Discussion Paper

Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2015 | 20

Ton de Waal
Wieger Coutinho
Natalie Shlomo

Content

Abstract

A common problem faced by statistical institutes is that some data of otherwise responding units may be missing. This is referred to as item non-response. Item-nonresponse is usually treated by imputing the missing data. The problem of imputing missing data is complicated by the fact that statistical data often have to satisfy so-called edit rules, which for numerical data usually take the form of linear restrictions. A further complication is that numerical data sometimes have to sum up to known totals. Standard imputation methods for numerical data as described in the literature generally do not take such linear edit restrictions on the data or known totals into account. In this paper we develop simple imputation methods that satisfy edits and preserve known totals. These methods are based on well-known hot deck approaches. Extension of our methods to other types of imputation, such as regression imputation or predictive mean matching, is straightforward.

1. Introduction

Missing data form a well-known problem that has to be faced by basically everyone who collects data on persons or enterprises. Missing data can arise due to unit non-response or item-nonresponse. Unit non-response occurs when units that are selected for data collection cannot be contacted, refuse 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, Knottnerus 2003, 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 giving 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 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), 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 (NSIs), data have to satisfy constraints in the form of edit restrictions, or edits for short. These edit restrictions complicate 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) and by Coutinho, De Waal and Shlomo (2013).

A further complication is that numerical data sometimes have to sum up to known totals. This situation can arise when a one-figure policy ("one figure for one phenomenon"), where one aims to publish only one estimate for the same phenomenon occurring in different tables, is pursued. At Statistics Netherlands, we, for instance, pursue a one-figure policy for the Dutch Census, as well as for many other statistics. When a one-figure policy is pursued imputations need to be calibrated on this previously estimated total (see also Section 2 of the current paper).

Pannekoek, Shlomo and De Waal (2013) have developed imputation methods for numerical data that ensure that edits are satisfied and already estimated or known totals are preserved. A drawback of these methods is that they are relatively complex and hard to implement.

Our goal in this paper is to develop simple imputation methods that achieve the same objectives namely satisfied edits and preserved totals. In order to keep our imputation methods in this paper as simple as possible we base them on well-known hot deck approaches, rather than on more complicated imputation techniques, such as regression imputation with a random residual or predictive mean matching (see, e.g., De Waal, Pannekoek and Scholtus 2011 for more about these techniques). Modifying the imputation approach developed in this paper to such imputation techniques is quite straightforward, however.

As in Pannekoek, Shlomo and De Waal (2013) the main objective of our imputation methods is to obtain accurate point estimates, rather than preserve variances and correlations.

The remainder of this paper is organized as follows. Section 2 describes why calibrated imputation could be beneficial for producers of statistical data. 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. With “sequential” we mean that the variables are imputed one by one in all records in which the value of the variable under consideration is missing. Section 5 describes an evaluation study and its results. Finally, Section 6 concludes with a brief discussion.

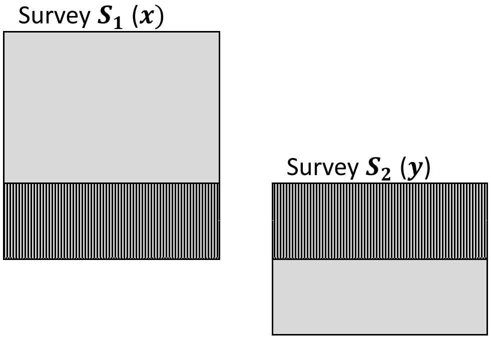
2. Why calibrated imputation?

We will start by illustrating why it is sometimes useful to calibrate estimates to known or previously estimated totals. 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 2.1.1 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.

In such a case, one could first use the data in survey S_1 to estimate the population total of x. When 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 like we propose in the current paper.

2.1.1 Two surveys


Survey $\boldsymbol{S_1}$ $(\boldsymbol{x})$
Survey $\boldsymbol{S_2}$ $(\boldsymbol{y})$

Below we argue why it could be beneficial to consider calibrated imputation instead of calibrated weighting. This is for a substantial part based on Pannekoek, Shlomo and De Waal (2013).

In the usual sample survey setting units are selected from a population according to a specified sampling design, where each population unit is included in the sample s with a certain non-zero probability. Estimates of population totals and other parameters of interest are then calculated by using sampling weights w_i that are the inverse of the inclusion probabilities. The total X_j of a variable x_j is estimated by

$$\hat{X}_j = \sum_{i \in s} w_i x_{ij}, \qquad (1)$$

with x_{ij} the value of variable x_j for unit i. In the simplest case where each population unit has the same sample probability $|s|/N$, with N the population size and $|s|$ the size of the sample selected, this weighting estimator is simply

$$\hat{X}_j = \sum_{i \in s} \frac{N}{|s|} x_{ij}. \qquad (2)$$

In practice, due to unit non-response, data are often only obtained for a subset r of the intended sample units, with an effective sample size $|r| < |s|$. The standard way to correct (1) for unit non-response is by inflating the weights w_i $(i \in r)$. For instance, a simple correction for (2) is by inflating the weights $N/|s|$ by the inverse of the non-response fraction,$|s|/|r|$. The weights are then simply given by $N/|r|$, i.e. the size of the population divided by the effective sample size.

If for some variables the population totals are known, the weights can be calibrated such that the estimated totals for these variables equal their known values. Such weights are generally not equal for all units (see e.g. Särndal and Lundström 2005). If an estimated total (1) for a certain variable is below the known population value, apparently not enough units with a high value on this variable were selected in the sample and too many units with a low value. Calibration weighting corrects for this unbalanced selection of units by increasing the weights w_i $(i \in r)$ for units with high values for that variable and decreasing the weights for units with low values. Note that correction for the unbalanced sample by calibrating the weights will affect estimates for all variables. This is not considered a problem as apparently the sample was unbalanced. For large samples and small unit non-response fractions calibration should have only minor effects on the weights.

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 units in r so that for the $|r|$ (partially) responding units a complete data set is obtained. Next, the $|r|$ units in this data set 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 population totals.

However, for variables with imputed values, differences between estimated totals and their known values are now not only caused by an unbalanced sample selection, but also by systematic errors in the imputed values (imputation bias). For large sample sizes and small unit non-response fractions the difference between estimated and known population totals will be mainly due to imputation bias as the total of imputed values may be substantial and calibration weights will be close to one. The weight adjustments due to calibration hence do not correct for an unbalanced selection of units but (mainly) for imputation bias in specific variables. There is no compelling reason to let this adjustment affect the estimates of all other variables. This makes calibration weighting less desirable in the case of item-nonresponse.

In this paper we therefore develop a different approach, where we leave the weights of the responding units unchanged so there will not be any unwanted effects due to adjusting these weights. Instead of calibrating the weights we will calibrate the imputations so weighted estimates will equal known population totals.

If one wants to avoid unwanted effects due to adjusting weights, our approach where we calibrate the imputations rather than the weights seems to be the most logical approach. As we already mentioned in the Introduction, our goal is to develop simple calibrated imputation methods that are easy to implement and maintain.

3. Constraints on the Imputed Data

3.1 Edit restrictions within records

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 $(k = 1, \dots, K)$ can then be written in either of the two following forms:

$$a_{1k}x_{i1} + \cdots + a_{pk}x_{ip} + b_k = 0 \tag{3a}$$

or

$$a_{1k}x_{i1} + \cdots + a_{pk}x_{ip} + b_k \geq 0 \tag{3b}$$

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 (3a) are referred to as balance edits. An example of such an edit is

$$T_i - C_i - P_i = 0 \tag{4}$$

where T_i is the turnover of an enterprise corresponding to the i-th record $(i = 1, \dots, |r|)$, P_i its profit, and C_i its costs. Edit (4) expresses that the profit of an enterprise equals its turnover minus its costs. A record not satisfying this edit is obviously incorrect.
Edits of type (3b) are referred to as inequality edits. An example is

$$T_i \geq 0 \tag{5}$$

expressing that the turnover of an enterprise should be non-negative.
Sum constraints across records
For equal survey weights, the sum constraints due to known or previously estimated population totals on the estimates can be expressed as

$$\sum_{i \in r} \frac{N}{|r|} x_{ij} = X_j^{pop}$$

with X_j^{pop} the known population total. In terms of the unweighted sample totals, these constraints then imply:

$$\sum_{i \in r} x_{ij} = \frac{|r|}{N} X_j^{pop} \equiv X_j$$

say.

In general, survey weights will often be unequal and sum constraints are given by

$$\sum_{i \in r} w_i x_{ij} = X_j^{pop} \tag{6}$$

with w_i the survey weights. In this paper we will consider general sum constraints of type (6).

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.

Usually, hot deck imputation is applied multivariately, that is several missing values in a recipient record are imputed simultaneously, using the same donor record. For our problem that approach is often not feasible. If an imputed record failed the edits, all one could do in such an approach were to reject the donor record and try to use 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 a different donor may be used.

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. Only potential donor values that can result in a record that satisfies all edits and preserves sum constraints, may be used for imputation. How we ensure that edits and sum constraints can be satisfied is explained in Section 4.2 below.

4.2 Using a Sequential Approach

In order to be able to use a sequential approach, 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

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

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.

Fourier-Motzkin elimination is closely related to the Fellegi-Holt method (see Fellegi and Holt 1976) for automatically detecting errors in a data set. A major difference is that in their article Fellegi and Holt mainly focus on categorical data instead of numerical data. Moreover, in our paper Fourier-Motzkin elimination is only used to impute the data in a manner consistent with the edits and totals, not to find any errors in the data. When used for automatic detection of errors, Fourier-Motzkin elimination and the related Fellegi-Holt approach can be very time-consuming to apply. As argued in Coutinho, De Waal and Remmerswaal (2011) and Coutinho, De Waal and Shlomo (2013), this is much less so for the case of imputation.

We illustrate how we apply Fourier-Motzkin elimination. 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:

$$l_{ij} \leq x_{ij} \leq u_{ij}, \tag{7}$$

where l_{ij} may be $-\infty$ and u_{ij} may be ∞.
We have such an interval constraint (7) 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 (7) are satisfied. For this we will use one of our sequential imputation algorithms as explained in Sections 4.3 and 4.4.

Example 1: To illustrate how a sequential approach can be used, 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 (4), (5),

$$P_i \leq 0.5T_i \tag{8}$$
$$-0.1T_i \leq P_i \tag{9}$$
$$T_i \leq 550N_i \tag{10}$$
$$N_i \geq 0 \tag{11}$$
$$C_i \geq 0 \tag{12}$$

We assume that the variables will be imputed in the following order: N, T, C and P. We also assume that variable N has already been imputed in all records in which its value was missing, and we are now ready to impute variable T.
Suppose that in a certain record i_0 we have $N = 5$ (this value may either have been observed or been imputed before), and the values of T, P and C are missing. We eliminate T, C and P in reverse order of imputation. We fill in the value for N into the edits and obtain the edit set (4), (5), (8), (9), (12) and

$$T_{i_0} \leq 2750. \tag{13}$$

To eliminate variable P, we use equation (4) to express P in terms of T and C. After elimination, we obtain the edit set (5), (12), (13)

$$T_{i_0} - C_{i_0} \leq 0.5T_{i_0} \tag{14}$$

and

$$-0.1T_{i_0} \leq T_{i_0} - C_{i_0} \tag{15}$$

To eliminate variable C from the current edit set (5), (12), (13), (14) and (15), we copy the edits not involving C (edits (5) and (13)) and eliminate C from the remaining edits. Eliminating C from (12), (14) and (15) leads to edits that are equivalent to (5). Hence the remaining edit set after elimination of C is given by (5) and (13), and so the admissible interval for T for record i_0 is given by

$$0 \leq T_{i_0} \leq 2750.$$

In a similar way we can derive admissible intervals for T_i for all records i $(i = 1, \dots, |r|)$ in which the value of T is missing. After we have done this, 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 can 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 in these admissible intervals.

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.

Nearest-neighbour hot deck imputation

Suppose we want to impute a certain variable x_j 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, that is the difference between the value of the 75% percentile of x_j and the 25% percentile of x_j.

In our evaluation study we have used three different distance functions.

$$d_1(\mathbf{x}_{i_o}^*, \mathbf{x}_i^*) = \sum_{j \in Obs} \gamma_j \, |x_{i_0 j}^* - x_{ij}^*| \tag{16}$$

$$d_2(\mathbf{x}_{i_o}^*, \mathbf{x}_i^*) = \sqrt{\sum_{j \in Obs} \gamma_j (x_{i_0 j}^* - x_{ij}^*)^2} \tag{17}$$

and

$$d_3\left(\mathbf{x}_{i_o}^*, \mathbf{x}_i^*\right) = \max\{j \in Obs: \gamma_j |x_{i_0 j}^* - x_{ij}^*|\} \tag{18}$$

where γ_j are nonnegative weights indicating how serious one considers a change in variable x_j to be, $\mathbf{x}_{i0}^*$ is the scaled recipient record and $\mathbf{x}_i^*$ a scaled potential donor record. In this paper we have set $\gamma_j = 1$ for all variables $(j = 1, \dots, |r|)$. Obs is the set of observed variables in the recipient record $\mathbf{x}_{i0}$.

If one or more values of variables in Obs of a potential donor record $\mathbf{x}_i$ are missing, we set the values of these variables equal to zero in both this potential donor record and the recipient record while calculating (16), (17), or (18). Note that this way of dealing with missing values in donor records deviates from the "conventional" way of dealing with such missing values, where donor pools are formed based on auxiliary variables that are observed for donors and recipients (see Andridge and Little 2010). This conventional way considers only donors where all so-called "matching variables", that is variables that one wants to use to match potential donor records to the recipient record, are observed. These records form a "donor pool" from which a donor record for the recipient record is subsequently drawn. If a donor pool is empty, some of the envisaged matching variables should not be used as matching variables in order to obtain a non-empty donor pool. For data sets containing a large number of missing values as in our evaluation data sets, constructing donor pools for each recipient can become quite cumbersome. As we aim to illustrate in this paper how the class of hot deck imputation techniques can be extended so that edits and totals are satisfied rather than develop "optimal" imputation methods for our data sets, we have chosen a simpler way to deal with missing values in donor records.

We construct an ordered list of potential donor values using distance function (16), (17) or (18). These potential donor values are ordered in increasing distance of the recipient record to the potential donor records. To impute a missing value, we will first select the first potential donor value of 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 an explanation of 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.

As a remark we note that if we had used the subset of variables that are observed for all records in (16), (17) or (18) instead of the set of all variables, the potential donor records for a certain recipient record would be ordered in the same way for each variable with missing values. In that case, if possible, multivariate imputation, using several values from the first potential donor record on this list, would be used. Only if a value of the first potential donor record could not be used because this were to lead to failing edits or a non-preserved total, a value from another potential donor record would be used. Since in our application in this paper we use the set of all variables to calculate (16), (17) and (18), the order of the records may differ per variable.

Random hot deck imputation

In our application of random hot deck imputation, we construct an ordered 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 we construct an ordered list of potential donor records for each recipient record, i.e. for each variable with a missing value in this recipient record we use the same ordered list.

To impute a missing value in a certain recipient record, we select the first donor record on the ordered list of donors for this recipient for which the corresponding value is observed. We then check whether that value is allowed according to the edits and totals (see Section 4.4). If so, we use it to impute this missing value. Otherwise, we select the next donor record on the ordered list of donors for this recipient for which the corresponding value is observed, and check whether that value is allowed according to the edits and totals and so on until we find a donor value that is allowed according to the edits and totals.

Note that, if possible, multivariate imputation using several values from the first potential donor record on this list, will be used. Only if a value from the first potential donor record is not observed or leads to failed edits or totals, we consider the other potential donor records on the ordered list.

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 the variable x_j under consideration in a certain record i_0 we basically apply the following procedure.

1. Set $t \coloneqq 1$.
2. 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.
3. 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 $t \coloneqq t + 1$ and return to Step 2.
4. 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 $t \coloneqq t + 1$ and go to Step 5.
5. If t does not exceed the number of potential donor values for the variable under consideration, go to Step 2. If t exceeds the number of potential donor values for the variable under consideration, we impute the value of the boundary of (20), respectively (21), below (depending on whether one wants to satisfy unweighted or weighted totals) that is closest to 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

$$\sum_{\substack{i \in M(j) \\ i > i_0}} l_{ij} \leq X_{j,imp} - \sum_{\substack{i \in M(j) \\ i < i_0}} \hat{x}_{ij} - x_{i_0 j}^d \leq \sum_{\substack{i \in M(j) \\ i > i_0}} u_{ij}, \tag{19}$$

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 minus the sum of the observed values for variable x_j.

In words, (19) 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 (19) can be rewritten as

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

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

$$\max\left(X_{j,imp} - \sum_{\substack{i \in M(j) \\ i < i_0}} \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} u_{ij}\, , l_{i_0 j} \right) \leq x_{i_0 j}^d \leq \min\left(X_{j,imp} - \sum_{\substack{i \in M(j) \\ i < i_0}} \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} l_{ij}\, , u_{i_0 j} \right) \tag{20}$$

(20) is our check for unweighted totals.

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

$$\max\left(X_{j,imp}^w - \sum_{\substack{i \in M(j) \\ i < i_0}} w_i \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} w_i u_{ij}\, , w_{i_0} l_{i_0 j} \right) \leq w_{i_0} x_{i_0 j}^d \leq \min\left(X_{j,imp}^w - \sum_{\substack{i \in M(j) \\ i < i_0}} w_i \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} w_i u_{ij}\, , w_{i_0} u_{i_0 j} \right) \tag{21}$$

Here $X_{j,imp}^w = X_j^{pop} - \sum_{i \in Obs(j)} w_i x_{ij}$. (21) is our check for weighted totals.

5. Evaluation Study

5.1 Evaluated methods

We have evaluated nearest-neighbour hot deck imputation for all three distance functions (16), (17) and (18). As the results are quite similar, we report only the results for distance function (17) in this paper. We give results for 4 versions of our imputation methods: nearest-neighbour hot deck imputation preserving unweighted totals (“NN hot deck without weights”), nearest-neighbour hot deck imputation preserving weighted totals (“NN hot deck with weights”), random hot deck imputation preserving unweighted totals (“random hot deck without weights”) and random hot deck imputation preserving weighted totals (“random hot deck with weights”). We have both an “unweighted” and a “weighted” version of our imputation methods in order to study the effect of using weights on our evaluation measures.

In our evaluation study we have compared our imputation methods to the standard multiple imputation routine implemented in SAS, which we will refer to as “MI-SAS” in this paper (see SAS 2015 for details on MI-SAS) and standard versions of nearest-neighbour imputation using distance function (17) (“standard NN”) and random hot deck (“standard random hot deck”). We have used MI-SAS with default options to produce 10 imputed data sets, and then pooled the results. The MI-SAS procedure and the standard versions of nearest-neighbour imputation and random hot deck do not take edits or known totals into account. In standard NN and standard random hot deck 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 one of our evaluation data sets (see Section 5.2 below).

Comparing our imputation methods to MI-SAS, standard NN and standard random hot deck enables us to compare our methods to a standard imputation method, 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 mainly used a data set with observed data from an annual structural business survey. 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, much more complicated data set. That data set was synthetically generated and contained 500 records and 10 variables. The number of missing values was much higher than in actually observed business surveys.

We will refer to that data set as data set 2.
The main characteristics of the data sets are presented in Table 5.2.1.

5.2.1 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	490
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) means of the variables of our data sets are given in Tables 5.2.2 and 5.2.3. 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 record for data set 2. The means are taken over all observations in the complete version of the data sets. Variable R8 in data set 1 does not contain any missing values and is only used as auxiliary variable. In our sequential imputation methods we have imputed the variables in reverse order: i.e. for data set 1 we have imputed the missing values for variable R7 first and the missing values for variable R1 last, and similarly for data set 2.
In our evaluation study we have a sum constraint due to a known total for every variable in the data sets to be imputed.

5.2.2 The numbers of missing values and the means in data set 1

Variable	Number of missing values	Mean
R1	76 (2.5%)	11,574.83
R2	68(2.2%)	777.56
R3	130 (4.2%)	8,978.70
R4	147 (4.8%)	1,034.07
R5	79 (2.6%)	10,012.77
R6	73 (2.4%)	169.24
R7	67 (2.2%)	209.86
R8	0 (0.0 %)	37.41

5.2.3 The numbers of missing values and the means in data set 2

Variable	Number of missing values	Mean
S1	120 (24%)	97.77
S2	180 (36%)	175,018.30
S3	240 (48%)	731.03
S4	120 (24%)	175,749.33
S5	180 (36%)	154,286.53
S6	180 (36%)	7,522.34
S7	180 (36%)	8,519.65
S8	180 (36%)	1,277.04
S9	120 (24%)	171,605.57
S10	120 (24%)	4,143.76

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 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 $\hat{x}_{ij}$ is the imputed value in record i of the variable x_j under consideration and x_{ij}^{true} the corresponding true value. Note that the value of d_{L1} depends on the variable. The same holds true for the evaluation measures mentioned below.
The m_1 measure, which measures the preservation of the first moment of the empirical distribution of the true values, is defined as

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

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

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

To remain consistent with the literature, in particular with the papers by Chambers (2003) and Pannekoek and De Waal (2005), we have not made an attempt to make the d_{L1} and the m_1 measures comparable across variables.

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 (also proposed by Chambers 2003). For unweighted data, the empirical distribution of the true values is defined as

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

with $|M(j)|$ the number of records with missing values for the variable at hand (i.e. the size of set $M(j)$), 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. Smaller absolute values of the evaluation measures indicate better imputation performance.

We also evaluate how well the imputation measures preserve medians and standard deviations. For this we use the percent difference defined by

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

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

5.4 Evaluation results

The evaluation results for data set 1 are presented in Tables 5.4.1 to 5.4.6. The “*” for variable R6 in Table 5.4.5 denotes that the median in the original, complete data set is zero, and hence that the percent difference is undefined. As variable R8 does not have any missing values, so no evaluation results for R8 are presented in the tables.

5.4.1 Results for the d_{L1} measure for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	2607	215	115	171	154	19	13
NN with weights	2236	125	120	110	16	3	19
Random hot deck without weights	6147	404	121	181	167	17	34
Random hot deck with weights	3374	199	117	111	50	3	17
MI-SAS	1073	230	360	347	21	17	26
Standard NN	4082	145	1186	175	1287	7	27
Standard random hot deck	1975	366	1480	201	1714	6	146

5.4.2 Results for the m_1 measure for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	288	74	11	72	132	16	2
NN with weights	0	0	0	0	0	0	0
Random hot deck without weights	3138	240	6	73	143	14	21
Random hot deck with weights	0	0	0	0	0	0	0
MI-SAS	127	84	123	126	21	17	12
Standard NN	3403	31	212	8	1268	4	1
Standard random hot deck	1047	185	1175	15	1295	4	142

5.4.3 Results for the rdm measure for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	-13.00	0.44	-0.01	0.44	0.08	12.96	-0.12
NN with weights	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Random hot deck without weights	1.47	1.41	0.00	0.45	0.08	11.36	1.25
Random hot deck with weights	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
Standard NN	1.95	-0.13	-0.27	-0.17	-0.75	0.10	-0.16
Standard random hot deck	-0.82	-0.64	-0.96	-0.81	-0.96	-0.74	1.92

5.4.4 Results for the *KS* measure for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	0.72	0.20	0.03	0.32	0.03	0.21	0.40
NN with weights	0.46	0.20	0.05	0.40	0.03	0.09	0.51
Random hot deck without weights	0.21	0.25	0.05	0.39	0.03	0.25	0.22
Random hot deck with weights	0.33	0.38	0.05	0.41	0.04	0.09	0.42
MI-SAS	0.16	0.25	0.12	0.17	0.03	0.13	0.12
Standard NN	0.02	0.03	0.04	0.05	0.02	0.02	0.02
Standard random hot deck	0.12	0.06	0.08	0.12	0.12	0.00	0.04

5.4.5 Results for medians for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	2.41	0.67	0.04	1.19	0.00	*	4.30
NN with weights	1.41	0.52	0.04	1.80	0.28	*	4.31
Random hot deck without weights	0.75	1.14	0.00	1.74	0.13	*	1.85
Random hot deck with weights	1.41	2.00	0.00	2.31	0.28	*	4.31
MI-SAS	0.00	0.03	0.00	1.86	0.13	*	0.41
Standard NN	0.98	1.03	3.64	1.60	3.64	*	0.00
Standard random hot deck	7.78	7,44	11.07	14.44	5.92	*	14.29

5.4.6 Results for standard deviations for data set 1

	R1	R2	R3	R4	R5	R6	R7
NN hot deck without weights	2.05	3.37	0.01	4.06	0.02	0.13	0.00
NN with weights	2.52	0.43	0.02	1.61	0.00	0.00	0.02
Random hot deck without weights	16.54	10.79	0.01	4.81	0.04	0.09	0.07
Random hot deck with weights	0.33	1.37	0.01	1.55	0.00	0.00	0.01
MI-SAS	0.14	3.28	0.11	4.07	0.00	0.10	0.01
Standard NN	2.99	0.30	0.16	0.16	0.75	0.00	0.00
Standard random hot deck	0.01	0.89	0.37	0.28	0.76	0.00	0.00

By simply counting per imputation method how often a variable is best (worst) for a certain evaluation measure, we find that NN hot deck without weights performs best 6 times (worst 5 times), NN hot deck with weights best 20 times (worst 1 time), random hot deck without weights best 2 times (worst 9 times), random hot deck with weights best 21 times (worst 2 times), MI-SAS best 2 times (worst 5 times), standard NN best 10 times (worst 1 time), and standard random hot deck best 4 times. This comparison is not completely fair, however, as NN hot deck with weights and random hot deck with weights have been designed to give perfect results for evaluation measures m_1 and *rdm*. Neglecting the results for these evaluation measures, we find that NN hot deck without weights is best 6 times (worst 3 times), NN hot deck with weights best 6 times (worst 1 time), random hot deck without weights best 2 times (worst 7 times), random hot deck with weights best 5 times (worst 2 times), MI-SAS best 4 times (worst 5 times), standard NN best 10 times (worst 0 times), and standard random hot deck best 4 times (worst 12).From this comparison we conclude that NN hot deck with weights and standard NN appear to perform best for data set 1, and random hot deck without weights and standard hot deck worst.

NN hot deck with weights performs especially well with respect to evaluation measures d_{L1}, m_1 and *rdm*, which all measure how well totals and individual values are preserved. Standard NN performs especially well for evaluation measure *KS* and the standard deviation, which both measure how well the statistical distribution is preserved. Random hot deck without weights and standard hot deck perform worst for measures d_{L1}, m_1, *rdm*, *KS* and the standard deviation, indicating that these methods are not good in preserving either individual values or the statistical distribution. MI-SAS performs with respect to the preservation of medians, whereas standard hot deck performs clearly worst.

Comparing the “unweighted” versions of our imputation methods to the “weighted” versions with respect to evaluation measures d_{L1} (Table 5.4.1), *KS* (Table 5.4.4), preservation of the median (Table 5.4.5) and preservation of the standard deviation (Table 5.4.6) – the “weighted” versions by design perform better on m_1 and *rdm* (Tables 5.4.2 and 5.4.3) – we see that the “weighted” versions generally perform better than the “unweighted” versions. An exception is random hot deck with respect to the preservation of the median, where the “unweighted” version seems to perform better.

Comparing NN with weights and random hot deck with weights to their standard versions standard NN respectively standard random hot deck, we see that taking edits and known totals into account leads to better preservation of totals and individual values (see Tables .5.4.1 to 5.4.3). However, taking edits and known totals into account leads to worse results without respect to preservation of the statistical distribution (see Tables 5.4.4 and 5.4.6). With respect to preservation of medians (Table 5.4.5) results are mixed for nearest-neighbour imputation: for some variables NN with weights performs better, for other variables standard NN.

For the complicated data set 2 our methods indeed succeeded in finding imputations that satisfy all edits and totals. The statistical quality of the imputations found by our imputation methods for that data set was rather low, however. This can be seen In Table 5.4.7 where we compare NN hot deck with weights to MI-SAS. “overall” in this table indicates the unweighted average over all 10 variables in the data set. “# better” indicates for how many variables one method performed better than the other method. For evaluation measures m_1 and *rdm* NN hot deck 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 hot deck with weights has m_1 and *rdm* equal to zero for all variables.

We conclude that, although our imputation methods managed to find imputations that were consistent with the edits and totals for data set 2, they did not find good imputations from a statistical point for view. For very complex data, i.e. data with many missing values or a restrictive set of edits, more advanced imputation methods appear to be required.

5.4.7 Comparison of NN hot deck with weights to MI-SAS for data set 2

		NN hot deck with weights	MI-SAS
d_{L1}	overall	1930	1261
	# better	0	10
m_1	overall	0.0	637.5
	# better	8	0
rdm	overall	0.00	0.01
	# better	8	0
KS	overall	0.10	0.06
	# better	4	6
median	overall	17.25	11.43
	# better	5	5
standard deviation	overall	163.79	25.93
	# better	4	6

Our imputation methods have been designed to satisfy edits and preserve known (weighted) totals. As can be seen from evaluation measures m_1 and *rdm* in Tables 5.4.2 and 5.4.3 the weighted versions of our imputation indeed succeeds in taking the weighted totals into account. Our imputation methods also succeed in satisfying edits. In none of the records of the two data sets we examined edits were violated. In this sense our imputation methods do exactly what they were designed for.

The MI-SAS procedure has not been designed to take edits and known totals into account. So, besides violating known totals (and hence means), which can be seen from evaluation measures m_1 and *rdm* in Tables 5.4.2 and 5.4.3, the MI-SAS procedure also leads to violated edits. The number of violated edits and violated records, i.e. records in which at least one edit is violated, is given in Table 5.4.8 below. In this table we see that especially standard NN and standard random hot deck 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.

5.4.8 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 hot deck without weights	0	0	0	0
NN hot deck with weights	0	0	0	0
Random hot deck without weights	0	0	0	0
Random hot deck with weights	0	0	0	0
MI-SAS	123	77 (2.5%)	25	25 (5.0%)
Standard NN	427	339 (10.9%)	1,212	488 (97.6%)
Standard random hot deck	394	292 (9.4%)	1,278	490 (98.0%)

6. Discussion

In this paper we have extended standard hot deck imputation methods so that the imputed data satisfy specified edits and preserve known totals, while taking survey weights into account. The hot deck imputation methods we have considered are random hot deck imputation and nearest neighbour hot deck imputation. 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. Novel aspects of our imputation methods are that they are able to take survey weights into account, and especially their simplicity, which may make them attractive from a practical point of view.

In our evaluation study we have used only two evaluation data sets. More evaluations studies on more 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.

In particular NN hot deck with weights 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 standard MI-SAS routine.

The preservation of the median might possibly be improved by extending the imputation procedure. The checks for unweighted totals (20), respectively for weighted totals (21), give an admissible interval for each value to be imputed. Any value between the lower and upper bound is acceptable. While imputing values, one might be able to utilize this freedom to preserve the median more accurately, assuming that the median of the observed data is a good approximation of the true median. When the median of the imputed so far data becomes too small (i.e. smaller than the observed median), one could "push" the median in the right direction by subsequently imputing some values larger than the observed median. Conversely when the median of the data imputed so far becomes too large, one could "push" the median in the right direction by subsequently imputing some values smaller than the observed median. We have not yet extended our imputation methods in this way and do not know to which extent this approach will improve the preservation of the median or to which extent the results for the other evaluation measures may be affected.

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 does lead to a deterioration of the preservation of the statistical distribution.

Since the evaluation data set that was used in Pannekoek, Shlomo and De Waal (2013) was no longer available to us, we did not compare the evaluation results of the imputation methods proposed in the current paper to the results of the imputation methods by Pannekoek, Shlomo and De Waal (2013). Since the imputation methods proposed in the current paper are based on simple hot deck techniques, we expect that the more advanced imputation methods by Pannekoek, Shlomo and De Waal (2013) are better from a statistical point of view. Whether this really is the case still has to be confirmed, though.

In this paper our goal was to develop simple imputation methods that satisfy edits, preserve totals and take survey weights into account. For the sake of simplicity, we decided to base our imputation methods on simple hot deck techniques. As we already mentioned in the Introduction our imputation methods can be modified in a straightforward way by using, for instance, regression imputation with a random residual or predictive mean matching. The only modification is that in Step 1 of our approach (see Section 4.4), regression imputation with a random residual or predictive mean matching instead of a hot deck technique should be used. The use of regression imputation with a random residual or predictive mean matching is (slightly) more complicated than the use of a hot deck technique as regression models have to be specified for all variables to be imputed and their model parameters estimated.

The imputation methods developed in this paper may distort correlations between variables. An important topic for future research is the development of imputation that preserve correlations, preserve totals and satisfy edits.

Our imputation methods were not designed to estimate the correct variance, including the variance due to imputation. In a future paper we hope to extend our methods to multiple imputation (see Rubin 1987). One way to do this would be to replace hot deck imputation by predictive mean matching (see, e.g., De Waal, Pannekoek and Scholtus 2011 and Vink 2015).

The 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. One way to overcome this drawback is by first imputing all missing data simultaneously without taking edit rules or known totals into account, and then adjust the imputed data so all edit rules are satisfied and totals preserved. This approach has the disadvantage that the adjustment step may have unfavourable effects on the imputation. A more principled way of overcoming the drawback of a sequential approach would be to use an imputation method that imputes all variables in each record simultaneously while taking edits, totals and survey weights into account. However, simultaneous imputation methods that satisfy all edits, preserve known totals and can take survey weights into account seem exceedingly hard to develop. 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.

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/).

Coutinho, W. and T. de Waal (2012), Hot Deck Imputation of Numerical Data under Edit Restrictions, Discussion paper 201223, Statistics Netherlands.

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.

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.

Fellegi, I.P. and D. Holt (1976), A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association 71, pp. 17-35.

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

Knottnerus, P. (2003), Sample Survey Theory. Springer-Verlag, New York.

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. 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.

Pannekoek, J. and T. de Waal (2005), Automatic Edit and Imputation for Business Surveys: the Dutch Contribution to the EUREDIT Project. Journal of Official Statistics 21, pp. 257-286.

Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. Annals of Applied Statistics 7, pp. 1983-2006.

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.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.

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.

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.

Vink, G. (2015), Restrictive Imputation of Incomplete Survey Data. PhD thesis, Utrecht University

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.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/’15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/’13–2014/’15	Crop year, financial year, etc., 2012/’13 to 2014/’15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information