Calibrated hot deck imputation for numerical data under edit restrictions

de Waal, A.G.; Coutinho, Wieger; Shlomo, Natalie

Published in:
Journal of Survey Statistics and Methodology

DOI:
10.1093/jssam/smw037

Publication date:
2017

Document Version
Peer reviewed version

Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions

Ton de Waal[1], Wieger Coutinho[2], Natalie Shlomo[3]

[1] Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands (email: t.dewaal@cbs.nl) & Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

[2] Loket Aangepast-Lezen, PO Box 84010, 2508 AA The Hague, The Netherlands

[3] Social Statistics, School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester, M13 9PL, United Kingdom (e-mail: natalie.shlomo@manchester.ac.uk)

Abstract: We develop a non-parametric imputation method for item non-response based on the well-known hot-deck approach. The proposed imputation method is developed for imputing numerical data that ensure that all record-level edit rules are satisfied and previously estimated or known totals are exactly preserved. We propose a sequential hot-deck imputation approach that takes into account survey weights. Original survey weights are not changed, rather the imputations themselves are calibrated so that weighted estimates will equal known or estimated population totals. Edit rules are preserved by integrating the sequential hot-deck imputation with Fourier-Motzkin elimination which defines the range of feasible values that can be used for imputation such that all record-level edits will be satisfied. We apply the proposed imputation method under different scenarios of random and nearest-neighbour hot-deck on two data sets: an annual structural business survey and a synthetically generated data set with a large proportion of missing data. We compare the proposed imputation methods to standard imputation methods based on a set of evaluation measures.

Keywords: Item non-response, Edit rules, Survey weights, Fourier-Motzkin elimination

Word Count 6891 (excluding abstract, figures, tables, references, and appendices)

Formatted: Line spacing: At least 14 pt

1. Introduction

Missing data form a well-known problem that has to be dealt with by agencies collecting data on persons or enterprises. Missing data can arise from unit non-response or item-nonresponse. Unit non-response occurs when units that are selected for data collection cannot be contacted, refuse or are unable to respond altogether, or respond to so few questions that their response is deemed useless for analysis or estimation purposes. Unit non-response is usually corrected by weighting the responding units (see, e.g., Särndal, Swensson and Wretman 1992 and Särndal and Lundström 2005).

Item non-response occurs when data on only some of the items in a record, i.e. the data of an individual respondent, are missing. Persons may, for instance, refuse to provide information on their income or on their sexual habits, while at the same time give answers to other, less sensitive questions on the questionnaire. Enterprises may not provide answers to certain questions, because they may consider it too complicated or too time-consuming to answer these specific questions. The most common solution to handle item non-response is imputation, where missing values are filled in with plausible estimates. There is an abundance of literature on imputation of missing data. We refer to Kalton and Kasprzyk (1986), Rubin (1987), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), Longford (2005), Andridge and Little (2010), De Waal, Pannekoek and Scholtus (2011), Van Buuren (2012) and references therein. In this paper we focus on item non-response for numerical data, and whenever we refer to missing data in this paper we will be referring to missing data due to item non-response, unless indicated otherwise.

In many cases, especially at National Statistical Institutes, data have to satisfy constraints in the form of edit restrictions, or edits for short. These edit restrictions constrict the imputation of missing data. Examples of such edits are that the profit of an enterprise equals its turnover minus its costs, and that the turnover of an enterprise should be at least zero. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. Despite the abundance of literature on imputation, imputation of numerical data under edit restrictions is a rather neglected research area. Approaches for

imputation of numerical data under edit restrictions have been developed by Geweke (1991), Raghunathan et al. (2001), Tempelman (2007), Holan et al. (2010), Coutinho, De Waal and Remmerswaal (2011), Coutinho and De Waal (2012), Pannekoek, Shlomo and De Waal (2013) and Kim et al. (2013). For categorical data under edit restrictions some work has been done by Winkler (2003 and 2008).

A further complication is that numerical data sometimes have to sum up to known or previously estimated totals. This situation can arise with a 'one-figure' policy, where an agency aims to publish only one estimate for the same phenomenon occurring in different tables. Statistics Netherlands pursues a 'one-figure' policy for the Dutch Census (Statistics Netherlands 2014), as well as for many other statistics. As an example, when budgets for municipalities are decided upon by national statistics, these statistics must be numerically consistent across all statistical outputs of the agency. Therefore, when a 'one-figure' policy is pursued estimates need to be calibrated to the previously estimated or known totals and this takes precedence over other desired statistical properties.

Cox (1980) proposed the weighted sequential hot deck method, which was designed so that means and proportions estimated using the imputed data will be equal in expectation to the weighted mean or proportion estimated using respondent data only. This differs from our aim where we want totals for the imputed data to be exactly equal to known or previously estimated totals as well as preserve record level edit restrictions.

For numerical data, Zhang and Nordbotten (2008) and Zhang (2009) have extended nearest hot deck imputation so that totals are (approximately) preserved. There is, however, no guarantee of numerical consistency for a 'one figure' policy. In addition, these authors do not consider edit restrictions. Also for numerical data, Beaumont (2005) has developed a calibrated imputation approach that in principle can deal with edit restrictions. This approach is based on solving a mathematical optimization problem that can be exceedingly large. Beaumont (2005) only reports results for a simulation study that does

not involve edit restrictions. The computational feasibility of the approach for data sets involving edit restrictions is as yet unknown. Pannekoek, Shlomo and De Waal (2013) have developed imputation methods for numerical data based on regression models with random residuals that ensure that edits are satisfied and previously estimated or known totals are preserved. Drawbacks of the methods developed by Pannekoek, Shlomo and De Waal (2013) are that they are relatively complex and are based on parametric model assumptions. In particular, these methods assume the data to be either normally or log-normally distributed. For data that are distributed otherwise these methods may give biased results.

Favre, Matei and Tillé (2005) and Coutinho, De Waal and Shlomo (2013) have developed methods for categorical data having to satisfy edits and to preserve totals. An obvious difference between those methods and our approach is we focus on numerical data.

The aim of this paper is to develop non-parametric imputation methods that lead to satisfied edits and preserved totals. In order to avoid problems of model misspecification and reduce the reliance on parametric assumptions, we base our approach on well-known hot deck approaches, similar to Zhang and Nordbotten (2008) and Zhang (2009). The main objective of our imputation methods is to obtain accurate point estimates while satisfying all edits and preserve totals.

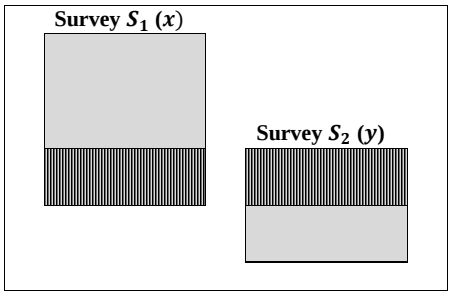
The remainder of this paper is organized as follows. Section 2 describes calibrated imputation in comparison to calibration via survey weights. Section 3 introduces the edit restrictions and sum constraints due to known or previously estimated totals we consider in this paper. Section 4 develops sequential hot deck imputation algorithms for our imputation problem. Section 5 describes an evaluation study and its results. Finally, Section 6 concludes with a brief discussion.

2. Calibrated Imputation versus Calibrated Weighting

Calibrated weighting is a well-known way of calculating survey weights for survey data which benchmarks known or estimated population totals on selected auxiliary variables. Calibration means that the weighted sample counts of the auxiliary variables equal the benchmarked population totals. Calibrated imputation is the lesser known equivalent of calibrated weighting. When calibrated imputation is used, the imputations are selected so that population estimates for selected variables obtained from the imputed data are equal to known or previously estimated totals. The survey design weights do not need to change.

We will discuss why in some cases it is useful to use calibrated imputation instead of calibrated weighting (see also Pannekoek, Shlomo and De Waal 2013). Let us suppose that first a survey S_1 with a numerical variable x becomes available and later a second survey S_2 with a categorical variable y. Figure 1 illustrates the case. In the rows we have the units in S_1 and S_2 and in the columns the variables in these data sets. The shaded parts in this figure indicate the overlapping records in these surveys.

Figure 1. Two partly overlapping surveys



In such a case, one could first use the data in survey S_1 to estimate the population total of x. When a later survey S_2 becomes available one could use the overlap between surveys S_1 and S_2 to estimate the population totals of the breakdown of x into categories of y. If a 'one-figure' policy is pursued, the

total estimate for x based on the overlap of S_1 and S_2 should be equal to the original estimate based on survey S_1. This can be achieved by calibrated weighting (see, e.g., Särndal and Lundström 2005) or by applying a calibrated imputation approach as we propose in the current paper.

In the usual sample survey setting, units are randomly selected from a population according to a specified sampling design, where each population unit is included in the sample s with a non-zero probability. Estimates of population totals and other parameters of interest are then calculated by using survey weights w_i that are the inverse of the inclusion probabilities adjusted for non-response and calibrated to known or estimated population totals. The correction for the unbalanced sample by calibrating the weights will affect estimates for all variables.

The situation with item-nonresponse differs from the situation with (only) unit non-response as item non-response fractions can vary greatly between variables. Adjustment to unit level weights only is therefore no longer an option. To deal with item non-response one usually first imputes the missing items for the responding units so that a complete data set is obtained. Next, the responding units are weighted to correct for unit non-response. In this weighting step, calibration weighting can again be used to ensure that estimates of totals will be equal to the known or estimated population totals.

However, for variables with imputed values, differences between estimated totals and their known values are now not only caused by unit non-response, but also by systematic errors in the imputed values due to misspecification of the imputation model. We will refer to these errors as *imputation bias*. The weight adjustments due to calibration hence do not correct for an unbalanced selection of units only but also for imputation bias in specific variables. There is no compelling reason to let this adjustment for imputation bias affect the estimates of all other variables. This makes calibration weighting less desirable in the case of item-nonresponse.

A drawback of weighting in general is that it can only be applied to so-called monotone missing data patterns and not to general missing data patterns. A missing data pattern is called monotone if the variables can be ordered such that the values of variables X_{j+1} to X_p are missing whenever the value of variable X_j is missing, for all $j = 1, \ldots, p-1$, where p is the number of variables.

In this paper we therefore develop a different approach, where we do not change the design weights of the responding units. Instead of calibrating the weights we will calibrate the imputations so weighted estimates will equal known or estimated population totals.

3. Constraints on Imputed Data

Edit rules imply restrictions within records on the values of the variables. Edits for numerical data are either linear equations or linear inequalities. We denote the number of variables by p. Edit k ($k = 1, \ldots, K$) can then be written in either of the two following forms:

$$a_{1k}x_{i1} + \cdots + a_{pk}x_{ip} + b_k = 0 \quad (1a)$$

or

$$a_{1k}x_{i1} + \cdots + a_{pk}x_{ip} + b_k \geq 0 \quad (1b)$$

where the a_{jk} and the b_k are certain constants, which define the edit. In many cases, the b_k will be equal to zero.

Edits of type (1a) are referred to as balance edits. An example of such an edit is

$$T_i - C_i - P_i = 0 \quad (2)$$

where T_i is the turnover of an enterprise corresponding to the i-th record $(i = 1, \ldots, m)$, P_i its profit, and C_i its costs. Edit (2) expresses that the profit of an enterprise equals its turnover minus its costs. Edits of type (1b) are referred to as inequality edits. An example of such an edit is

$$T_i \geq 0 \tag{3}$$

expressing that the turnover of an enterprise should be non-negative.

Sum constraints due to known or previously estimated population totals can be expressed as

$$\sum_{i \in r} w_i x_{ij} = X_j^{pop}$$

with w_i the survey weights that have been adjusted for unit nonresponse, r the set of respondents and X_j^{pop} the known population total of variable X_j . ~~and r the set of respondents.~~

4. Sequential Hot Deck Imputation Satisfying Edits and Totals

4.1 The basic idea

The imputation methods we apply in this paper are all based on a hot deck approach. When hot deck imputation is used, for each record containing missing values, the so-called recipient record, one uses the values of one or more other records where these values are observed, the so-called donor record(s), to impute the missing values (see Andridge and Little 2010).

Usually, hot deck imputation is applied in a multivariate fashion,~~multivariately,~~ that is several missing values in a recipient record are imputed simultaneously, using the same donor record. This approach aims to preserve the correlation structure in the data. For our problem that approach is often not

feasible. If an imputed record failed the edits, all one could do in such an approach is to reject the donor record and try another donor record. For a relatively complicated set of edits, one may have to test many different potential donor records until a donor record is found that leads to an imputed record satisfying all edits. Moreover, for some recipient records one may not be able to find any donor records such that the resulting imputed records satisfy all edits, let alone imputed records that at the same time also preserve totals.

Our imputation methods, in principle, aim for multivariate hot deck imputation where all imputations are taken from the same donor. When that is not possible, our imputation methods automatically switch to sequential imputation, where for different variables in a recipient record a different donor may be used. In particular, for variables with missing values involved in balance edits it is rare to be able to find a donor satisfying all balance edits. This means that for some of those variables our imputation methods switch to sequential imputation. Generally, the first missing fields in a record are imputed using the same donor, whereas the value of the last variable involved in a balance edit will not be based on that donor (or another donor), but will be determined deterministically from the other values involved in this balance edit.

The hot deck imputation methods we apply are described in Subsection 4.3. These hot deck imputation methods are used to order the potential imputation values for a certain missing field. Whether a value is actually used to impute the missing field depends on whether the edits and totals can be satisfied. The approach for hot deck imputation that ensures edits and sum constraints is described in Section 4.2 below.

4.2 Using a Sequential Approach to Imputation

In order to be able to use a sequential approach to imputation, we apply Fourier-Motzkin elimination (Duffin 1974 and De Waal, Pannekoek and Scholtus 2011). Fourier-Motzkin elimination is a technique to project a set of linear constraints involving q variables onto a set of linear constraints involving $q - 1$

variables. It is guaranteed to terminate after a finite number of steps. The essence of Fourier-Motzkin elimination is that two constraints, say $L(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq}) \leq x_{ij}$ and $x_{ij} \leq U(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq})$, where x_{ij} is the variable to be eliminated in a record i and $L(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq})$ and $U(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq})$ are linear expressions in the other variables, lead to a constraint

$$L(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq}) \leq U(x_{i1,}\dots,x_{i,j-1},x_{i,j+1},\dots,x_{iq})$$

involving these other variables. The main property of Fourier-Motzkin elimination is that the original set of constraints involving q variables can be satisfied if and only if the corresponding projected set of constraints involving q-1 variables can be satisfied.

Now, suppose we want to impute a record with some missing items. By repeated application of Fourier-Motzkin elimination we can derive an admissible interval for one of the values to be imputed in this record. The main property of Fourier-Motzkin guarantees that if we impute a value within this admissible interval, the remaining missing items in this record can be imputed in a manner consistent with the constraints, i.e. such that all constraints are satisfied.

Say we want to impute a variable x_j. We consider all the records in which the value of variable x_j is missing. In order to impute a missing field x_{ij} in a record i, we first fill in the observed and previously imputed values (if any) for the other variables in record i into the edits. This leads to a reduced set of edits involving only the remaining variables to be imputed in record i.

Next, we eliminate all equations from this reduced set of edits for record i. That is, we sequentially select any equation and one of the variables x $(x \neq x_j)$ involved in the selected equation. We then express x in terms of the other variables in the selected equation, and substitute this expression for x into the other edits in which x is involved. In this way we obtain a set of edits involving only inequality

restrictions for the remaining variables in record i. Once we have obtained imputation values for the variables involved in this set of inequalities, it is guaranteed that we can later find values consistent with the edits for the variables that were used to eliminate the equations in record i by means of back-substitution.

From the set of inequality restrictions we eliminate any remaining variables except x_{ij} itself by means of Fourier-Motzkin elimination. Using this elimination technique guarantees that the eliminated variables can later be imputed themselves such that all edits become satisfied.

After Fourier-Motzkin elimination the restrictions for x_{ij} can be expressed as interval constraints:

$$l_{ij} \leq x_{ij} \leq u_{ij}, \tag{4}$$

where l_{ij} may be $-\infty$ and u_{ij} may be ∞..We have such an interval constraint (4) for each record i in which the value of variable x_j is missing. Now, the problem for variable x_j is to fill in the missing values with imputations, such that the sum constraint for variable x_j and the interval constraints (4) are satisfied. For this we will use one of our sequential imputation algorithms as explained in Sections 4.3 and 4.4.

In Appendix A we illustrate how a sequential approach can be used.

4.3 Hot deck imputation methods

In this paper we apply two classes of hot deck imputation methods: nearest-neighbour imputation and random hot deck imputation. We describe these methods below.

4.3.1 Nearest-neighbour hot deck imputation

Suppose we want to impute a certain variable x_j $(j = 1, \dots, q)$ in a record i_o. In the nearest-neighbour approach we calculate for each other record i for which the value of x_j is not missing a distance to record i_o given by some distance function.

Before we calculate these distance functions, we first scale the values. We denote the scaled value of variable x_j in record i by x_{ij}^*. We determine the scaled value x_{ij}^* by

$$x_{ij}^* = \frac{x_{ij} - med_j}{s_j}$$

where med_j is the median of the observed values for variable x_j and s_j the interquartile distance, i.e. the difference between the value of the 75% percentile and the 25% percentile of x_j. We have used the median and the interquartile distance rather than the mean and standard deviation for scaling the variables in order to be more robust against possible ~~outliers~~ influential observations ~~units~~ in the data. Note that if one wants to apply outlier-robust imputation in practice, one should use a more complicated approach that on the one hand prevents influential observations from having too much effect on estimates, but on the other hand also ensures that these influential observations are taken into account to a sufficient extent (see, e.g., Chambers and Ren 2004). Since outlier-robust imputation is not the topic of the current paper, we have opted for our simple approach.

In our evaluation study we have used ~~a~~ the ~~different~~ distance function L_2 defined by

$$L_2\left(\mathbf{x}_{i_o}^*, \mathbf{x}_i^*\right) = \sqrt{\sum_{j \in Obs} (x_{i_o j}^* - x_{ij}^*)^2} \qquad (5)$$

where $\mathbf{x}_{i0}^{*}$ is the scaled recipient record and $\mathbf{x}_{i}^{*}$ a scaled potential donor record. Obs is the set of observed variables in the recipient record $\mathbf{x}_{i0}$.

An alternative to using nearest-neighbour imputation would be to apply predictive mean matching (see, e.g., Little 1988). We have not examined this option and leave this to future research.

In our implementation of nearest-neighbour imputation, we first apply a standard imputation routine in order to obtain a complete data set for calculating the distance function. In this case, we implemented the multiple imputation routine in SAS, which we will refer to as "MI-SAS" (see SAS 2015 for details on MI-SAS), and obtain a complete data set by averaging the imputations across 10 replicates. MI-SAS uses a Markov Chain Monte Carlo method for arbitrary missing data patterns assuming multivariate normal data. See Schafer (1997) for more details of this method. The reason for selecting MI-SAS rather than another imputation method is simply that this is a frequently used imputation routine albeit based on parametric assumptions. Alternatively, we could have applied other good imputation routines, such as "surveyimpute", also available in SAS.

Having now a complete dataset after the prior imputation by means of MI-SAS, we can compute distances between all pairs of records. Note that the imputation by MI-SAS introduces some uncertainty. This uncertainty may influence the order of the potential donor records. However, without this prior imputation step we would have had to find a way for dealing with missing values in our distance functions which would have increased the complexity.

Further motivations for carrying out the prior imputation step to obtain a complete dataset for calculating distance functions are:

- If the parametric assumptions of the initial imputation step resembles the model for the true data, then this step has the potential to improve the final imputations.

- The potential donor records for a certain recipient record are ordered in the same way for each variable with missing values. This means that, if possible, multivariate imputation, using several values from the first potential donor record on the list, is used. Only if a value of the first potential donor record is missing or cannot be used because this were to lead to failing edits or a non-preserved total, a value from another potential donor record is considered.

The imputed complete data set is only used to compute distances between recipient records and potential donor records. Only records with an observed value for the current variable to be imputed are considered while calculating these distances, so only actually observed values will be selected for actual imputation.

For each recipient record we construct a list of potential donor values in increasing order of the distance function (5). To impute a missing value, we will first select the first potential donor value on this list, i.e. the potential donor value from the record with the smallest distance to the recipient. If the value is allowed according to the edits and totals (see Section 4.4 for when a value is allowed according to the edits and totals), we will use it to impute the missing value. If that value is not allowed according to the edits or totals, we will try the second potential donor value on the list and so on until we find a donor value that is allowed according to the edits and totals.

4.3.2 Random hot deck imputation

In our application of random hot deck imputation, we construct a list of potential donor records for each record with missing values by randomly drawing (without replacement) potential donor records, until all potential donor records have been drawn and put on the list for this recipient record. Note that since we construct a list of potential donor records for each recipient record, for each variable with a missing value in this recipient record the potential donor records are in the same order.

To impute a missing value in a certain recipient record, we follow the same procedure as for nearest-neighbour imputation. Note that again, if possible, multivariate imputation using several values from the first potential donor record on this list will be used.

4.4 The imputation algorithm

We now explain how we check whether a potential donor value for a certain record is allowed according to the edits and totals.

We first examine the case where all survey weights are equal. When we want to impute a missing value for variable x_j in a record i_0 we apply the following procedure.

0. Set $t := 1$.
1. Select the t-th observed value on the list of potential donor values obtained from one of the hot deck methods described in Section 4.3.
2. We check whether this value lies in the admissible interval for $x_{i_0 j}$. If so, we continue with Step 3. Otherwise, we set $t := t + 1$ and go to Step 4.
3. We check whether the potential donor value would enable us to preserve the total for variable x_j. If so, we use this potential donor value to impute the missing value. Otherwise, we set $t := t + 1$ and go to Step 4.
4. If t does not exceed the number of potential donor values for variable x_j, go to Step 1. Otherwise we impute the first potential donor value.

We can efficiently combine the checks in Steps 2 and 3. The check in Step 2 is simply whether $l_{i_0 j} \leq x_{i_0 j}^d \leq u_{i_0 j}$, where x_{ij}^d is the potential donor value drawn in Step 1, $l_{i_0 j}$ is the lower bound according to the edits for variable x_j in record i_0 and $u_{i_0 j}$ the corresponding upper bound (see Section 4.2). The check in Step 3 amounts to checking whether

$$\sum_{\substack{i\in M(j)\\ i>i_0}} l_{ij} \leq X_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} \hat{x}_{ij} - x^d_{i_0 j} \leq \sum_{\substack{i\in M(j)\\ i>i_0}} u_{ij}, \tag{6}$$

where $M(j)$ is the set of records with missing values for variable x_j, $\hat{x}_{ij}$ $(i \in M(j), i < i_0)$ are the already imputed values, and $X_{j,imp}$ is the total to be imputed for variable x_j. This total to be imputed equals the total X_j^{pop} minus the sum of the observed values for variable x_j.

In words, (6) simply says that the remaining total to be imputed for variable x_j should lie between the sum of the lower bounds for the remaining records to be imputed and the corresponding sum of upper bounds. That this check is necessary and sufficient in order to be able to preserve the total follows from the observation that the sum of the lower bounds for the remaining records to be imputed is the minimum amount that has to be imputed, and the corresponding sum of upper bounds is the maximum that can be imputed.

Check (6) can be rewritten as

$$\left(X_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} \hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} u_{ij} \right) \leq x^d_{i_0 j} \leq \left(X_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} \hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} l_{ij} \right)$$

The checks in Steps 2 and 3 can be combined into one check:

$$\max\left(X_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} \hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} u_{ij}, l_{i_0 j} \right) \leq x^d_{i_0 j} \leq \min\left(X_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} \hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} l_{ij}, u_{i_0 j} \right) \tag{7}$$

Equation (7) is our check for unweighted totals.

We can easily extend this to the case of unequal sampling weights w_i for each record i

$$\max\left(X^w_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} w_i\hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} w_i u_{ij}\,, w_{i_0} l_{i_0 j}\right) \le w_{i_0} x^d_{i_0 j} \le \min\left(X^w_{j,imp} - \sum_{\substack{i\in M(j)\\ i<i_0}} w_i\hat{x}_{ij} - \sum_{\substack{i\in M(j)\\ i>i_0}} w_i u_{ij}\,, w_{i_0} u_{i_0 j}\right) \quad (8)$$

Here $X^w_{j,imp} = X^{pop}_j - \sum_{i\in Obs(j)} w_i x_{ij}$. Equation (8) is our check for weighted totals.

If t exceeded the number of potential donor values for variable x_j in Step 4 of the algorithm, we adjust the imputed value by changing it to the closest boundary of (7) (unweighted) or (8) (weighted) totals.

Note that (7) and (8) imply that the value to be imputed lastly for a certain variable will be equal to the known total minus the sum (unweighted in the case of (7) and weighted in the case of (8)) of the observed and imputed values of this variable for the other records. The lastly imputed value is guaranteed to satisfy the edits. Given that the edits are sufficiently strict, they will offer some protection against the imputation of an unreasonable value for the lastly imputed value.

5. Evaluation Study

5.1 Methods Evaluated

We give results for 4 versions of our imputation methods: nearest-neighbour hot deck imputation preserving unweighted totals ("NN HD without weights"), nearest-neighbour hot deck imputation preserving weighted totals ("NN HD with weights"), random hot deck imputation preserving unweighted totals ("Random HD without weights") and random hot deck imputation preserving weighted totals ("Random HD with weights"). We have both an "unweighted" and a "weighted"

version of our imputation methods in order to study the effect of using survey weights on our evaluation measures.

In our evaluation study we have compared our imputation methods to the MI-SAS imputations that were described in Section 4.3.1 to calculate the distance functions. This represents a commonly used multivariate parametric imputation procedure. We also compare our methods to standard versions of random hot deck ("Standard Random HD") and nearest-neighbour imputation using distance function (6) ("Standard NN HD") and using imputations obtained from MI-SAS to calculate this distance function. For each of our proposed imputation methods, Standard NN HD and Standard Random HD we have produced only one imputed data set. The MI-SAS procedure, Standard NN HD and Standard Random HD do not take edits or known totals into account. In Standard NN HD and Standard Random HD we have, in principle, applied multivariate donor imputation, where the donor is either the nearest record or a randomly selected record. If a selected donor record did not have observed values for all missing values in a recipient record, additional donors for the remaining missing values were selected. We have applied this procedure because of the high number of missing values in evaluation data set 2 (see Section 5.2 below).

Comparing our imputation methods to MI-SAS, Standard NN HD and Standard Random HD enables us to compare our methods to commonly used standard imputation methods, and at the same time to some extent examine the effect of taking edits and totals into account.

5.2 Evaluation Data

For our evaluation study we have used a data set with observed data from an annual structural business survey of Statistics Netherlands from 2003. This data set contains survey weights that differ across different (strata of) records. We will refer to this data set as data set 1.

To test whether our imputation methods also produce imputations that satisfy edits and totals in exceedingly difficult cases, we have applied them to another, more complicated data set. This data set was synthetically generated and contained 500 records and 10 variables. The number of missing values was higher than typically observed in business surveys. We will refer to this data set as data set 2.

The main characteristics of the data sets are presented in Table 1.

Table 1. The characteristics of the evaluation data sets

	data set 1	data set 2
Total number of records	3,096	500
Number of records with missing values	544 (17.5%)	490 (98.0%)
Total number of variables	8	10
Total number of edits	14	16
Number of balance edits	1	3
Total number of inequality edits	13	13
Number of non-negativity edits	8	9

The actual values for the data in the two data sets are all known. In the completely observed data sets values were deleted by a third party, using a mechanism unknown to us. For each of our evaluation data sets we have two versions available: a version with missing values and a version with complete records. The former version is imputed. The resulting data set is then compared to the version with complete records, which we consider as a data set with the true values.

The numbers of missing values and (unweighted) basic statistics of the variables of our data sets are given in Tables 2 and 3. The percentages in brackets are the percentages of records with a missing

value for the corresponding variable out of the total number of 3,096 records for data set 1 and 500 records for data set 2. The basic statistics are taken over all observations in the complete versions of the data sets. Variable R_8 in data set 1 does not contain any missing values and is only used as auxiliary variable. In our evaluation study we have a sum constraint due to a known total for every variable in the data sets to be imputed.

We have not made an attempt to optimize the order in which the variables are imputed. We have simply imputed the variables in reverse order, i.e. for data set 1 we have imputed the missing values for variable R_7 first and the missing values for variable R_1 last, and similarly for data set 2.

Table 2. The numbers of missing values, mean, median, standard deviation and range in data set 1

Variable	Number of missing values	Mean	Median	Standard deviation	Range
R_1	76 (2.5%)	11,574.8	1997.5	51,747	[0 ; 1,264,082]
R_2	68 (2.2%)	777.6	242.0	1,723	[0 ; 60,563]
R_3	130 (4.2%)	8,978.7	1070.5	48,857	[0 ; 1,257,348]
R_4	147 (4.8%)	1,034.1	187.0	3,791	[0 ;152,814]
R_5	79 (2.6%)	10,012.8	1496.0	49,862	[0 ; 1,258,837]
R_6	73 (2.4%)	169.2	0.0	4,885	[0 ; 205,210]
R_7	67 (2.2%)	209.9	7.0	4,927	[0 ; 206,327]
R_8	0 (0.0 %)	37.4	14.0	58	[0 ; 689]

Table 3. The numbers of missing values, mean, median, standard deviation and range in data set 2

Variable	Number of missing values	Mean	Median	Standard deviation	Range
S_1	120 (24%)	97.8	95.0	39	[16 ; 223]
S_2	180 (36%)	175,018.3	159,045.5	101,787	[7,256 ; 550,332]
S_3	240 (48%)	731.0	657.5	503	[0 ; 3,386]
S_4	120 (24%)	175,749.3	159,627	101,847	[7,496 ; 550,856]
S_5	180 (36%)	154,286.5	140,321	98,117	[1,480 ; 525,490]
S_6	180 (36%)	7,522.3	7,369.0	3,263	[431 ; 1,8061]
S_7	180 (36%)	8,519.7	8,355.0	4,327	[112 ; 21,355]
S_8	180 (36%)	1,277.0	1,206.5	785	[0 ; 3,745]
S_9	120 (24%)	171,605.6	156,868.5	101,435	[10,481 ; 547,168]
S_{10}	120 (24%)	4,143.8	4,126	3,240	[-5,911 ; 13,563]

As explained in Section 4.1, variables with missing values involved in balance edits will only rarely all be imputed by means of multivariate imputation. In data set 1 there are 3 variables involved in one balance edit, and in data set 2 there are in total 9 variables involved in 3 balance edits. This means that our imputation methods will switch to sequential imputation instead of multivariate imputation quite regularly. In data set 1, 49% of the records with missing values were imputed by means of multivariate imputation and for the remaining 51% our imputation methods switched to sequential imputation. In data set 2, where almost all variables are involved in one or more balance edits, 99% of the records with missing values were switched to sequential imputation and only 1% were imputed by means of multivariate imputation.

5.3 Evaluation measures

To measure the performance of our imputation approaches we use a d_{L1} measure, an m_1 measure, an rdm measure, and the Kolmogorov-Smirnov distance (KS). The first two criteria and the

Kolmogorov-Smirnov distance have been proposed by Chambers (2003). The d_{L1} measure is the average distance between the imputed and true values defined as

$$d_{L1} = \frac{\sum_{i \in M(j)} w_i |\hat{x}_{ij} - x_{ij}^{\text{true}}|}{\sum_{i \in M(j)} w_i}$$

where from Section 4.4, $\hat{x}_{ij}$ is the imputed value in record i of the variable x_j under consideration and x_{ij}^{true} the corresponding true value. *M(j)* is the set of records where the value of variable x_j is missing.

The m_1 measure, which measures the preservation of the first moment of the empirical distribution of the true values, is defined as

$$m_1 = \left| \frac{\sum_{i \in M(j)} w_i (\hat{x}_{ij} - x_{ij}^{\text{true}})}{\sum_{i \in M(j)} w_i} \right|$$

The rdm (relative difference in means) measure has been used in an evaluation study by Pannekoek and De Waal (2005), and is defined as

$$rdm = \frac{\sum_{i \in M(j)} w_i \hat{x}_{ij} - \sum_{i \in M(j)} w_i x_{ij}^{\text{true}}}{\sum_{i \in M(j)} w_i x_{ij}^{\text{true}}}$$

The rdm measure is the weighted bias due to imputation.

Finally, we use the KS Kolmogorov-Smirnov distance to compare the empirical distribution of the original values to the empirical distribution of the imputed values. For weighted data, the empirical distribution of the true values is defined as

$$F_{x_j}(t) = \sum_{i \in M(j)} I(w_i x_{ij} \leq t)/|M(j)|$$

with $|M(j)|$ the number of records with missing values for variable x_j and I the indicator function. Similarly, we define $F_{\hat{x}_j}(t)$. The KS distance is defined as

$$KS = \max_k |F_{x_j}(t_k) - F_{\hat{x}_j}(t_k)|,$$

where the t_k values are the $2|M(j)|$ jointly ordered true and imputed values.

We also evaluate how well the imputation measures preserve medians. For this we use the percent relative absolute difference defined by

$$PD(X) = 100 \times \frac{|\phi_{\text{orig}} - \phi_{\text{imp}}|}{\phi_{\text{orig}}}$$

where ϕ denotes the median of the variable under consideration, ϕ_{orig} its value in the original complete data set calculated using survey weights, and ϕ_{imp} its value in the imputed data set for the imputation method under consideration again calculated using survey weights.

Smaller absolute values of the evaluation measures indicate better imputation performance.

5.4 Evaluation results

The evaluation results for data set 1 are presented in Tables 4 to 8. As variable R_8 does not have any missing values in data set 1, no evaluation results for R_8 are presented in the tables. The column "Average" in Tables 4 to 7 is the average of the absolute results over all 7 variables, however in Table 8 "Average" is over 6 variables not including R_6 which has a median value of 0.

Table 4. Results for the d_{L1} measure for data set 1

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Average
NN HD without weights	2,607	215	115	171	154	19	13	98
NN HD with weights	2,236	125	120	110	16	3	19	56
Random HD without weights	6,147	404	121	181	167	17	34	132
Random HD with weights	3,374	199	117	111	50	3	17	71
MI-SAS	1,073	230	360	347	21	17	26	143
Standard NN HD	438	81	232	266	234	2	16	181
Standard Random HD	25,199	819	4,454	587	4,703	7	49	5,117

Table 5. Results for the m_1 measure for data set 1

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Average
NN HD without weights	288	74	11	72	132	16	2	44
NN HD with weights	0	0	0	0	0	0	0	0
Random HD without weights	3,138	240	6	73	143	14	21	71
Random HD with weights	0	0	0	0	0	0	0	0
MI-SAS	127	84	123	126	21	17	12	55
Standard NN HD	72	5	122	186	14	1	5	58
Standard Random HD	22,691	622	2,366	412	2,172	4	21	4,041

Table 6. Results for the rdm measure for data set 1

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Average†
NN HD without weights	-13.00	0.44	-0.01	0.44	0.08	12.96	-0.12	3.86
NN HD with weights	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Random HD without weights	1.47	1.41	0.00	0.45	0.08	11.36	1.25	2.29
Random HD with weights	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MI-SAS	-0.06	-0.49	-0.08	0.78	0.01	13.10	0.69	2.17
Standard NN HD	-0.03	0.03	-0.08	1.15	0.01	-0.72	-0.29	0.33
Standard Random HD	10.62	3.65	1.59	2.55	1.24	3.44	1.23	3.48

† "Average" in this table denotes the average over the absolute values

Table 7. Results for the KS measure for data set 1

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Average
NN HD without weights	0.72	0.20	0.03	0.32	0.03	0.21	0.40	0.17
NN HD with weights	0.46	0.20	0.05	0.40	0.03	0.09	0.51	0.18
Random HD without weights	0.21	0.25	0.05	0.39	0.03	0.25	0.22	0.17
Random HD with weights	0.33	0.38	0.05	0.41	0.04	0.09	0.42	0.20
MI-SAS	0.16	0.25	0.12	0.17	0.03	0.13	0.12	0.12
Standard NN HD	0.02	0.07	0.02	0.11	0.06	0.92	0.16	0.19
Standard Random HD	0.52	0.26	0.42	0.18	0.54	0.92	0.28	0.44

Table 8. Results for medians for data set 1 ("*" denotes that the median in the original, complete data set is zero, and hence the percent difference is undefined)

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Average[♯]
NN HD without weights	2.41	0.67	0.04	1.19	0.00	*	4.30	1.03
NN HD with weights	1.41	0.52	0.04	1.80	0.28	*	4.31	1.16
Random HD without weights	0.75	1.14	0.00	1.74	0.13	*	1.85	0.81
Random HD with weights	1.41	2.00	0.00	2.31	0.28	*	4.31	1.48
MI-SAS	0.00	0.03	0.00	1.86	0.13	*	0.41	0.41
Standard NN HD	0.23	0.00	0.00	4.28	0.00	*	0.00	0.75
Standard Random HD	0.98	0.83	0.61	1.60	0.80	*	0.00	0.80

[♯] Average taken over 6 variables.

NN HD with weights performs especially well with respect to evaluation measures d_{L1} (Table 4), m_1 (Table 5) and *rdm* (Table 6), which all measure how well totals and individual values are preserved. MI-SAS performs especially well for evaluation measure *KS* (Table 7) and the preservation of the medians (Table 8), which both measure how well the statistical distribution is preserved. Standard Random HD performs worst for measures d_{L1}, m_1 and *KS* (Tables 4, 5 and 7), indicating that this method is not good in preserving either individual values or the statistical distribution of data set 1.

The bad results of Standard Random HD are partly caused by outliers in the data. For instance, in one record with a relatively large survey weight a very large value is imputed for variable R_1, whereas the true value is small. Excluding this outlier, d_{L1} for R_1 drops to 5,370, m_1 to 2,862 and *rdm* to 1.34. These numbers are still quite large, but better than Random HD without weights.

Comparing the unweighted versions of our imputation methods to the weighted versions with respect to evaluation measures d_{L1} (Table 4), *KS* (Table 7) and preservation of the median (Table 8), the

unweighted versions perform worse for d_{L1}, but perform better on *KS* and the preservation of the median. The weighted versions by design perform better on m_1 and *rdm* (Tables 5 and 6).

Comparing NN HD with weights and Random HD with weights to Standard NN HD and Standard Random HD, we see that taking edits and known totals into account leads to better preservation of totals and individual values (Tables 4 to 6). NN HD with weights and Random HD with weights also appear to perform better with respect to the *KS* measure, but this is partly caused by outliers in the data. For instance, the large values for the *KS* measure for variable R_6 for Standard NN HD and Standard Random HD are caused by a missing field that has an unusually large value in the true data. Variable R_6 equals zero in most records, and both Standard NN HD and Standard Random HD impute zero in the corresponding record. Medians (Table 8) seem to be better preserved by the Standard NN HD and Standard Random HD than by our proposed methods.

The general conclusion we can draw is that standard multivariate hot deck imputation appears to preserves distributional aspects better, whereas our sequential imputation approach preserves point estimates better. MI-SAS performs best with respect to distributional aspects. With respect to point estimates MI-SAS performs slightly better than Standard NN HD (except for *rdm*, see Table 6) and much better than Standard Random HD. It performs worse than the weighted versions of our imputation methods with respect to point estimates.

For the more complex data set 2, the statistical distribution was less well preserved by our imputation methods than by MI-SAS. This can be seen in Table 9 where we compare NN HD with weights to MI-SAS. The numbers in this table are the unweighted averages over all 10 variables in the data set. For evaluation measures m_1 and *rdm*, NN HD with weights by design performs better than MI-SAS for all variables. For MI-SAS, m_1 ranged from 4.5 to 1860.6, and the absolute value of *rdm* from 0.005 to 0.058. NN HD with weights has m_1 and *rdm* equal to zero for all variables.

Table 9. Comparison of NN HD with weights to MI-SAS for data set 2

	NN HD with weights	MI-SAS
d_{L1}	1,930	1261
m_1	0.0	637.5
rdm	0.00	0.01
KS	0.10	0.06
median	17.25	11.43

In Appendix B we examine the preservation of correlations.

Our imputation methods have been designed to satisfy edits and preserve known (weighted) totals. As can be seen from the evaluation measures m_1 and *rdm* in Tables 5 and 6 the weighted versions of our imputation indeed succeed in taking the weighted totals into account. Our imputation methods also succeed in satisfying edits. Of the two data sets examined, none of the records had violated edits.

Besides violating known totals (and hence means), which can be seen from evaluation measures m_1 and *rdm* in Tables 5 and 6, MI-SAS, Standard NN HD and Standard Random HD also violate edits. The number of violated edits and violated records, i.e. records in which at least one edit is violated, is given in Table 10. In this table we see that Standard NN HD and Standard Random HD violate a relatively large number of edits and records. MI-SAS performs better in this respect as the multivariate regression ensures that balance restrictions are satisfied after imputation. Still 2.5% of the records of data set 1 and 5% of the records of data set 2 are violated when MI-SAS is used.

Table 10. Numbers of violated edit rules and records in both data sets

	data set 1		data set 2	
	violated edits	violated records	violated edits	violated records
NN HD without weights	0	0	0	0
NN HD with weights	0	0	0	0
Random HD without weights	0	0	0	0
Random HD with weights	0	0	0	0
MI-SAS	123	77 (2.5%)	25	25 (5.0%)
Standard NN HD	313	267 (8.6%)	1,125	485 (97.0%)
Standard Random HD	419	331 (10.7%)	1,256	488 (97.6%)

6. Discussion

In this paper we have extended standard hot deck imputation methods so that the imputed data satisfy edits and preserve known totals, while taking survey weights into account. The hot deck imputation methods we have considered are random hot deck and nearest neighbour hot deck. To ensure that edits are satisfied and known totals are preserved after imputation, we have applied these hot deck imputation methods in a sequential manner and have used a check based on Fourier-Motzkin elimination for determining admissible intervals for each value to be imputed. Strong aspects of our imputation methods are their simplicity and that they are non-parametric, which make them attractive from a practical point of view. For data sets that are approximately normally, or log-normally, distributed we may opt to use model-based imputation methods developed by Pannekoek, Shlomo and De Waal (2013).

In our evaluation study we have used two evaluation data sets. More evaluation studies on other data sets are required before firm conclusions can be drawn. The results of our evaluation study are indicative that our imputation methods may give acceptable results. Obviously, our imputation

methods give perfect results with respect to preservation of means and totals as they have been designed to do so according to the 'one-figure' policy.

NN hot deck with weights in particular gives good results for most of the evaluation measures we have examined, while satisfying edits and preserving known totals at the same time. An exception is the preservation of the median. NN hot deck with weights gives only mediocre results in that respect, and is outperformed by the MI-SAS imputation. Future work will investigate how the imputations can be controlled in such a way as to have a better preservation of the median as well as the mean.

Taking edits and known totals into account while imputing missing data improves the preservation of individual values and, obviously, of means and totals. However, taking edits and known totals into account may lead to a deterioration of the preservation of the joint statistical distribution. A practical advantage of taking edits into account is that this offers some protection against outliers in the data. Some results for Standard NN HD and Standard Random HD were quite bad due to such outliers. In our imputation methods, the edit rules often will not allow clearly outlying values to be imputed, simply since this would lead to edit violations.

In Section 4.4 we noted that the value to be imputed lastly for a certain variable will be equal to the known total minus the sum (unweighted in the case of (7) and weighted in the case of (8)) of the observed and imputed values for this variable for the other records. If this would lead to an unreasonable value for the lastly imputed value, one could try some simple approaches to remedy this, such as adding some additional edits or subdividing the data into several groups, each with an estimated group total, and impute these groups separately. As the group totals are smaller than the overall total, an unreasonable value for the variable to be imputed lastly is less likely to occur.

As mentioned in Section 5.2 we have not made an attempt to optimize the order in which the variables are imputed. As the results of our imputation methods depend on this order, optimizing the order is an

interesting research topic for future research. As a practical guideline, we suggest to impute the variables that are considered to be important first and the less important variables last.

In this paper we have not considered the issue of estimating the correct variance after imputation, including the variance due to nonresponse and imputation. In addition, if estimated totals are used for the benchmarking instead of known totals, there will be an additional component to be added to the variance. In order to estimate variance correctly after imputation we can consider the extension of our approach to multiple imputation (see Rubin 1987) or use resampling techniques such as the bootstrap or jackknife adapted to account for the fact that were dealing with a finite population (see e.g. ~~Efron and Tibshirani 1979~~Mashreghi, Haziza and Léger 2016).

A drawback of a sequential imputation method is that optimal choices for individual variables to be imputed may not lead to overall optimality for all variables. An imputation method that imputes all variables in each record simultaneously while taking edits, totals and survey weights into account would be preferable to sequential imputation. Further research is required to develop such simultaneous imputation methods that are computationally tractable and easy to apply in practical situations.

References

Andridge R.R. and R.J.A. Little (2010). A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review 78*, pp. 40-64.

Beaumont, J.-F. (2005), Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *Journal of the Statistical Society B 67*, pp. 445-458.

Chambers, R. (2003), Evaluation Criteria for Statistical Editing and Imputation. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (available on http://www.cs.york.uk/euredit/).

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

Coutinho, W., T. de Waal and M. Remmerswaal (2011), Imputation of Numerical Data under Linear Edit Restrictions. *Statistics and Operations Research Transactions 35*, pp. 39-62.

Coutinho, W., T. de Waal and N. Shlomo (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. *Journal of Official Statistics 29*, pp. 299-321.

Cox, B.G. (1980), *The Weighted Sequential Hot Deck Imputation Procedure*. ASA SRMS Proceedings, pp. 721-726.

De Waal, T., J. Pannekoek and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New York.

Duffin, R.J. (1974), On Fourier's Analysis of Linear Inequality Systems. *Mathematical Programming Studies 1*, pp. 71-95.

~~Efron, B. and R.J. Tibshirani (1979), An Introduction to the Bootstrap. Chapman & Hall/CRC, Boka Raton.~~

Geweke, J. (1991), *Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities*. Report, University of Minnesota.

Holan, S.H., D. Toth, M.A.R. Ferreira and A.F. Karr (2010), Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality, *Journal of the American Statistical Association 105*, pp. 564-577.

Kalton, G. and D. Kasprzyk (1986), The Treatment of Missing Survey Data. *Survey Methodology 12*, pp. 1-16.

Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox and A.F. Karr (2014), Multiple Imputation of Missing or Faulty Values under Linear Constraints. *Journal of Business and Economic Statistics 32*, pp. 375-386

Kovar, J. and P. Whitridge (1995), Imputation of Business Survey Data. In: *Business Survey Methods* (ed. Cox, Binder, Chinnappa, Christianson & Kott), John Wiley & Sons, New York, pp. 403-423.

Little, R.J.A. (1988), Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics 6*, pp. 287-296.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data (second edition).* John Wiley & Sons, New York.

Longford, N.T. (2005), *Missing Data and Small-Area Estimation*. Springer-Verlag, New York.

Mashreghi, Z., ~~,~~ D. Haziza~~, D.~~ and C. Léger~~, C.~~ (2016), A Survey of Bootstrap Methods in Finite Population Sampling, *Statistics Surveys 10,* pp. 1-52.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk and P. Solenberger (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology 27*, pp. 85-95.

Särndal, C-E. and S. Lundström (2005), *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Chichester.

Särndal, C.E., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAS (2015), *SAS/STAT 14.1 User's Guide The MI Procedure*. Cary, NC, USA.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Statistics Netherlands (2014), *Dutch Census 2011: Analysis and Methodology*. Report, Statistics Netherlands.

Tempelman, C. (2007), *Imputation of Restricted Data*. Doctorate thesis, University of Groningen.

Van Buuren, S. (2012), *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, Florida.

Winkler, W.E. (2003), *A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. Research Report Series 2003-07, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Winkler, W.E. (2008), *General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints*. Research Report Series 2008-08, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Zhang, L.C. (2009), *A Triple-Goal Imputation Method for Statistical Registers*. UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Zhang, L.C. and S. Nordbotten (2008), *Prediction and Imputation in ISEE: Tools for More Efficient Use of Combined Data Sources*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Appendix A

We illustrate how a sequential approach can be used with a simple example taken from Coutinho, De Waal and Remmerswaal (2011). We consider a case where we have $|r|$ records with four variables, T (turnover), P (profit), C (costs), and N (number of employees in fulltime equivalents). The edits are given by:

$$T_i - C_i - P_i = 0 \quad \text{(A.1)}$$

$$T_i \geq 0 \quad \text{(A.2)}$$

$$P_i \leq 0.5T_i \quad \text{(A.3)}$$

$$-0.1T_i \leq P_i \quad \text{(A.4)}$$

$$T_i \leq 550N_i \quad \text{(A.5)}$$

$$N_i \geq 0 \quad \text{(A.6)}$$

$$C_i \geq 0 \quad \text{(A.7)}$$

We impute the variables in the following order: N, T, C and P. We assume that variable N has already been imputed and that we now want to impute variable T.

Suppose that in a certain record i_0 $N = 5$. This value may either have been observed or been imputed before. The values of of T, C and P in record i_0 are missing. We eliminate T, C and P in reverse order of imputation. We first fill in N into the edits. This gives us the edit set (A.1), (A.2), (A.3), (A.4), (A.7) and

$$T_{i_0} \leq 2{,}750 \quad \text{(A.8)}$$

We use equation (A.1) to express P in terms of T and C, and use that expression to eliminate variable P from edits (A.2), (A.3), (A.4), (A.7) and (A.8). This gives us the edit set (A.2), (A.7), (A.8),

$$T_i - C_i \leq 0.5T_i \quad \text{(A.9)}$$

and

$$-0.1T_i \leq T_i - C_i \qquad \text{(A.10)}$$

To eliminate variable C from the edit set (A.2), (A.7), (A.8), (A.9) and (A.10), we first copy the edits not involving C (edits (A.2) and (A.8)) and then eliminate C from the other edits. Eliminating C from (A.7), (A.9) and (A.10) gives us edits that are equivalent to (A.2). So, the edit set after elimination of C is given by (A.2) and (A.8). The admissible interval for T for record i_0 is hence given by

$$0 \leq T_{i_0} \leq 2{,}750.$$

Similarly, we can derive admissible intervals for T_i for all records i $(i = 1, \ldots, |r|)$ in which the value of T is missing. Once we have derived these admissible intervals, we impute values for T_i in all these records by means of one of our sequential imputation algorithms (see Sections 4.3 and 4.4).

After variable T has been imputed in all records in which its value was missing, we derive admissible intervals for variable C, and later variable P, in a similar manner. The main property of Fourier-Motzkin elimination guarantees that the original edits will be satisfied, if we select donor values lying inside these admissible intervals.

Appendix B

In Table B.1 we examine the effect of taking edits and known totals into account on (weighted) correlations between variables. In this table we give the average absolute deviation of the correlations in the imputed data from the correlations in the true data taken over all pairs of different variables for NN HD with weights and Standard NN HD. That is, for data set 1 we take the average absolute deviation of the correlations over 28 pairs of different variables and for data set 2 over 45 pairs of different variables. Between brackets we give the average of the absolute percent differences between the correlations in the true data and in the imputed data.

Table B.1. Average absolute deviation from true correlations

	Data set 1	Data set 2
NN HD without weights	0.0047 (1.19%)	0.1090 (42.77%)
NN HD with weights	0.0037 (0.90%)	0.0861 (26.09%)
Random HD without weights	0.0072 (0.66%)	0.1083 (35.63%)
Random HD with weights	0.0044 (1.13%)	0.0866 (26.84%)
SAS-MI	0.0007 (0.20%)	0.0142 (13.80%)
Standard NN HD	0.0015 (0.45%)	0.0167 (9.49%)
Standard Random HD	0.0528 (10.58%)	0.2151 (51.78%)

From Table B.1 we conclude that MI-SAS performs best with respect to preservations of correlations due to the multivariate imputation procedure under the MCMC approach. For data set 1 NN HD without weights, NN HD with weights, Random HD without weights, Random HD with weights, and Standard NN HD also give good results for practical purposes. The correlations after imputation are very close to the original correlations for these methods. For data set 2 all imputation methods, even SAS-MI, perform quite badly.