}_{yIh} = \textstyle\sum_{k \in A_h} u_{hk}^* = \hat{t}_{y0h}, \\ &\textstyle\sum_{k \in A_h} (u_{hk}^* - \bar{u}_h^*)^2 = \sum_{k \in A_h} u_{hk}^{*2} - \hat{t}_{y0h}^2 / n_h = (n_h - 1)\hat{v}_{y0h} / n_h. \end{aligned} \tag{20}$$

We propose to obtain a solution of this problem in two stages. First, the initial imputed PSU totals are adjusted minimally subject to the two constraints above, yielding the adjusted PSU total u_{hk}^*. Second, the initial imputed values $\tilde{y}_j$ are adjusted, separately within each PSU, to agree with the corresponding calibrated PSU total from the first step.

For the first stage, we can apply Theorem 1 within the hth stratum similarly as in Section 2. Let $A_{h0} = \{k \in A_h : \#(A_{mhk}) = 0\}$, $\hat{u}_{hk} = u_{hk}$ for $k \in A_{h0}$ and $\hat{u}_{hk} = \tilde{u}_{hk}(\hat{t}_{y0h} - \sum_{k \in A_{h0}} u_{hk}) / (\sum_{\ell \in A_h \setminus A_{h0}} \tilde{u}_{h\ell})$ for $k \in A_h \setminus A_{h0}$. Then, take $D = D_h = A_h \setminus A_{h0}$,

$d_k = d_{hk} = 1$, $\hat{a}_k = \hat{a}_{hk} = \hat{u}_{hk}$, $t_0 = \sum_{k\in A_h\setminus A_{h0}} d_k \equiv m_h$, $t_1 = t_{1h} = \hat{t}_{y0h} - \sum_{k\in A_{h0}} u_{hk}$ and $t_2 = t_{2h} = (n_h - 1)\hat{t}_{y0h}/n_h + \hat{t}^2_{y0h}/n_h - \sum_{k\in A_{h0}} u^2_{hk}$. For each $h = 1, \dots, H$, the optimal solution that minimizes the squared distance $\Delta_h = \sum_{A_h}(u^*_{hk} - \hat{u}_{hk})^2 = \sum_{A_h\setminus A_{h0}}(u^*_{hk} - \hat{u}_{hk})^2$ subject to (20) are given by $u^*_{hk} = u_{hk}$ for $k \in A_{h0}$, and

$$u^*_{hk} = \frac{\hat{t}_{1h}}{m_h} + \hat{\beta}_h\Big(\hat{u}_{hk} - \frac{\hat{t}_{1h}}{m_h}\Big) \tag{21}$$

for $k \in A_h\setminus A_{h0}$, where

$$\hat{\beta}_h = \left\{\frac{(n_h - 1)\hat{v}_{y0h}/n_h - \sum_{A_{h0}}(u_{hk} - \hat{t}_{y0h}/n_h)^2 - m_h(\hat{t}_{y0h}/n_h - t_{1h}/m_h)^2}{\sum_{k\in A_h\setminus A_{h0}}(\hat{u}_{hk} - t_{1h}/m_h)^2}\right\}^{\frac{1}{2}}.$$

Having thus obtained u^*_{hk}, we adjust the $\tilde{y}_i$'s separately within each PSU so that $u^*_{hk} = \sum_{i\in A_{rhk}} w_i y_i + \sum_{i\in A_{mhk}} w_i y^*_i$, which is a single constraint. For given h and $k \in A_h\setminus A_{h0}$, the values y^*_i that minimize the distance $\sum_{i\in A_{mhk}}(y^*_i - \tilde{y}_i)^2/2$ subject to $\sum_{j\in A_{mhk}} w_i y^*_i = u^*_{hk} - \sum_{i\in A_{rhk}} w_i y_i \equiv u_{hk0}$ are

$$y^*_i = \tilde{y}_i\Big\{1 + \Big(\frac{w_i}{\tilde{y}_i}\Big)\frac{\big(u_{hk0} - \sum_{i\in A_{mhk}} w_i\tilde{y}_i\big)}{\sum_{i\in A_{mhk}} w^2_i}\Big\} \quad (i \in A_{mhk}). \tag{22}$$

4 | CALIBRATION OF MULTIPLE VARIABLES

Let $\boldsymbol{y}_k = (y_{k1}, \dots, y_{kp})^\top$ denote a p-dimensional vector of values for the k-th unit and $\boldsymbol{u}^*_k = w_k\boldsymbol{y}^*_k$, where $\boldsymbol{y}^*_k = (y^*_{k1}, \dots, y^*_{kp})^\top$ denote the calibrated imputed values having the restriction that $y^*_{k\ell} = y_{k\ell}$ if $y_{k\ell}$ is observed and fixed ($k \in A$ and $\ell = 1, \dots, p$). Following the basic algorithm of Section 2, consider the problem of finding the $\boldsymbol{u}^*_k$ satisfying

$$\sum_{k\in A}\boldsymbol{u}^*_k = \hat{\boldsymbol{t}}_0, \quad \frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2} = \hat{\boldsymbol{V}}_0,$$

where $\hat{\boldsymbol{t}}_0$ denotes a p-dimensional vector of target estimates for the population total $\boldsymbol{t}_y = \sum_{k\in U}\boldsymbol{y}_k$, $\hat{\boldsymbol{V}}_0$ denotes the target estimated variance-covariance matrix of $\hat{\boldsymbol{t}}_0$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\top$. To obtain $\hat{\boldsymbol{V}}_0$ in the presence of multivariate missing data is a difficult issue. See, e.g., Skinner & Rao (2002) and Chauvet & Haziza (2012) for a fully efficient approach in the bivariate case, and Im et al. (2018) and Sang & Kim (2018) for two fractional imputation methods in the multivariate setting. Below we propose a two-phase calibration procedure, where at the first phase the problem is solved for transformed vectors $\boldsymbol{v}^*_k$'s ($k \in A$), and at the second phase the results are back-transformed to $\boldsymbol{u}^*_k$'s as required. Assume without loss of generality the weighted values

$$\hat{\boldsymbol{u}}_k = w_k\hat{\boldsymbol{y}}_k \equiv (\hat{u}_{k1}, \dots, \hat{u}_{kp})^\top$$

satisfy $\sum_{k\in A}\hat{\boldsymbol{u}}_k = \hat{\boldsymbol{t}}_0$ preserving all the observed and fixed values. This can be achieved for instance by separate imputation of $y_{k\ell}$, for $\ell = 1, ..., p$, up to Step 1 in Section 2.2.

Let D be the subset of units with all items missing and suppose $m = |D| > 0$. Denote by $\hat{\boldsymbol{V}}_0 = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$ the spectral decomposition of $\hat{\boldsymbol{V}}_0$, where $\boldsymbol{P}$ is the $p \times p$ orthogonal matrix of the eigenvectors of $\hat{\boldsymbol{V}}_0$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of the corresponding eigenvalues. Consider the orthogonal principal components

$$\hat{\boldsymbol{v}}_k = \boldsymbol{P}^\top(\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{t}}_0/n) \quad (k \in A)$$

and set

$$\boldsymbol{v}^*_k = \hat{\boldsymbol{v}}_k \quad (k \in R = A \setminus D). \tag{23}$$

Now, choose the $\{\boldsymbol{v}^*_k : k \in D\}$ which minimize the squared Frobenius norm

$$\sum_{k\in D}(\boldsymbol{v}^*_k - \hat{\boldsymbol{v}}_k)^\top(\boldsymbol{v}^*_k - \hat{\boldsymbol{v}}_k)$$

subject to

$$\sum_{k\in D}\boldsymbol{v}^*_k = -\sum_{k\in R}\hat{\boldsymbol{v}}_k \equiv m\bar{\boldsymbol{v}}^*_D, \quad \sum_{k\in D}\boldsymbol{v}^{*\otimes 2}_k = \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k.$$

The solution to this constrained optimization problem can be obtained by Theorem 2 below. The proof of this theorem is presented in the Appendix.

Theorem 2. Consider a set of vectors $\hat{\boldsymbol{a}}_k$ $(k \in D)$, satisfying $\sum_{k\in D}\hat{\boldsymbol{a}}_k = 0$ and $\boldsymbol{C} = \sum_{k\in D}\hat{\boldsymbol{a}}^{\otimes 2}_k$ being positive definite. Let $\boldsymbol{B}$ be a pre-specified positive definite matrix. Multivariate calibrated vectors that minimize $\sum_{k\in D}(\boldsymbol{a}^*_k - \hat{\boldsymbol{a}}_k)^\top(\boldsymbol{a}^*_k - \hat{\boldsymbol{a}}_k)$ subjected to

$$\sum_{k\in D}\boldsymbol{a}^*_k = 0, \quad \sum_{k\in D}\boldsymbol{a}^*_k\boldsymbol{a}^{*\top}_k = \boldsymbol{B}$$

are $\boldsymbol{a}^*_k = \boldsymbol{\beta}\hat{\boldsymbol{a}}_k$, where $\boldsymbol{\beta} = \boldsymbol{B}^{1/2}\boldsymbol{C}^{-1/2}$ and $\boldsymbol{B}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are the square root and inverse square root matrices of $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively.

In the first phase of the calibration method, we let $\boldsymbol{a}^*_k = \boldsymbol{v}^*_k - \bar{\boldsymbol{v}}^*_D$ and $\hat{\boldsymbol{a}}_k = \hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D$ in Theorem 2. If the matrices

$$\boldsymbol{B} \equiv \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k - m\bar{\boldsymbol{v}}^{*\otimes 2}_D, \quad \boldsymbol{C} \equiv \sum_{k\in D}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D)^{\otimes 2}$$

are both positive definite, then the optimal calibrated $\boldsymbol{v}^*_k$ vectors by Theorem 2 are

$$\boldsymbol{v}^*_k = (\boldsymbol{I}_p - \boldsymbol{\beta})\bar{v}^*_D + \boldsymbol{\beta}\hat{\boldsymbol{v}}_k = \bar{\boldsymbol{v}}^*_D + \boldsymbol{\beta}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D) \quad (k \in D), \tag{24}$$

where $\boldsymbol{I}_p$ is the identity matrix of order p. Now, from $\sum_{k\in D}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D) = \sum_{k\in A}\hat{\boldsymbol{v}}_k = 0$ by (23), it follows that $\sum_{k\in D}\boldsymbol{v}^*_k = m\bar{\boldsymbol{v}}^*_D$ and $\sum_{k\in D}\boldsymbol{v}^{*\otimes 2}_k = m\bar{\boldsymbol{v}}^{*\otimes 2}_D + \boldsymbol{\beta}\boldsymbol{C}\boldsymbol{\beta}^\top = m\bar{\boldsymbol{v}}^{*\otimes 2}_D + \boldsymbol{B} = (n-1)n^{-1}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k$. We thus enter the second phase of imputation, where we transform the $\boldsymbol{v}^*_k$ $(k \in D)$ given by (23) and (24) back into

$$\boldsymbol{u}^*_k = \boldsymbol{P}\boldsymbol{v}^*_k + \hat{\boldsymbol{t}}_0/n = \begin{cases} \hat{\boldsymbol{u}}_k, & (k \in R = A \setminus D), \\ \boldsymbol{P}\{\bar{\boldsymbol{v}}^*_D + \boldsymbol{\beta}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D)\} + \hat{\boldsymbol{t}}_0/n, & (k \in D), \end{cases} \tag{25}$$

which are the final calibrated imputed vectors. The transformation (25) implies that

$$\sum_{k\in A}\boldsymbol{u}^*_k = \boldsymbol{P}\sum_{k\in A}\boldsymbol{v}^*_k + \hat{\boldsymbol{t}}_0 = \hat{\boldsymbol{t}}_0$$

and

$$\frac{n}{n-1}\sum_{i\in D}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2} = \frac{n}{n-1}\boldsymbol{P}\Big\{\sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k + \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k\Big\}\boldsymbol{P}^\top = \hat{\boldsymbol{V}}_0,$$

as intended.

Notice that the vectors $\boldsymbol{u}^*_k$ in (25) can be computed directly from $\hat{\boldsymbol{u}}_k$ by the linear transformation

$$\boldsymbol{u}^*_k = \tilde{\boldsymbol{\alpha}}\mathbb{1}(k \in D) + \{\boldsymbol{I}_p\mathbb{1}(k \in A \setminus D) + \tilde{\boldsymbol{\beta}}\mathbb{1}(k \in D)\}\hat{\boldsymbol{u}}_k, \tag{26}$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{\beta}\boldsymbol{P}^\top$, $\tilde{\boldsymbol{\alpha}} = -(\boldsymbol{I}_p - \tilde{\boldsymbol{\beta}})(\sum_{k\in R}\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{t}}_0)/m$ and $\mathbb{1}(\mathcal{A})$ is the indicator function of a set $\mathcal{A}$. All the observed and fixed data values are associated with the units in $R = A \setminus D$, and thus are preserved in the respective $\boldsymbol{u}^*_k$ values. The $\boldsymbol{u}^*_k$ and $\hat{\boldsymbol{u}}_k$ vectors share the same distance as that of the corresponding $\boldsymbol{v}^*_k$ and the $\hat{\boldsymbol{v}}_k$ vectors, for $k \in D$.

One of the benefits of the proposed calibrated imputation approach above is to facilitate inferences for a nonlinear function of population total vector $\boldsymbol{t}_y$. For example, suppose $\sqrt{n}(\hat{\boldsymbol{t}}_0 - \boldsymbol{t}_y)$ has an asymptotic multivariate normal distribution with mean zero and positive-definite variance-covariance matrix $\boldsymbol{V}$ and that the target variance $\hat{\boldsymbol{V}}_0$ is consistent for the variance of $\hat{\boldsymbol{t}}_y$. Thus, an approximate $100(1-\alpha)\%$ confidence interval for $g(\boldsymbol{t}_y)$, where $g : \mathbb{R}^p \longrightarrow \mathbb{R}$ is smooth and has nonzero gradient $\nabla g(\cdot)$ at $\boldsymbol{t}_y$, is

$$g\left(\hat{\boldsymbol{t}}_{yI}\right) \pm z_{1-\alpha/2}\left[\nabla^\top g\left(\hat{\boldsymbol{t}}_{yI}\right)\left\{\frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2}\right\}\nabla g\left(\hat{\boldsymbol{t}}_{yI}\right)\right]^{1/2},$$

where $\hat{\boldsymbol{t}}_{yI} = \sum_{k\in A}\boldsymbol{u}^*_k = \hat{\boldsymbol{t}}_0$, which is Wald-type interval that could alternatively be computed by

$$g\left(\hat{\boldsymbol{t}}_{yI}\right) \pm z_{1-\alpha/2}\left[\hat{v}_F\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\}\right]^{1/2}, \tag{27}$$

where

$$\hat{v}_F\left\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\right\} = \frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_{gk} - \hat{\boldsymbol{t}}_{g0}/n)^{\otimes 2},$$

$\boldsymbol{u}^*_{gk} = \nabla^\top g\left(\hat{\boldsymbol{t}}_{yI}\right)\boldsymbol{u}^*_k$ and $\hat{\boldsymbol{t}}_{g0} = \nabla^\top g\left(\hat{\boldsymbol{t}}_{yI}\right)\hat{\boldsymbol{t}}_0$. The use of the estimate $\hat{v}_F\left\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\right\}$ above does not merely simplifies the computation of the estimated variance of $g\left(\hat{\boldsymbol{t}}_{yI}\right)$ by a complete sample formula, but it also prevents the secondary users to having to apply a specific variance estimation method and corresponding software to estimate the variance of $g\left(\hat{\boldsymbol{t}}_{yI}\right)$.

5 | CONCLUDING REMARKS

In this paper we propose a calibrated imputation approach to be used routinely in the preparation of imputed sample survey data, under a stratified multistage sampling design. It allows secondary users to estimate specified population parameters and associated variances and covariances by simple complete-sample formulae, regardless of how complicatedly it may be when these are to be derived from the original incomplete data. Domain estimation can be accommodated in addition. The approach avoids the need to disseminate multiple imputed (or replicated) datasets.

A topic for future research is the imputation of categorical variables. The fractional imputation of Kim & Fuller (2004) provides a fully efficient imputation method, where K imputed values are created for a K-category variable. Favre et al. (2005) propose a fully efficient single-sample imputation method, under which a categorical variable is treated as a vector of dummy-indicators during the imputation. However, variance estimation under these methods may require a different approach than by using simple complete-sample formulae. It seems possible to adapt the multivariate calibrated imputation approach in this paper, where the imputed values of the dummy-vector can be continuous instead of just 0 and 1, given the target estimates of the population totals of each category and the associated variances and covariances.

Another research topic of interest is calibrated imputation for quantile estimation. For instance, given the target quantile estimate $\hat{t}_\alpha$ and its variance estimate $\hat{v}_\alpha$, for $0 < \alpha < 1$. It may be possible to adapt the proposed algorithm in Section 2.2. as follows. At Step 1, adjust the initial imputed values $\{\tilde{y}_k : k \in A_m\}$ to $\{\hat{y}_k : k \in A_m\}$, which satisfy

$$\sum_{k\in A_r} w_i I(y_k \leqslant \hat{t}_\alpha) + \sum_{k\in A_m} w_i I(\hat{y}_k \leqslant \hat{t}_\alpha) = \alpha \sum_{k\in A} w_i.$$

Then, at Step 2, adjust the negative and positive deviations $\hat{y}_k - \hat{t}_\alpha$, for $k \in A_m$, while maintaining their signs to satisfy a suitably chosen complete-sample variance formula.

ACKNOWLEDGMENTS

This research was developed at the Southampton Statistical Sciences Research Institute of the University of Souhampton, UK. The research of the first author was supported by a grant from the Brazilian Research Council (CNPq). We are thankful to the Associate Editor and two anonymous referees for their useful comments.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. J. R. Stat. Soc. Ser. B Stat. Methodol., 67(3), 445–458.

Chambers, R. L., & Ren, R. (2004). Outlier robust imputation of survey data. In *JSM Proceedings*, the Survey Research Methods Section, 3336–3344, Alexandria, VA: American Statistical Association.

Chauvet, G., Deville, J.-C., & Haziza, D. (2011). On balanced random imputation in surveys. Biometrika, 98(2), 459–471.

Chauvet, G., & Haziza, D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data. Canad. J. Statist., 40(1), 124–149.

Chen, J., Rao, J. N. K., & Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. Statist. Sinica, 10(4), 1153–1169.

Deville, J.-C., & Särndal, C.-E. (1992a). Calibration estimators in survey sampling. J. Amer. Statist. Assoc., 87, 376–382.

Deville, J.-C., & Särndal, C.-E. (1994b). Variance estimation for the regression imputed horvitz-thompson estimator. Journal of Official Statistics, 10, 381–394.

Favre, A.-C., Matei, A., & Tillé, Y. (2005). Calibrated random imputation for qualitative data. J. Statist. Plann. Inference, 128(2), 411 – 425.

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, 429-440, Washington, D.C.: U.S. Dept. of Commerce, Bureau of the Census.

Fay, R. E. (1992). When are inferences from multiple imputation valid? In *JSM Proceedings*, the Survey Research Methods Section, 227-232, Alexandria, VA: American Statistical Association.

Fuller, W. A., & Kim, J. K. (2005). Hot deck imputation for the response model. Survey Methodology, 31(2), 139–149.

Gelein, B., Haziza, D., & Causeur, D. (2014). Preserving relationships between variables with MIVQUE based imputation for missing survey data. J. Multivariate Anal., 131(0), 197 – 208.

Ghosh, M. (1992). Constrained bayes estimation with applications. J. Amer. Statist. Assoc., 87(418), 533–540.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics, Volume 29A, Sample surveys: Design, methods and applications* (pp. 215-246). Amsterdam: Elsevier.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc., 47, 663–685.

Im, J., Cho, I. H., & Kim, J. K. (2018). FHDI: An R package for fractional hot deck imputation. The R Journal, 10(1), 140–154.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. In *JSM Proceedings*, the Survey Research Methods Section, 22–31, Alexandria, VA: American Statistical Association.

Kim, J. K. (2001a). Variance estimation afeter imputation. Survey Methodology, 27(1), 75–83.

Kim, J. K. (2011b). Parametric fractional imputation for missing data analysis. Biometrika, 98(1), 119–132.

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. Biometrika, 91(3), 559–578.

Kim, J. K., & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. Biometrika, 96(4), 917–932.

Lanke, J. (1983). Hot deck imputation techniques that permit standard methods for assessing precision of estimates. Statistical Review, 21(5), (Essays in Honour of Tore E. Dalenius), 105–110.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statist. Sci., 9(4), 538–558.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. J. Amer. Statist. Assoc., 91(434), 499–506.

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika, 79, 811–822.

Rubin, D. B. (1978a). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. In *JSM Proceedings*, the Survey Research Methods Section, 20-28, Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987b). Multiple imputation for nonresponse in surveys. Hoboken, NJ: Wiley.

Rubin, D. B. (1996c). Multiple imputation after 18+ years. J. Amer. Statist. Assoc., 91(434), 473–489.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc., 81(394), 366–374.

Sang, H., & Kim, J. K. (2018). Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. Unpublished manuscript. arXiv:1809.05976. Retrieved from https://arxiv.org/abs/1809.05976.
URL https://arxiv.org/abs/1809.05976

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18, 241–252.

Sedransk, J. (1985). The objectives and practice of imputation. In *Proceedings of the First Annual Research Conference*, 445–452, Washington, D.C.: Bureau of the Census.

Shao, J., & Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. J. Amer. Statist. Assoc., 91(435), 1278–1288.

Shao, J., & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. J. Amer. Statist. Assoc., 94(445), 254–265.

Skinner, C., & Rao, J. (2002). Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. J. Statist. Plann. Inference, 102(1), 149 – 167.

Skinner, C. J. (1989). Introduction to part A. In C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 23-58). Chichester: Wiley.

Wolter, K. M. (2007). Introduction to variance estimation (2nd ed.). New York: Springer.

SUPPORTING INFORMATION

The supporting information, available as part of the online article, provides numerical illustrations for the proposed approach to estimate a population total and to preserve variability. Examples consider element sampling, stratified multistage sampling and also calibration of imputed multivariate data.

□

APPENDIX

A PROOF OF THEOREM 1

The proof of Theorem 1 is shown below. In the special case of $d_k = 1$, the problem addressed by Theorem 1 reduces essentially to that of the small-area constrained Bayes estimators (Ghosh, 1992, Theorem 1), which minimize an unweighted quadratic function of the prediction errors. The two constraints in Ghosh (1992, Theorem 1), despite being formulated in a small area estimation setting, are equivalent in form to (8).

Proof. Notice that

$$\sum_{k\in D} d_k(a_k - \hat{a}_k)^2 = t_2 - 2\sum_{k\in D} d_k a_k \hat{a}_k + \hat{t}_2.$$

Since t_2 and $\hat{t}_2$ are positive and fixed, one needs to maximize $\sum_{k\in D} d_k a_k \hat{a}_k$ or, equivalently, $\sum_{k\in D} q_k a_k \hat{a}_k$, where $q_k = d_k/t_0$ and $\sum_{k\in D} q_k = 1$. Consider then $(q_1, ..., q_m)$ as a discrete distribution function, denoted by $\mathcal{F}_q$, which puts probability q_k on each pair of $(a_k, \hat{a}_k)$, for $k \in D$. We have $\sum_{k\in D} q_k a_k \hat{a}_k = E_q(a_k \hat{a}_k)$, i.e., the expectation over $\mathcal{F}_q$. Notice that $E_q(a_k) = \sum_{k\in D} q_k a_k = t_1/t_0$ and $E_q(\hat{a}_k) = t_1/t_0$ are both fixed, as well as $V_q(a_k) = t_2/t_0 - (t_1/t_0)^2$ and $V_q(\hat{a}_k) = \hat{t}_2/t_0 - (t_1/t_0)^2$. Thus, the maximum of $E_q(a_k \hat{a}_k)$ is the same as that of $Cov_q(a_k, \hat{a}_k)$, which is given when the correlation between a_k and $\hat{a}_k$ is equal to 1 over $\mathcal{F}_q$, i.e., $a_k = \alpha + \beta \hat{a}_k$ with probability one, for some constants α and β. Solving (α, β) for (8) yields then $\alpha = (1 - \beta)t_1/t_0$ and $\beta = [(t_2 - t_1^2/t_0)/(\hat{t}_2 - t_1^2/t_0)]^{1/2}$. This completes the proof. □

B PROOF OF THEOREM 2

Proof. Let the Lagrangian of this constrained optimisation problem be

$$L = \frac{1}{2}\sum_{k\in D}(\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*)^\top(\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*) - \boldsymbol{\psi}^T\sum_{k\in D}\tilde{\boldsymbol{a}}_k^* - \mathbf{1}^\top[\boldsymbol{\Lambda}\circ(\sum_{k\in D}\tilde{\boldsymbol{a}}_k^*\tilde{\boldsymbol{a}}_k^{*\top} - \boldsymbol{S}_D^*)]\mathbf{1}$$

for some vector $\boldsymbol{\psi}$ and matrix $\boldsymbol{\Lambda}$, where $\mathbf{1}$ denotes a vector of ones and "$\circ$" denotes the Hadamard (element-wise) product of two matrices. We have

$$(\partial L/\partial\tilde{\boldsymbol{a}}_k^*)^\top = (\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*) - \boldsymbol{\psi} - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)\tilde{\boldsymbol{a}}_k^* = 0 \quad\Rightarrow\quad \tilde{\boldsymbol{a}}_k^* = [\boldsymbol{I}_p - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)]^{-1}(\boldsymbol{a}_k^* + \boldsymbol{\psi})$$

provided $\boldsymbol{W} = [\boldsymbol{I}_p - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)]^{-1}$ exists. Notice that $\boldsymbol{W}$ is symmetric. Substitution of $\tilde{\boldsymbol{a}}_k^* = \boldsymbol{W}(\boldsymbol{a}_k^* + \boldsymbol{\psi})$ into $\sum_{k\in D}\boldsymbol{a}_k^* = 0$ yields $\boldsymbol{\psi} = 0$. Next, substitution of $\tilde{\boldsymbol{a}}_k^* = \boldsymbol{W}\boldsymbol{a}_k^*$ into $\sum_{k\in D}\boldsymbol{a}_k^*\boldsymbol{a}_k^{*\top} = \boldsymbol{B}$ yields

$$\boldsymbol{B}\boldsymbol{B}^\top = \sum_D \boldsymbol{W}\boldsymbol{a}_k^*\boldsymbol{a}_k^{*\top}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{W} \quad\Rightarrow\quad \boldsymbol{W} = \boldsymbol{B}\boldsymbol{C}^{-1}.$$

This completes the proof. □

Received ; | Revised ; | Accepted

DOI: xxx/xxxx

ARTICLE TYPE

A Calibrated Imputation Method for Secondary Data Analysis of Survey Data.

Damião Nóbrega Da Silva*[1] | Li-Chun Zhang[2,3,4]

[1]Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil
[2]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, Hampshire, UK
[3]Statistisk sentralbyrå, Oslo, Norway
[4]University of Oslo, Oslo, Norway

Correspondence
*Damião Nóbrega Da Silva, Departamento de Estatística, Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte, Natal, RN, 59078-970 Brazil. Email: damiao@ccet.ufrn.br

Funding Information

This research was supported by the Brazilian Research Council (CNPq), Grant Number: 211518/2013-1

Summary

In practical survey sampling, missing data are unavoidable due to nonresponse, rejected observations by editing, disclosure control or outlier suppression. We propose a calibrated imputation approach so that valid point and variance estimates of the population (or domain) totals can be computed by the secondary users using simple complete-sample formulae. This is especially helpful for variance estimation, which generally require additional information and tools that are unavailable to the secondary users. Our approach is natural for continuous variables, where the estimation may be either based on reweighting or imputation, including possibly their outlier-robust extensions. We also propose a multivariate procedure to accommodate the estimation of the covariance matrix between estimated population totals, which facilitates variance estimation of the ratios or differences among the estimated totals. We illustrate the proposed approach using simulation data in supplementary materials that are available online.

KEYWORDS:
analysis of incomplete data, item nonresponse, missing data, variance estimation

1 | INTRODUCTION

In the preparation of survey data for use by secondary analysts, some or all of the sample units are usually assigned estimation weights that can be applied to all the survey variables. In addition to these weights, imputed values may be needed for the units that are subjected to item missingness. It is often possible to choose the imputed values for each survey variable so that, together with the observed and retained values of this variable, the corresponding population total can be estimated by weighting as if

[0]**Abbreviations:** ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

the sample were completely observed. However, applying standard complete-sample variance estimator formulae to the same imputed sample would generally cause bias (see, e.g., Wolter, 2007, pp. 419, 421).

Variance estimation in the presence of imputed data needs to appropriately account for the underlying data generation mechanism. Some common techniques are Fay's reverse framework (e.g. Fay, 1991, 1992; Shao & Steel, 1999; Kim & Rao, 2009), two-phase sampling (Särndal, 1992; Deville & Särndal, 1994b) and replication methods (Rao & Shao, 1992; Rao, 1996; Shao & Sitter, 1996; Chen, Rao, & Sitter, 2000). To choose and apply any of these methods may be difficult for secondary users who are non-specialists, even impossible if the relevant information about the sampling design and data processing is lacking.

The needs for easier secondary analyses ensuring that different users of the imputed data could obtain the same results by simple estimation methods have long been recognized (Kalton & Kasprzyk, 1982). Early work under this estimation perspective for the imputed data was addressed by Lanke (1983), Sedransk (1985) and Kim (2001a). Some later works considered the use of constrained or calibrated imputation for point estimation in different situations (Chambers & Ren, 2004; Beaumont, 2005; Chauvet, Deville, & Haziza, 2011; Gelein, Haziza, & Causeur, 2014). Multiple imputation (Rubin, 1978a, 1987b, 1996c; Rubin & Schenker, 1986) and fractional imputation (Kim & Fuller, 2004; Fuller & Kim, 2005) are two methods based on multiply imputed values. For instance, provided the multiple imputation procedure is proper or the congeniality condition (Meng, 1994) holds, one can compute valid point and variance estimates by combining the results obtained from applying standard complete-sample formulae to each imputed sample.

In this paper, we propose a calibrated imputation approach that allows for valid point and variance estimation of the population and domain totals (or means), by applying simple complete-sample formulae to the imputed sample. Although this accommodates a more limited scope than multiple or fractional imputation, secondary users can achieve the intended analyses based on a single imputed dataset using standard software. Moreover, we provide a multivariate procedure for the estimation of a vector of totals (or means) and the associated covariance matrix, using simple complete-sample formulae. This allows one to estimate the variance of ratios or differences of the estimated totals. Finally, the proposed approach has also benefits to the data producer, such as avoiding the dissemination of multiply imputed datasets, the freedom to choose a suitable inference outlook and apply different missing data treatments from one variable to another.

The rest of the paper is organized as follows. The proposed approach is outlined in Section 2. We explain in Section 3 how it can be applied in some common situations of reweighting and imputation-based estimation, as well as domain estimation and estimation under stratified multistage sampling. The multivariate calibrated imputation procedure is described in Section 4. Some concluding remarks and future research topics are given in Section 5.

2 | CALIBRATION OF A SINGLE VARIABLE

2.1 | Estimation Setup

Consider a finite population of N units denoted by $U = \{1, 2, ..., N\}$ and let $Y_N = (y_1, \dots, y_N)$, where y_k is the value of a survey variable y for the kth unit, $k \in U$. Let A be a sample from U selected by a probability sampling design $p(\boldsymbol{I}_N)$, where $\boldsymbol{I}_N = (i_1, \dots, i_N)$ and $i_k = 1$ if the kth unit is selected to the sample A and $i_k = 0$ otherwise, and let $\boldsymbol{R}_N = (r_1, \dots, r_N)$, where $r_k = 1$ if y_k is observed and $r_k = 0$ if the y_k is unobserved or rejected during data processing ($k \in U$), $A_r = \{k : k \in A, r_k = 1\}$ be the set of units for which the observations are to be preserved and $A_m = \{k : k \in A, r_k = 0\}$ be the set for which imputation is needed. We assume the missing information of the variable y for the units in A_m is filled in by imputation and denote the corresponding imputed values by $\{y_k^* : k \in A_m\}$. We assume, in addition, that the imputed dataset will be accompanied by a set of survey weights $\{w_k : k \in A\}$, as for instance, the inverse of the inclusion probabilities (Horvitz & Thompson, 1952), or weights suitably calibrated for auxiliary population totals (Deville & Särndal, 1992a).

Suppose it is of interest to estimate the population total of the variable y, that is $t_y = \sum_{k \in U} y_k$. When it comes to complete-sample estimation of t_y using the imputed data $\{(w_k, y_k^*) : k \in A\}$, where y_k^* is the value for unit k in the imputed full sample A with $y_k^* = y_k$ for $k \in A_r$, a natural and simple choice for the imputed estimator is

$$\hat{t}_{yI} = \sum_{k \in A} w_k y_k^*. \tag{1}$$

Statistical properties of $\hat{t}_{yI}$ are studied by adopting an *inference approach* for the imputed data, which is usually specified by postulating explicitly a model for the distribution of the response indicators or a superpopulation model for the values of the variables of interest in the population (Haziza, 2009, pp. 222-223). The properties of $\hat{t}_{yI}$ are then evaluated with respect to the joint distribution of the sampling design and the assumed model, allowing the unconditional variance of the imputed estimator to be decomposed into variance components which, when estimated, lead to the estimated variance of $\hat{t}_{yI}$.

Here we consider instead the estimation of the variance of the imputed estimator $\hat{t}_{yI}$ by means of the complete-sample estimator

$$\hat{v}_F(\hat{t}_{yI}) \equiv \frac{n}{n-1} \sum_{k \in A} (u_k^* - \bar{u}^*)^2, \tag{2}$$

where $u_k^* = w_k y_k^*$ $(k \in A)$ and $\bar{u}^* = \sum_{k \in A} u_k^*/n = \hat{t}_{yI}/n$, which amounts to the with-replacement *pps sampling* variance formula and, hence, may be computed more easily by secondary users using standard software. For example, when $w_k = N/n$ then $\hat{t}_{yI} = N\bar{y}_I$ and $Var(\hat{t}_{yI}) = N^2 s_{yI}^2/n$, where $\bar{y}_I$ and s_{yI}^2 are the sample mean and variance of the imputed variable.

Clearly, naive application of estimators (1) and (2) would lead to incorrect inference generally. In order for these estimators to yield valid estimates, the imputed values need to be created in a controlled manner, as it will be discussed in the next section.

2.2 | The calibrated imputation approach

The main goal of the following calibration method for the imputed data is to provide imputed values y_k^* so that the complete-sample estimators (1) and (2) satisfy

$$\hat{t}_{yI} \equiv \sum_{k \in A} w_k y_k^* = \hat{t}_{y0}, \quad \hat{v}_F(\hat{t}_{yI}) \equiv \frac{n}{n-1} \sum_{k \in A} \left(w_k y_k^* - \hat{t}_{yI}/n\right)^2 = \hat{v}_{y0}, \tag{3}$$

where $\hat{t}_{y0}$ and $\hat{v}_{y0}$ are valid *target* estimates for the population total and its corresponding variance estimate. The method requires the data producer to choose and calculate such targets for the variable specified, as well as to calibrate the imputed values to attain the conditions in (3). These targets should incorporate all the aspects of the sampling design, response mechanism and inference approach for the imputed estimator. However, as a benefit of the calibration method, the suitability of these target estimates is a matter of concern only for the data producer and not for the secondary users, who are no longer exposed to the theoretical and computational complications involved.

The calibration method can be described by the following two-step algorithm.

Calibration algorithm:

Step 1 (Imputation): Using a standard imputation procedure, obtain a set of initial imputed values $\{\tilde{y}_k : k \in A_m\}$. For each $\tilde{y}_k$, obtain a corresponding adjusted imputed value $\hat{y}_k$ so that

$$\sum_{k \in A_m} w_k \hat{y}_k = \hat{t}_{y0} - \sum_{k \in A_r} w_k y_k. \tag{4}$$

Step 2 (Calibration): For each $k \in A_r$, set $u_k^* = w_k y_k$. For $k \in A_m$, obtain an imputed value u_k^* by a minimal adjustment to $\hat{u}_k = w_k \hat{y}_k$, where $\hat{y}_k$ is computed in Step 1, so that

$$\sum_{k \in A_m} u_k^* = \hat{t}_{y0} - \sum_{k \in A_r} w_k y_k \tag{5}$$

and

$$\sum_{k \in A_m} u_k^{*2} = \frac{n-1}{n} \hat{v}_{y0} + \frac{1}{n} \hat{t}_{y0}^2 - \sum_{k \in A_r} w_k^2 y_k^2, \tag{6}$$

where $\hat{t}_{y0}$ and $\hat{v}_{y0}$ are the targets in (3). Take $y_k^* = u_k^* / w_k$ for $k \in A$.

The algorithm initiates in Step 1 by choosing an imputation scheme to provide preliminary imputed values $\hat{y}_k$, for $k \in A_m$, such that applying (1) with these values yields $\hat{t}_{y0}$. Provided the initial imputed values $\tilde{y}_k$ ($k \in A_m$) already yields $\hat{t}_{y0}$ by (1), one can simply take $\hat{y}_k = \tilde{y}_k$, for $k \in A_m$. An example is given in Section 3.1. Otherwise, the $\tilde{y}_k$ values need to be adjusted. One simple ratio adjustment of the initial imputed values is

$$\hat{y}_k = \frac{\left(\hat{t}_{y0} - \sum_{\ell \in A_r} w_\ell y_\ell\right)}{\sum_{\ell \in A_m} w_\ell \tilde{y}_\ell} \tilde{y}_k \quad (k \in A_m), \tag{7}$$

which is a special case of the *reverse calibration* approach of Chambers & Ren (2004), originally proposed for the estimation of t_y in the presence of survey outliers. Then, in Step 2, the calibration of the imputed values is made. Optimal imputed values that are calibrated to (5) and (6) could be computed in closed-form by applying Theorem 1 below. The proof of this theorem is shown in the Appendix.

Theorem 1. Consider initial values $\hat{a}_k$ and $d_k > 0$ for all k in a non-null set $D \subset A$. Suppose $\sum_{k \in D} d_k \hat{a}_k = t_1$ for some fixed constant t_1 and $\sum_{k \in D} d_k (\hat{a}_k - t_1/t_0)^2 > 0$, where $t_0 = \sum_{k \in D} d_k > 0$. Let $t_2 > t_1^2/t_0$ be a fixed constant . Then, the adjusted a_k values that minimize $\Delta = \sum_{k \in D} d_k (a_k - \hat{a}_k)^2$ subjected to the constraints

$$\sum_{k \in D} d_k a_k = t_1, \quad \sum_{k \in D} d_k a_k^2 = t_2, \tag{8}$$

are given by

$$a_k = t_1/t_0 + \beta(\hat{a}_k - t_1/t_0), \tag{9}$$

where

$$\beta = \left(\frac{t_2 - t_1^2/t_0}{\hat{t}_2 - t_1^2/t_0} \right)^{1/2}$$

and $\hat{t}_2 = \sum_{k \in D} d_k \hat{a}_k^2$.

The optimal calibrated imputed values y_k^* of Step 2 are obtained as follows. First, we take the values of the calibration conditions t_1 and t_2 of (8) as the right-hand sides of (5) and (6), namely

$$t_1 = \hat{t}_{y0} - \sum_{k \in A_r} w_k y_k \tag{10}$$

and

$$t_2 = (n-1)\hat{v}_{y0}/n + \hat{t}_{y0}^2/n - \sum_{k \in A_r} w_k^2 y_k^2.$$

Then, we set $D = A_m$, $d_k = 1$ and $\hat{a}_k = \hat{u}_k = w_k \hat{y}_k$ for all $k \in A_m$, where the $\hat{y}_k$ $(k \in A_m)$ are obtained in Step 1. Thus, it follows from (9) that the u_k^* values of Step 2 are

$$u_k^* \equiv a_k = t_1/m + \hat{\beta}(\hat{u}_k - t_1/m) \ \ (k \in A_m), \tag{11}$$

where t_1 is defined in (10) and

$$\hat{\beta} = \left\{ \frac{(n-1)\hat{v}_{y0}/n - \sum_{k \in A_r}(u_k - \hat{t}_{y0}/n)^2 - m\left(\hat{t}_{y0}/n - t_1/m\right)^2}{\sum_{k \in A_m}(\hat{u}_k - t_1/m)^2} \right\}^{1/2}.$$

The resulting calibrated imputed variable is

$$y_k^* = \begin{cases} y_k, & k \in A_r, \\ u_k^*/w_k, & k \in A_m. \end{cases} \tag{12}$$

Remark 1. The calibrated imputation method in (12) does not modify the observed values for units in the respondent set (A_r). The values that are actually modified are the calibrated $y_k^* = u_k^*/w_k$ values ($k \in A_m$), where the u_k^* values minimize the squared distance to the imputed values $\hat{u}_k = w_k \hat{y}_k$ $(k \in A_m)$ obtained in Step 1, that is, $\Delta = \sum_{k \in A_m}(u_k^* - \hat{u}_k)^2$. The resulting u_k^* values are obtained analytically as the "best" linear predictor of u_k based on the $\hat{u}_k$ $(k \in A_m)$, where the slope $\hat{\beta}$ of the regression line, given in (11), dictates how the empirical variance of the u_k^* relates to that of the $\hat{u}_k$ $(k \in A_m)$. In practice, unless the $\hat{y}_k$ values are created to have greater empirical variance over A_m than A_r, one may expect $\hat{\beta} > 1$. This is because the formula (2) is ostensibly aimed at a variance of the order n^{-1}, whereas the target $\hat{v}_{y0}$ is generally aimed at a variance of the order r^{-1}, where r is the size of A_r. Thus, in order for the two to be equal to each other, the imputed y_k^* values will need to have greater variation over A_m than the observed y_k over A_r.

Remark 2. Given the set of missing units A_m, the application of Theorem 1 to obtain the optimal solution (11) requires that

$$\hat{v}_{y0} > \frac{n}{n-1}\left\{ \sum_{k \in A_r}(u_k - \hat{t}_{y0}/n)^2 + m\left(t_1/m - \hat{t}_{y0}/n\right)^2 \right\} \tag{13}$$

and

$$\sum_{k \in A_m} (\hat{u}_k - t_1/m)^2 > 0. \tag{14}$$

Comparing (13) to (2), it is readily seen that, for the solution to the optimization problem in Step 2 to exist, the target estimate $\hat{v}_{y0}$ needs to be larger than the full-sample variance estimate (2) that would have been obtained had the missing values been imputed by the common value t_1/m. The second condition (14) demands that the sampling weights and the imputation scheme are such that the $\hat{u}_k = w_k \hat{y}_k$ values are different from t_1/m for at least one $k \in A_m$. This is not the case when mean imputation is used at Step 1 to fill in the missing values of an equal probability sample. In such a situation, the proposed approach could still be applied by adding some initial zero-mean noise to each mean imputed value. The calibration constraints ensure that this added variability will not affect the variance of the imputed estimator.

3 | SOME APPLICATIONS

We explain below how the two-step approach and Theorem 1 proposed in Section 2 can be applied in some general situations, which comprise reweighting and imputation-based estimation, as well as domain estimation and estimation under stratified multistage sampling.

3.1 | Ratio imputation

Suppose that, in addition to the survey variable y, there is an auxiliary variable x which is not affected by nonresponse. Assume a population ratio model ξ of the pairs $\{(x_k, y_k) : k \in U\}$, under which

$$E_\xi(y_k \mid x_k) = \beta_0 x_k, \;\; Var_\xi(y_k \mid x_k) = \sigma^2 x_k,$$

for some unknown parameters β_0 and σ^2. By ratio imputation under the model ξ, the missing y_k values are imputed as

$$\tilde{y}_k = \hat{\beta}_{0r} x_k \quad (k \in A_m),$$

where $\hat{\beta}_{0r} = \sum_{k\in A_r} w_k y_k / \sum_{k\in A_r} w_k x_k$, and $w_k = 1/\pi_k$, and π_k is the sample inclusion probability, for $k \in A$. The resulting imputed estimator of the population total t_y is

$$\hat{t}_{y0} = \sum_{k\in A_r} w_k y_k + \sum_{k\in A_m} w_k \tilde{y}_k = \hat{\beta}_{0r}\hat{t}_x,$$

where $\hat{t}_x = \sum_{k\in A} w_k x_k$ is the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) of the population total $t_x = \sum_{k\in U} x_k$. Mean imputation is a special case of ratio imputation with $x_k = 1$ for all $k \in A$, by which the imputed estimator $\hat{t}_{y0}$ reduces to $\hat{t}_{y0} = N\bar{y}_r$. Under the conditions of Theorem 1 of Kim & Rao (2009), a design and model consistent estimator of the variance of $\hat{t}_{y0}$ can be expressed as

$$\hat{v}_{y0} = \hat{v}_1 + \hat{v}_2, \tag{15}$$

where

$$\hat{v}_1 = \sum_{k\in A}\sum_{\ell\in A} \frac{(\pi_{k\ell} - \pi_k\pi_\ell)}{\pi_{k\ell}} w_k\hat{\eta}_k w_\ell \hat{\eta}_\ell, \quad \hat{v}_2 = \left(\frac{\hat{t}_x}{\hat{t}_{xr}}\right)^2 \sum_{k\in A_r} w_k \hat{e}_k^2,$$
$$\hat{\eta}_k = \hat{\beta}_{0r} x_k + \frac{\hat{t}_x}{\hat{t}_{xr}} r_k \hat{e}_k \quad (k \in A)$$

$\hat{t}_{xr} = \sum_{k\in A_r} w_k x_k$ and $\hat{e}_k = y_k - \hat{\beta}_{0r} x_k$.

However, to compute $\hat{v}_1$, the secondary user needs to have access to the matrix of the second-order inclusion probabilities $\{\pi_{k\ell} : k\ell \in A\}$, which are almost never disseminated together with the imputed sample. The proposed approach avoids this complication. To calibrate the ratio imputed values $\tilde{y}_k = \hat{\beta}_{0r} x_k$ $(k \in A_m)$, we notice that $\hat{y}_k = \tilde{y}_k$ already satisfies Step 1, since $t_1 = \hat{\beta}_{0r}\hat{t}_{xm}$ in Theorem 1. For Step 2, by (12) and (11), the calibrated imputed values are

$$y_k^* = u_k^* / w_k, \tag{16}$$

where $u_k^* = w_k y_k$ $(k \in A_r)$, and for $k \in A_m$,

$$u_k^* = \frac{\hat{\beta}_{0r}\hat{t}_{xm}}{m} + \hat{\beta}\hat{\beta}_{0r}\left(w_k x_k - \frac{\hat{t}_{xm}}{m}\right) \quad (k \in A_m)$$

and

$$\hat{\beta}^2 = \frac{\frac{(n-1)}{n}(\hat{v}_1 + \hat{v}_2) - \sum\limits_{k\in A_r}\left(w_k y_k - \frac{\hat{\beta}_{0r}\hat{t}_x}{n}\right)^2 - m\hat{\beta}_{0r}^2\left(\frac{\hat{t}_x}{n} - \frac{\hat{t}_{xm}}{m}\right)^2}{\hat{\beta}_{0r}^2 \sum\limits_{k\in A_m}\left(w_k x_k - \frac{\hat{t}_{xm}}{m}\right)^2}.$$

In the case of mean imputation and simple random sampling without replacement, (15) reduces to

$$\hat{v}_{y0} = N^2\left(\frac{1}{r} - \frac{1}{N}\right)s_{yr}^2, \tag{17}$$

where $s_{yr}^2 = \sum_{k \in A_r} (y_k - \bar{y}_r)^2/(r-1)$ and $\bar{y}_r$ is the observed respondent mean.

3.2 | Domain estimation

As a realistic setting for domain total estimation, in addition to the population total, consider a domain population partition $U = U_1 \cup \cdots \cup U_D$. Let the population total of domain U_d be

$$t_{dy} = \sum_{k \in U_d} y_k = \sum_{k \in U} \delta_{kd} y_k,$$

where the domain indicator δ_{kd}, $\delta_{kd} = 1$ if $k \in U_d$ and $\delta_{kd} = 0$ otherwise, is observed for all units in the sample A ($d = 1, \ldots, D$). Let $\hat{t}_{dy}$ be the target domain total estimator and $\hat{v}_{dy}$ its variance estimate. Domain estimation can be handled by separate calibration for each domain by the producer and application of the domain complete-data formulae by the secondary users, yielding $\hat{t}_{dyI} = \hat{t}_{dy}$ and $\hat{v}_F(\hat{t}_{dyI}) = \hat{v}_{dy}$, as explained in Section 2.

However, one is still interested in estimating the population total, in addition to the domain totals. Directly applying the complete-sample formula (1) to the domain-calibrated imputed sample would correctly estimate the population total. One can combine the domain variance estimates, as if the sampling were stratified by the domains. However, the resulting variance estimate is incorrect even when the domain total estimators are independent of each other, due to the additional term

$$v_b = \frac{n}{n-1} \sum_{d=1}^{D} n_d (\hat{t}_{dyI}/n_d - \hat{t}_{y0}/n)^2 = \frac{n^2}{n-1} V_n(\hat{t}_{dyI}/n_d),$$

where $V_n(\hat{t}_{dyI}/n_d)$ is the variance of $\hat{t}_{dyI}/n_d$ with respect to the empirical sample domain distribution function $(n_1/n, \ldots, n_D/n)$, since

$$V_n(\hat{t}_{dyI}/n_d) = \sum_{d=1}^{D} \frac{n_d}{n} (\hat{t}_{dyI}/n_d - \hat{t}_{y0}/n)^2 \quad \text{and} \quad \hat{t}_{y0}/n = E_n(\hat{t}_{dyI}/n_d) = \sum_{d=1}^{D} \frac{n_d}{n} (\hat{t}_{dyI}/n_d).$$

We propose to introduce a *domain estimation effect factor*, denoted by γ, and use

$$\hat{v}_F(\hat{t}_{yI}) = \gamma^2 \frac{n}{n-1} \sum_{k \in A} (w_k y_k^* - \hat{t}_{y0}/n)^2 = \hat{v}_{y0}. \tag{18}$$

The factor γ can be calculated after domain-calibrated imputation, and disseminated together with imputed sample.

In the separate domain calibration above, $\hat{v}_F(\hat{t}_{dyI})$ is built on the squared errors around $\hat{t}_{dyI}/n_d$. Consider using another complete-sample formula $\hat{v}_F(\hat{t}_{dyI})$, built around $\hat{t}_{yI}/n$ instead, where

$$\hat{v}_F(\hat{t}_{dyI}) = \frac{n_d}{n_d - 1} \sum_{k \in A} \delta_{kd} (w_k y_{kd}^* - \hat{t}_{y0}/n)^2.$$

We need to extend the calibration constraints as follows:

$$\begin{cases} \hat{t}_{dyI} = \sum_{k\in A} \delta_{kd} w_k y_k^* = \hat{t}_{dy} & \text{for } d = 1, ..., D \\ \hat{v}_F(\hat{t}_{dyI}) = \frac{n_d}{n_d-1} \sum_{k\in A} \delta_{kd}(w_k y_{kd}^* - \hat{t}_{y0}/n)^2 = \hat{v}_{dy} & \text{for } d = 1, ..., D \\ \hat{v}_F(\hat{t}_{yI}) = \gamma^2 \frac{n}{n-1} \sum_{k\in A}(w_k y_k^* - \hat{t}_{y0}/n)^2 = \hat{v}_{y0}. \end{cases} \tag{19}$$

In other words, we use δ_{kd} to identify the relevant observations for domain estimation, including the special case of $U_d = U$ and $\delta_{kd} \equiv 1$, and use $\hat{t}_{y0}/n$ in all the ultimate variance estimators, including domain variance estimation. We refer to (19) as the centred domain calibration approach.

Minimum adjustments of $\{\hat{y}_k; k \in A_m\}$ from Step 2 of the proposed approach can be achieved by Theorem 1 as well. To focus the idea, suppose negligible $1/n$ and $1/n_d$. Let $\{u_k^*; k \in A_{md}\}$ be the calibrated imputations in domain d, given by

$$u_k^* = t_{1d}/m + \beta_d(\hat{u}_k - t_{1d}/m),$$

where $t_{1d} = \hat{t}_{dy} - \sum_{k\in A_{rd}} w_k y_k$ is the constrained total of $u_k^* = w_k y_k^*$ in A_{md}. However, instead of choosing β_d such that

$$\beta_d^2 \sum_{k\in A_{md}} \left(\hat{u}_k - \frac{t_{1d}}{m}\right)^2 = \hat{v}_{dy} - \sum_{k\in A_{rd}} \left(u_k - \frac{\hat{t}_{dy}}{n_d}\right)^2 - m_d\left(\frac{\hat{t}_{dy}}{n_d} - \frac{t_{1d}}{m}\right)^2,$$

as under separate domain calibration, we should now choose β_d such that

$$\beta_d^2 \sum_{k\in A_{md}} \left(\hat{u}_k - \frac{t_{1d}}{m}\right)^2 = \hat{v}_{dy} - \sum_{k\in A_{rd}} \left(u_k - \frac{\hat{t}_{y0}}{n}\right)^2 - m_d\left(\frac{\hat{t}_{y0}}{n} - \frac{t_{1d}}{m}\right)^2.$$

This allows us to estimate the domain variance $\hat{v}_{dy}$ as in (19). The domain estimation effect factor γ can be calculated afterwards to satisfy (19). The conditions for the existence of solution are formally the same as discussed in Section 2.2. Provided domain-specific calibration, it is feasible as long as $\hat{t}_{y0}/n$ does not differ too much from $\hat{t}_{dy}/n_d$ in the different domains.

In practice one may be interested in multiple sets of (overlapping) domains. For example, a user may want to have estimates by region as well as estimates by industry. Insofar as the need is known in advance, the producer can apply the approach above to the ‘atomic domains’, which arise from crossing region and industry. In addition to the separate atomic-domain calibrated sample, one can supply a domain estimation factor for the population total, a set of domain estimation factors for each of the regions, and another set of factors for each industry.

3.3 | Stratified Multistage Sampling

Let the population U be partitioned into H strata of n primary sampling units (PSUs), where a sample A_h of n_h PSUs is selected separately within the hth stratum ($h = 1, \dots, H$; $n_1 + \cdots + n_H = n$). From each PSU in A_h, additional stages of sampling are undertaken until the selection of the ultimate sampling units (USUs). Let w_i, y_i and r_i be, respectively, the weight, the y-value and the response indicator for the ith USU. Let A_{hk} be the set of USUs in the kth selected PSU of the hth stratum, where $A_{rhk} = \{i : i \in A_{hk}, r_i = 1\}$ and $A_{mhk} = \{i : i \in A_{hk}, r_i = 0\}$.

By setting $y_i^* = y_i$ if $r_i = 1$ and letting y_i^* be the calibrated imputation value if $r_i = 0$, the imputed estimate of the population total t_y can be written as $\hat{t}_{yI} = \sum_{h=1}^{H} \hat{t}_{yIh}$, where $\hat{t}_{yIh} = \sum_{k\in A_h} u_{hk}^*$ and $u_{hk}^* = \sum_{i\in A_{hk}} w_i y_i^* = \sum_{i\in A_{rhk}} w_i y_i + \sum_{i\in A_{mhk}} w_i y_i^*$. For calibrated imputation that enables (3), we can apply Theorem 1 and the 2-step approach directly at the level of USUs, ignoring the clustering structure of the multistage sampling.

Survey data analysis softwares (such as STATA, R, SAS) commonly use the stratified ultimate variance formula for variance estimation. It is therefore convenient if the secondary user can simply input the imputed sample, and let the software carry on as usual. Thus, as another possibility of full-sample variance estimator, we consider

$$\hat{v}_F(\hat{t}_{yI}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{k\in A_h} (u_{hk}^* - \bar{u}_h^*)^2,$$

where $\bar{u}_h^* = \sum_{k\in A_h} u_{hk}^*/n_h = \hat{t}_{yIh}/n_h$. This choice fits naturally with the standard approach of ultimate-cluster variance estimation under stratified multistage sampling (e.g. Skinner, 1989, Section 2.13).

Given $\hat{t}_{y0h}$ for the population total in the hth stratum and its associated variances $\hat{v}_{y0h} = \hat{v}(\hat{t}_{y0h})$ ($h = 1, \dots, H$), consider the problem of finding the values y_j^* starting with $\tilde{y}_j$, $j \in \cup_{k\in A_h} A_{mhk}$, so that

$$\begin{aligned} &\hat{t}_{yIh} = \textstyle\sum_{k\in A_h} u_{hk}^* = \hat{t}_{y0h}, \\ &\textstyle\sum_{k\in A_h} (u_{hk}^* - \bar{u}_h^*)^2 = \sum_{k\in A_h} u_{hk}^{*2} - \hat{t}_{y0h}^2/n_h = (n_h - 1)\hat{v}_{y0h}/n_h. \end{aligned} \tag{20}$$

We propose to obtain a solution of this problem in two stages. First, the initial imputed PSU totals are adjusted minimally subject to the two constraints above, yielding the adjusted PSU total u_{hk}^*. Second, the initial imputed values $\tilde{y}_j$ are adjusted, separately within each PSU, to agree with the corresponding calibrated PSU total from the first step.

For the first stage, we can apply Theorem 1 within the hth stratum similarly as in Section 2. Let $A_{h0} = \{k \in A_h : \#(A_{mhk}) = 0\}$, $\hat{u}_{hk} = u_{hk}$ for $k \in A_{h0}$ and $\hat{u}_{hk} = \tilde{u}_{hk}(\hat{t}_{y0h} - \sum_{k\in A_{h0}} u_{hk})/(\sum_{\ell\in A_h\setminus A_{h0}} \tilde{u}_{h\ell})$ for $k \in A_h\setminus A_{h0}$. Then, take $D = D_h = A_h\setminus A_{h0}$,

$d_k = d_{hk} = 1$, $\hat{a}_k = \hat{a}_{hk} = \hat{u}_{hk}$, $t_0 = \sum_{k\in A_h\setminus A_{h0}} d_k \equiv m_h$, $t_1 = t_{1h} = \hat{t}_{y0h} - \sum_{k\in A_{h0}} u_{hk}$ and $t_2 = t_{2h} = (n_h-1)\hat{t}_{y0h}/n_h + \hat{t}^2_{y0h}/n_h - \sum_{k\in A_{h0}} u^2_{hk}$. For each $h = 1, \dots, H$, the optimal solution that minimizes the squared distance $\Delta_h = \sum_{A_h}(u^*_{hk} - \hat{u}_{hk})^2 = \sum_{A_h\setminus A_{h0}}(u^*_{hk} - \hat{u}_{hk})^2$ subject to (20) are given by $u^*_{hk} = u_{hk}$ for $k \in A_{h0}$, and

$$u^*_{hk} = \frac{\hat{t}_{1h}}{m_h} + \hat{\beta}_h\Big(\hat{u}_{hk} - \frac{\hat{t}_{1h}}{m_h}\Big) \tag{21}$$

for $k \in A_h\setminus A_{h0}$, where

$$\hat{\beta}_h = \left\{\frac{(n_h-1)\hat{v}_{y0h}/n_h - \sum_{A_{h0}}(u_{hk} - \hat{t}_{y0h}/n_h)^2 - m_h(\hat{t}_{y0h}/n_h - t_{1h}/m_h)^2}{\sum_{k\in A_h\setminus A_{h0}}(\hat{u}_{hk} - t_{1h}/m_h)^2}\right\}^{\frac{1}{2}}.$$

Having thus obtained u^*_{hk}, we adjust the $\tilde{y}_i$'s separately within each PSU so that $u^*_{hk} = \sum_{i\in A_{rhk}} w_i y_i + \sum_{i\in A_{mhk}} w_i y^*_i$, which is a single constraint. For given h and $k \in A_h\setminus A_{h0}$, the values y^*_i that minimize the distance $\sum_{i\in A_{mhk}}(y^*_i - \tilde{y}_i)^2/2$ subject to $\sum_{j\in A_{mhk}} w_i y^*_i = u^*_{hk} - \sum_{i\in A_{rhk}} w_i y_i \equiv u_{hk0}$ are

$$y^*_i = \tilde{y}_i\left\{1 + \Big(\frac{w_i}{\tilde{y}_i}\Big)\frac{\big(u_{hk0} - \sum_{i\in A_{mhk}} w_i\tilde{y}_i\big)}{\sum_{i\in A_{mhk}} w^2_i}\right\} \quad (i \in A_{mhk}). \tag{22}$$

4 | CALIBRATION OF MULTIPLE VARIABLES

Let $\boldsymbol{y}_k = (y_{k1}, \dots, y_{kp})^\top$ denote a p-dimensional vector of values for the k-th unit and $\boldsymbol{u}^*_k = w_k\boldsymbol{y}^*_k$, where $\boldsymbol{y}^*_k = (y^*_{k1}, \dots, y^*_{kp})^\top$ denote the calibrated imputed values having the restriction that $y^*_{k\ell} = y_{k\ell}$ if $y_{k\ell}$ is observed and fixed ($k \in A$ and $\ell = 1, \dots, p$). Following the basic algorithm of Section 2, consider the problem of finding the $\boldsymbol{u}^*_k$ satisfying

$$\sum_{k\in A}\boldsymbol{u}^*_k = \hat{\boldsymbol{t}}_0, \quad \frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2} = \hat{\boldsymbol{V}}_0,$$

where $\hat{\boldsymbol{t}}_0$ denotes a p-dimensional vector of target estimates for the population total $\boldsymbol{t}_y = \sum_{k\in U}\boldsymbol{y}_k$, $\hat{\boldsymbol{V}}_0$ denotes the target estimated variance-covariance matrix of $\hat{\boldsymbol{t}}_0$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\top$. To obtain $\hat{\boldsymbol{V}}_0$ in the presence of multivariate missing data is a difficult issue. See, e.g., Skinner & Rao (2002) and Chauvet & Haziza (2012) for a fully efficient approach in the bivariate case, and Im et al. (2018) and Sang & Kim (2018) for two fractional imputation methods in the multivariate setting. Below we propose a two-phase calibration procedure, where at the first phase the problem is solved for transformed vectors $\boldsymbol{v}^*_k$'s ($k \in A$), and at the second phase the results are back-transformed to $\boldsymbol{u}^*_k$'s as required. Assume without loss of generality the weighted values

$$\hat{\boldsymbol{u}}_k = w_k\hat{\boldsymbol{y}}_k \equiv (\hat{u}_{k1}, \dots, \hat{u}_{kp})^\top$$

satisfy $\sum_{k\in A}\hat{\boldsymbol{u}}_k = \hat{\boldsymbol{t}}_0$ preserving all the observed and fixed values. This can be achieved for instance by separate imputation of $y_{k\ell}$, for $\ell = 1, ..., p$, up to Step 1 in Section 2.2.

Let $\boldsymbol{D}$ be the subset of units with all items missing and suppose $m = |D| > 0$. Denote by $\hat{\boldsymbol{V}}_0 = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$ the spectral decomposition of $\hat{\boldsymbol{V}}_0$, where $\boldsymbol{P}$ is the $p \times p$ orthogonal matrix of the eigenvectors of $\hat{\boldsymbol{V}}_0$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of the corresponding eigenvalues. Consider the orthogonal principal components

$$\hat{\boldsymbol{v}}_k = \boldsymbol{P}^\top(\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{t}}_0/n) \quad (k \in A)$$

and set

$$\boldsymbol{v}_k^* = \hat{\boldsymbol{v}}_k \quad (k \in \boldsymbol{R} = A \setminus D). \tag{23}$$

Now, choose the $\{\boldsymbol{v}_k^* : k \in \boldsymbol{D}\}$ which minimize the squared Frobenius norm

$$\sum_{k\in D}(\boldsymbol{v}_k^* - \hat{\boldsymbol{v}}_k)^\top(\boldsymbol{v}_k^* - \hat{\boldsymbol{v}}_k)$$

subject to

$$\sum_{k\in D}\boldsymbol{v}_k^* = -\sum_{k\in R}\hat{\boldsymbol{v}}_k \equiv m\bar{\boldsymbol{v}}_D^*, \quad \sum_{k\in D}\boldsymbol{v}_k^{*\otimes 2} = \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}_k^{\otimes 2}.$$

The solution to this constrained optimization problem can be obtained by Theorem 2 below. The proof of this theorem is presented in the Appendix.

Theorem 2. Consider a set of vectors $\hat{\boldsymbol{a}}_k$ $(k \in D)$, satisfying $\sum_{k\in D}\hat{\boldsymbol{a}}_k = 0$ and $\boldsymbol{C} = \sum_{k\in D}\hat{\boldsymbol{a}}_k^{\otimes 2}$ being positive definite. Let $\boldsymbol{B}$ be a pre-specified positive definite matrix. Multivariate calibrated vectors that minimize $\sum_{k\in D}(\boldsymbol{a}_k^* - \hat{\boldsymbol{a}}_k)^\top(\boldsymbol{a}_k^* - \hat{\boldsymbol{a}}_k)$ subjected to

$$\sum_{k\in D}\boldsymbol{a}_k^* = 0, \quad \sum_{k\in D}\boldsymbol{a}_k^*\boldsymbol{a}_k^{*\top} = \boldsymbol{B}$$

are $\boldsymbol{a}_k^* = \boldsymbol{\beta}\hat{\boldsymbol{a}}_k$, where $\boldsymbol{\beta} = \boldsymbol{B}^{1/2}\boldsymbol{C}^{-1/2}$ and $\boldsymbol{B}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are the square root and inverse square root matrices of $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively.

In the first phase of the calibration method, we let $\boldsymbol{a}_k^* = \boldsymbol{v}_k^* - \bar{\boldsymbol{v}}_D^*$ and $\hat{\boldsymbol{a}}_k = \hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}_D^*$ in Theorem 2. If the matrices

$$\boldsymbol{B} \equiv \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}_k^{\otimes 2} - m\bar{\boldsymbol{v}}_D^{*\otimes 2}, \quad \boldsymbol{C} \equiv \sum_{k\in D}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}_D^*)^{\otimes 2}$$

are both positive definite, then the optimal calibrated $\boldsymbol{v}_k^*$ vectors by Theorem 2 are

$$\boldsymbol{v}_k^* = (\boldsymbol{I}_p - \boldsymbol{\beta})\bar{v}_D^* + \boldsymbol{\beta}\hat{\boldsymbol{v}}_k = \bar{\boldsymbol{v}}_D^* + \boldsymbol{\beta}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}_D^*) \quad (k \in D), \tag{24}$$

where $\boldsymbol{I}_p$ is the identity matrix of order p. Now, from $\sum_{k\in D}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D) = \sum_{k\in A}\hat{\boldsymbol{v}}_k = 0$ by (23), it follows that $\sum_{k\in D}\boldsymbol{v}^*_k = m\bar{\boldsymbol{v}}^*_D$ and $\sum_{k\in D}\boldsymbol{v}^{*\otimes 2}_k = m\bar{\boldsymbol{v}}^{*\otimes 2}_D + \boldsymbol{\beta}\boldsymbol{C}\boldsymbol{\beta}^\top = m\bar{\boldsymbol{v}}^{*\otimes 2}_D + \boldsymbol{B} = (n-1)n^{-1}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k$. We thus enter the second phase of imputation, where we transform the $\boldsymbol{v}^*_k$ $(k \in D)$ given by (23) and (24) back into

$$\boldsymbol{u}^*_k = \boldsymbol{P}\boldsymbol{v}^*_k + \hat{\boldsymbol{t}}_0/n = \begin{cases} \hat{\boldsymbol{u}}_k, & (k \in R = A \setminus D), \\ \boldsymbol{P}\{\bar{\boldsymbol{v}}^*_D + \boldsymbol{\beta}(\hat{\boldsymbol{v}}_k - \bar{\boldsymbol{v}}^*_D)\} + \hat{\boldsymbol{t}}_0/n, & (k \in D), \end{cases} \tag{25}$$

which are the final calibrated imputed vectors. The transformation (25) implies that

$$\sum_{k\in A}\boldsymbol{u}^*_k = \boldsymbol{P}\sum_{k\in A}\boldsymbol{v}^*_k + \hat{\boldsymbol{t}}_0 = \hat{\boldsymbol{t}}_0$$

and

$$\frac{n}{n-1}\sum_{i\in D}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2} = \frac{n}{n-1}\boldsymbol{P}\Big\{\sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k + \frac{n-1}{n}\boldsymbol{\Lambda} - \sum_{k\in R}\hat{\boldsymbol{v}}^{\otimes 2}_k\Big\}\boldsymbol{P}^\top = \hat{\boldsymbol{V}}_0,$$

as intended.

Notice that the vectors $\boldsymbol{u}^*_k$ in (25) can be computed directly from $\hat{\boldsymbol{u}}_k$ by the linear transformation

$$\boldsymbol{u}^*_k = \tilde{\boldsymbol{\alpha}}\mathbb{1}(k \in D) + \{\boldsymbol{I}_p\mathbb{1}(k \in A \setminus D) + \tilde{\boldsymbol{\beta}}\mathbb{1}(k \in D)\}\hat{\boldsymbol{u}}_k, \tag{26}$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{\beta}\boldsymbol{P}^\top$, $\tilde{\boldsymbol{\alpha}} = -(\boldsymbol{I}_p - \tilde{\boldsymbol{\beta}})(\sum_{k\in R}\hat{\boldsymbol{u}}_k - \hat{\boldsymbol{t}}_0)/m$ and $\mathbb{1}(\mathcal{A})$ is the indicator function of a set $\mathcal{A}$. All the observed and fixed data values are associated with the units in $R = A \setminus D$, and thus are preserved in the respective $\boldsymbol{u}^*_k$ values. The $\boldsymbol{u}^*_k$ and $\hat{\boldsymbol{u}}_k$ vectors share the same distance as that of the corresponding $\boldsymbol{v}^*_k$ and the $\hat{\boldsymbol{v}}_k$ vectors, for $k \in D$.

One of the benefits of the proposed calibrated imputation approach above is to facilitate inferences for a nonlinear function of population total vector $\boldsymbol{t}_y$. For example, suppose $\sqrt{n}(\hat{\boldsymbol{t}}_0 - \boldsymbol{t}_y)$ has an asymptotic multivariate normal distribution with mean zero and positive-definite variance-covariance matrix $\boldsymbol{V}$ and that the target variance $\hat{\boldsymbol{V}}_0$ is consistent for the variance of $\hat{\boldsymbol{t}}_y$. Thus, an approximate $100(1-\alpha)\%$ confidence interval for $g(\boldsymbol{t}_y)$, where $g : \mathbb{R}^p \longrightarrow \mathbb{R}$ is smooth and has nonzero gradient $\nabla g(\cdot)$ at $\boldsymbol{t}_y$, is

$$g\left(\hat{\boldsymbol{t}}_{yI}\right) \pm z_{1-\alpha/2}\left[\nabla^\top g\left(\hat{\boldsymbol{t}}_{yI}\right)\left\{\frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_k - \hat{\boldsymbol{t}}_0/n)^{\otimes 2}\right\}\nabla g\left(\hat{\boldsymbol{t}}_{yI}\right)\right]^{1/2},$$

where $\hat{\boldsymbol{t}}_{yI} = \sum_{k\in A}\boldsymbol{u}^*_k = \hat{\boldsymbol{t}}_0$, which is Wald-type interval that could alternatively be computed by

$$g\left(\hat{\boldsymbol{t}}_{yI}\right) \pm z_{1-\alpha/2}\left[\hat{v}_F\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\}\right]^{1/2}, \tag{27}$$

where

$$\hat{v}_F\left\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\right\} = \frac{n}{n-1}\sum_{k\in A}(\boldsymbol{u}^*_{gk} - \hat{\boldsymbol{t}}_{g0}/n)^{\otimes 2},$$

$\boldsymbol{u}^*_{gk} = \nabla^{\top} g\left(\hat{\boldsymbol{t}}_{yI}\right)\boldsymbol{u}^*_k$ and $\hat{\boldsymbol{t}}_{g0} = \nabla^{\top} g\left(\hat{\boldsymbol{t}}_{yI}\right)\hat{\boldsymbol{t}}_0$. The use of the estimate $\hat{v}_F\left\{g\left(\hat{\boldsymbol{t}}_{yI}\right)\right\}$ above does not merely simplifies the computation of the estimated variance of $g\left(\hat{\boldsymbol{t}}_{yI}\right)$ by a complete sample formula, but it also prevents the secondary users to having to apply a specific variance estimation method and corresponding software to estimate the variance of $g\left(\hat{\boldsymbol{t}}_{yI}\right)$.

5 | CONCLUDING REMARKS

In this paper we propose a calibrated imputation approach to be used routinely in the preparation of imputed sample survey data, under a stratified multistage sampling design. It allows secondary users to estimate specified population parameters and associated variances and covariances by simple complete-sample formulae, regardless of how complicatedly it may be when these are to be derived from the original incomplete data. Domain estimation can be accommodated in addition. The approach avoids the need to disseminate multiple imputed (or replicated) datasets.

A topic for future research is the imputation of categorical variables. The fractional imputation of Kim & Fuller (2004) provides a fully efficient imputation method, where K imputed values are created for a K-category variable. Favre et al. (2005) propose a fully efficient single-sample imputation method, under which a categorical variable is treated as a vector of dummy-indicators during the imputation. However, variance estimation under these methods may require a different approach than by using simple complete-sample formulae. It seems possible to adapt the multivariate calibrated imputation approach in this paper, where the imputed values of the dummy-vector can be continuous instead of just 0 and 1, given the target estimates of the population totals of each category and the associated variances and covariances.

Another research topic of interest is calibrated imputation for quantile estimation. For instance, given the target quantile estimate $\hat{t}_\alpha$ and its variance estimate $\hat{v}_\alpha$, for $0 < \alpha < 1$. It may be possible to adapt the proposed algorithm in Section 2.2. as follows. At Step 1, adjust the initial imputed values $\{\tilde{y}_k : k \in A_m\}$ to $\{\hat{y}_k : k \in A_m\}$, which satisfy

$$\sum_{k\in A_r} w_i I(y_k \leqslant \hat{t}_\alpha) + \sum_{k\in A_m} w_i I(\hat{y}_k \leqslant \hat{t}_\alpha) = \alpha \sum_{k\in A} w_i.$$

Then, at Step 2, adjust the negative and positive deviations $\hat{y}_k - \hat{t}_\alpha$, for $k \in A_m$, while maintaining their signs to satisfy a suitably chosen complete-sample variance formula.

ACKNOWLEDGMENTS

This research was developed at the Southampton Statistical Sciences Research Institute of the University of Souhampton, UK. The research of the first author was supported by a grant from the Brazilian Research Council (CNPq). We are thankful to the Associate Editor and two anonymous referees for their useful comments.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. J. R. Stat. Soc. Ser. B Stat. Methodol., 67(3), 445–458.

Chambers, R. L., & Ren, R. (2004). Outlier robust imputation of survey data. In *JSM Proceedings*, the Survey Research Methods Section, 3336–3344, Alexandria, VA: American Statistical Association.

Chauvet, G., Deville, J.-C., & Haziza, D. (2011). On balanced random imputation in surveys. Biometrika, 98(2), 459–471.

Chauvet, G., & Haziza, D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data. Canad. J. Statist., 40(1), 124–149.

Chen, J., Rao, J. N. K., & Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. Statist. Sinica, 10(4), 1153–1169.

Deville, J.-C., & Särndal, C.-E. (1992a). Calibration estimators in survey sampling. J. Amer. Statist. Assoc., 87, 376–382.

Deville, J.-C., & Särndal, C.-E. (1994b). Variance estimation for the regression imputed horvitz-thompson estimator. Journal of Official Statistics, 10, 381–394.

Favre, A.-C., Matei, A., & Tillé, Y. (2005). Calibrated random imputation for qualitative data. J. Statist. Plann. Inference, 128(2), 411 – 425.

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, 429-440, Washington, D.C.: U.S. Dept. of Commerce, Bureau of the Census.

Fay, R. E. (1992). When are inferences from multiple imputation valid? In *JSM Proceedings*, the Survey Research Methods Section, 227-232, Alexandria, VA: American Statistical Association.

Fuller, W. A., & Kim, J. K. (2005). Hot deck imputation for the response model. Survey Methodology, 31(2), 139–149.

Gelein, B., Haziza, D., & Causeur, D. (2014). Preserving relationships between variables with MIVQUE based imputation for missing survey data. J. Multivariate Anal., 131(0), 197 – 208.

Ghosh, M. (1992). Constrained bayes estimation with applications. J. Amer. Statist. Assoc., 87(418), 533–540.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics, Volume 29A, Sample surveys: Design, methods and applications* (pp. 215-246). Amsterdam: Elsevier.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc., 47, 663–685.

Im, J., Cho, I. H., & Kim, J. K. (2018). FHDI: An R package for fractional hot deck imputation. The R Journal, 10(1), 140–154.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. In *JSM Proceedings*, the Survey Research Methods Section, 22–31, Alexandria, VA: American Statistical Association.

Kim, J. K. (2001a). Variance estimation afeter imputation. Survey Methodology, 27(1), 75–83.

Kim, J. K. (2011b). Parametric fractional imputation for missing data analysis. Biometrika, 98(1), 119–132.

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. Biometrika, 91(3), 559–578.

Kim, J. K., & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. Biometrika, 96(4), 917–932.

Lanke, J. (1983). Hot deck imputation techniques that permit standard methods for assessing precision of estimates. Statistical Review, 21(5), (Essays in Honour of Tore E. Dalenius), 105–110.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statist. Sci., 9(4), 538–558.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. J. Amer. Statist. Assoc., 91(434), 499–506.

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. Biometrika, 79, 811–822.

Rubin, D. B. (1978a). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. In *JSM Proceedings*, the Survey Research Methods Section, 20-28, Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987b). Multiple imputation for nonresponse in surveys. Hoboken, NJ: Wiley.

Rubin, D. B. (1996c). Multiple imputation after 18+ years. J. Amer. Statist. Assoc., 91(434), 473–489.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc., 81(394), 366–374.

Sang, H., & Kim, J. K. (2018). Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. Unpublished manuscript. arXiv:1809.05976. Retrieved from https://arxiv.org/abs/1809.05976. URL https://arxiv.org/abs/1809.05976

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18, 241–252.

Sedransk, J. (1985). The objectives and practice of imputation. In *Proceedings of the First Annual Research Conference*, 445–452, Washington, D.C.: Bureau of the Census.

Shao, J., & Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. J. Amer. Statist. Assoc., 91(435), 1278–1288.

Shao, J., & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. J. Amer. Statist. Assoc., 94(445), 254–265.

Skinner, C., & Rao, J. (2002). Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. J. Statist. Plann. Inference, 102(1), 149 – 167.

Skinner, C. J. (1989). Introduction to part A. In C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 23-58). Chichester: Wiley.

Wolter, K. M. (2007). Introduction to variance estimation (2nd ed.). New York: Springer.

SUPPORTING INFORMATION

The supporting information, available as part of the online article, provides numerical illustrations for the proposed approach to estimate a population total and to preserve variability. Examples consider element sampling, stratified multistage sampling and also calibration of imputed multivariate data.

□

APPENDIX

A PROOF OF THEOREM 1

The proof of Theorem 1 is shown below. In the special case of $d_k = 1$, the problem addressed by Theorem 1 reduces essentially to that of the small-area constrained Bayes estimators (Ghosh, 1992, Theorem 1), which minimize an unweighted quadratic function of the prediction errors. The two constraints in Ghosh (1992, Theorem 1), despite being formulated in a small area estimation setting, are equivalent in form to (8).

Proof. Notice that

$$\sum_{k \in D} d_k (a_k - \hat{a}_k)^2 = t_2 - 2 \sum_{k \in D} d_k a_k \hat{a}_k + \hat{t}_2.$$

Since t_2 and $\hat{t}_2$ are positive and fixed, one needs to maximize $\sum_{k \in D} d_k a_k \hat{a}_k$ or, equivalently, $\sum_{k \in D} q_k a_k \hat{a}_k$, where $q_k = d_k / t_0$ and $\sum_{k \in D} q_k = 1$. Consider then $(q_1, ..., q_m)$ as a discrete distribution function, denoted by $\mathcal{F}_q$, which puts probability q_k on each pair of $(a_k, \hat{a}_k)$, for $k \in D$. We have $\sum_{k \in D} q_k a_k \hat{a}_k = E_q(a_k \hat{a}_k)$, i.e., the expectation over $\mathcal{F}_q$. Notice that $E_q(a_k) = \sum_{k \in D} q_k a_k = t_1 / t_0$ and $E_q(\hat{a}_k) = t_1 / t_0$ are both fixed, as well as $V_q(a_k) = t_2 / t_0 - (t_1 / t_0)^2$ and $V_q(\hat{a}_k) = \hat{t}_2 / t_0 - (t_1 / t_0)^2$. Thus, the maximum of $E_q(a_k \hat{a}_k)$ is the same as that of $Cov_q(a_k, \hat{a}_k)$, which is given when the correlation between a_k and $\hat{a}_k$ is equal to 1 over $\mathcal{F}_q$, i.e., $a_k = \alpha + \beta \hat{a}_k$ with probability one, for some constants α and β. Solving (α, β) for (8) yields then $\alpha = (1 - \beta) t_1 / t_0$ and $\beta = [(t_2 - t_1^2 / t_0) / (\hat{t}_2 - t_1^2 / t_0)]^{1/2}$. This completes the proof. □

B PROOF OF THEOREM 2

Proof. Let the Lagrangian of this constrained optimisation problem be

$$L = \frac{1}{2}\sum_{k\in D}(\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*)^\top(\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*) - \boldsymbol{\psi}^T \sum_{k\in D}\tilde{\boldsymbol{a}}_k^* - \mathbf{1}^\top[\boldsymbol{\Lambda}\circ(\sum_{k\in D}\tilde{\boldsymbol{a}}_k^*\tilde{\boldsymbol{a}}_k^{*\top} - \boldsymbol{S}_D^*)]\mathbf{1}$$

for some vector $\boldsymbol{\psi}$ and matrix $\boldsymbol{\Lambda}$, where $\mathbf{1}$ denotes a vector of ones and "$\circ$" denotes the Hadamard (element-wise) product of two matrices. We have

$$(\partial L/\partial\tilde{\boldsymbol{a}}_k^*)^\top = (\tilde{\boldsymbol{a}}_k^* - \boldsymbol{a}_k^*) - \boldsymbol{\psi} - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)\tilde{\boldsymbol{a}}_k^* = 0 \quad \Rightarrow \quad \tilde{\boldsymbol{a}}_k^* = [\boldsymbol{I}_p - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)]^{-1}(\boldsymbol{a}_k^* + \boldsymbol{\psi})$$

provided $\boldsymbol{W} = [\boldsymbol{I}_p - (\boldsymbol{\Lambda}+\boldsymbol{\Lambda}^\top)]^{-1}$ exists. Notice that $\boldsymbol{W}$ is symmetric. Substitution of $\tilde{\boldsymbol{a}}_k^* = \boldsymbol{W}(\boldsymbol{a}_k^* + \boldsymbol{\psi})$ into $\sum_{k\in D}\boldsymbol{a}_k^* = 0$ yields $\boldsymbol{\psi} = 0$. Next, substitution of $\tilde{\boldsymbol{a}}_k^* = \boldsymbol{W}\boldsymbol{a}_k^*$ into $\sum_{k\in D}\boldsymbol{a}_k^*\boldsymbol{a}_k^{*\top} = \boldsymbol{B}$ yields

$$\boldsymbol{B}\boldsymbol{B}^\top = \sum_D \boldsymbol{W}\boldsymbol{a}_k^*\boldsymbol{a}_k^{*\top}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{W} \quad \Rightarrow \quad \boldsymbol{W} = \boldsymbol{B}\boldsymbol{C}^{-1}.$$

This completes the proof. □