Preserving Logical Relations while Estimating Missing Values

Ton de Waal (t.dewaal@cbs.nl)
Statistics Netherlands, & Tilburg University

Wieger Coutinho
Statistics Netherlands

ABSTRACT

Item-nonresponse is often treated by means of an imputation technique. In some cases, the data have to satisfy certain constraints, which are frequently referred to as edits. An example of an edit for numerical data is that the profit of an enterprise equals its turnover minus its costs. Edits place restrictions on the imputations that are allowed and hence complicate the imputation process. In this paper we explore an adjustment approach. This adjustment approach consists of three steps. In the first step, the imputation step, nearest neighbour hot deck imputation is used to find several pre-imputed values. In a second step, the adjustment step, these pre-imputed values are adjusted so the resulting records satisfy all edits. In a third step, the best donor record is selected. The adjusted record corresponding to that donor record is the final imputed record. In principle, a potential donor that is not the closest to the record to be imputed may still give the best results after adjustment. In this paper we therefore focus on the number of potential donor records that are considered in the imputation step.

Keywords: *Nearest-neighbour Imputation, Edit restrictions, Linear programming, Data adjustment*

JEL classification: *C13, C14, C61, C81, C83*

INTRODUCTION

Item-nonresponse is a frequently occurring problem in survey data. An often used approach to treat missing data due to item-nonresponse is imputation, where individual values are estimated and filled into the data set.

In some cases, the data have to satisfy certain constraints, which are often referred to as edit rules or edits for short. Examples of such edits for numerical data 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. These edits place restrictions on the imputations that are allowed and hence complicate the imputation process.

In this paper we describe an imputation approach that satisfies the edits. We will refer to this approach as adjusted nearest-neighbour imputation.

The remainder of this paper is organized as follows. Section 2 gives a literature review. Section 3 describes the edits we consider in this paper and the proposed imputation approach. Section 4 gives the results of several variations of the imputa-

tion approach and discusses these results. Section 5 ends the paper by drawing some conclusions and identifying some possible topics for future research.

LITERATURE REVIEW

Many imputation methods have been developed for many different situations and kinds of data sets, and are discussed in a large number of articles and books, such as Andridge and Little (2010), Kalton and Kasprzyk (1986), Rubin (1987), Schafer (1997), Little and Rubin (2002), De Waal, Pannekoek and Scholtus (2011) and Van Buuren (2012).

In particular several imputation methods satisfying edits have been developed, see Geweke (1991), Raghunathan, Solenberger and Van Hoewyk, (2002), Tempelman (2007), Coutinho, De Waal and Remmerswaal (2011), Coutinho and De Waal (2012), Pannekoek, Shlomo and De Waal (2013), Kim et al. (2014) and De Waal, Coutinho and Shlomo (forthcoming). These methods are generally quite complicated, and are based either on sequentially imputing the variables with missing values or on somehow truncating a statistical distribution, such as the multivariate normal, to the region defined by the edits. We note that, besides satisfying edits, the methods proposed by Pannekoek, Shlomo and De Waal (2013 and De Waal, Coutinho and Shlomo (forthcoming) also preserve known or previously estimated totals.

Pannekoek and Zhang (2015) proposed a much simpler adjustment approach for satisfying edits, consisting of two steps. In the first step, the imputation step, nearest neighbour hot deck imputation is used to find a donor record that is closest to the record to be imputed. Missing values in the record to be imputed are pre-imputed with values from that donor record. In a second step, the adjustment step, these pre-imputed values are adjusted so the resulting record satisfies all edits.

In the current paper we explore this approach in more detail. In particular, we will focus on the number of potential donor records that are considered in the imputation step. In principle, a potential donor that is not the closest to the record to be imputed may still give the best results after adjustment. Whereas Pannekoek and Zhang (2015) considered only the closest donor record to the record to be imputed in the imputation step, we will consider the k closest records, where $k = 1, 2, 5$ or 10. We will also provide more evaluation results than Pannekoek and Zhang (2015).

METHODOLOGY

Before we describe the proposed methodology, we first need to describe the kind of edits that we consider as this clarifies the problem we are trying to solve. This is the topic of subsection 3.1 below. Our adjusted nearest-neighbour method itself is described in subsection 3.2.

Edits

In this paper we consider edits for numerical data. Such edits are generally either linear equations or linear inequalities. In other words, an edit can be written as:

$$a_{1e}x_1 + \cdots + a_{ne}x_n + b_e = 0 \quad (1a)$$

or as

$$a_{1e}x_1 + \cdots + a_{ne}x_n + b_e \geq 0 \quad (1b)$$

Here the a_{je} and the b_e are constants and the x_j $(j = 1, \ldots, n)$ are the variables.

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

$$T - C - P = 0 \quad (2)$$

where *T* is the turnover of an enterprise, *P* its profit, and *C* 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 is

$$T \geq 0$$

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

Adjusted nearest-neighbour imputation

We start by describing the idea of our adjusted nearest-neighbour imputation method. For each record to be imputed our method consists of 3 steps. In the first step we use observed values from potential donor records to construct a pre-imputed record for each potential donor record. These pre-imputed records may, and often will, violate some of the edits. In a second step we therefore adjust the imputed values in these pre-imputed records so these adjusted records satisfy all edits. In a third step we select the donor record for which the distance of the donor record to the record to be imputed plus the distance of the pre-imputed record to the adjusted record is minimal. The adjusted record corresponding to that donor record is our final imputed record.

Since using all potential donor records may be too time-consuming, in practice we restrict ourselves to the *k* nearest-neighbour donor records instead of all potential donor records.

We will now describe the above idea in mathematical terms. Let us suppose that we want to impute a certain record $\boldsymbol{x}_{r_0}$, the recipient record. Let $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ denote the record that would be obtained if the missing values in record $\boldsymbol{x}_{r_0}$ were imputed with the corresponding (observed) values in record $\boldsymbol{x}_r$. Also, let $\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ denote the record after adjusting the record $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ so the adjusted record satisfies the edit restrictions

We aim to find the donor record that we will use to impute the recipient by minimizing the distance of the donor record $\boldsymbol{x}_r$ to the recipient record $\boldsymbol{x}_{r_0}$ plus the distance of $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ to $\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$. That is, we solve the following problem

$$\min_{\boldsymbol{x}_r \in D(\boldsymbol{x}_{r_0})} \left[d_{\text{obs}(\boldsymbol{x}_{r_0})}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r) + d_{\text{mis}(\boldsymbol{x}_{r_0})}(\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r), \boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)) \right] \quad (3)$$

where $D(\boldsymbol{x}_{r_0})$ is the set of potential donor records for $\boldsymbol{x}_{r_0}$, (the so-called donor pool), $d_{\text{obs}(\boldsymbol{x}_{r_0})}$ the distance function used to measure the distance between records restricted to the variables observed in $\boldsymbol{x}_{r_0}$, and $d_{\text{mis}(\boldsymbol{x}_{r_0})}$ the distance function used restricted to the variables with missing values in $\boldsymbol{x}_{r_0}$.

$\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ itself is obtained by solving a constrained optimization problem. Let

$$\boldsymbol{x}_{\text{opt}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r) = \underset{\boldsymbol{x}}{\text{argmin}} \left[d_{\text{mis}(\boldsymbol{x}_{r_0})}\left(\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r), \boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x})\right) | \boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}) \in A \right]$$

where A denotes the set of records that satisfy all edit restrictions. x is not restricted to be a record in the donor pool, but may be any (observed or unobserved) record satisfying all edit restrictions.

The adjusted record is then given by

$$\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r) \equiv \boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_{\text{opt}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)),$$

In words, the record $\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ is obtained by adjusting the donor values from record $\boldsymbol{x}_r$ to values from a synthetic record $\boldsymbol{x}_{\text{opt}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_r)$ so that the adjusted donor values together with the observed values in $\boldsymbol{x}_{r_0}$ satisfy all edits.

As we already mentioned at the beginning of this subsection, solving (3) to optimality is in many cases quite time-consuming since all potential donor records have to be considered. Instead of solving the above problem to optimality we will instead apply a less time-consuming approach. This approach is likely to find the optimal solution for most records in any case. In our approach we first select the k nearest-neighbours of $\boldsymbol{x}_{r_0}$ according to the distance function d, say $\boldsymbol{x}_i^* \ (i = 1, \ldots, k)$. For each $\boldsymbol{x}_i^*$ we then solve

$$\boldsymbol{x}_{\text{opt},i}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*) = \underset{\boldsymbol{x}_a}{\text{argmin}} \left[d_{\text{mis}(\boldsymbol{x}_{r_0})}\left(\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*), \boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_a)\right) | \boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_a) \in A \right] \quad (4)$$

Next, we calculate

$$d_{\text{obs}(\boldsymbol{x}_{r_0})}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*) + d_{\text{mis}(\boldsymbol{x}_{r_0})}\left(\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*), \boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*)\right), \quad (5)$$

and select the record $\boldsymbol{x}_i^* \ (i = 1, \ldots, k)$ for which (5) is the smallest. Our imputed record is then given by $\boldsymbol{x}_{\text{adj}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_i^*)$. This heuristic returns the optimal solution to (3), unless the optimal donor record is not among the k nearest-neighbours of $\boldsymbol{x}_{r_0}$.

In this paper we will use the sum of absolute differences to measure the distance between two records $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. i.e. we will use

$$d_S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j \in S} u_j |x_{1j} - x_{2j}| \quad (6)$$

with S the set of variables that are used to calculate the distance function, and u_j a weight for the j-th variable indicating the importance of a change in that variable.

Substituting (6) into (4), we see that in step 2 of our approach, i.e. the adjustment step, we need, for each $\boldsymbol{x}_i^* \ (i = 1, \ldots, k)$, to solve a linear programming problem with objective function given by

$$\sum_{j \in \text{mis}(\boldsymbol{x}_{r_0})} u_j |x_{ij}^* - x_{aj}| \quad (7)$$

subject to the constraint that record $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_0}, \boldsymbol{x}_a)$ satisfies all edits. This can be formulated as a linear programming problem and can, for instance, be solved by the well-known simplex algorithm.

A technical problem is that many implementations of the simplex algorithm cannot minimize a sum of absolute distances directly. We overcome this technical problem by introducing variables λ_j ($j \in \text{mis}(\boldsymbol{x}_{r_n})$) that have to satisfy

$$\lambda_j \geq x_{ij}^* - x_{aj} \text{ for } j \in \text{mis}(\boldsymbol{x}_{r_n}) \tag{8}$$

$$\lambda_j \geq x_{aj} - x_{ij}^* \text{ for } j \in \text{mis}(\boldsymbol{x}_{r_n}) \tag{9}$$

Our minimization problem is now given by:

Minimize $\sum_{j \in \text{mis}(x_{r_0})} u_j \lambda_j$ (10)

subject to (8), (9) and the constraint that $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_n}, \boldsymbol{x}_a)$ satisfies all edits. Since in an optimal solution $\lambda_j = x_{ij}^* - x_{aj}$ or $\lambda_i = x_{aj} - x_{ij}^*$ (for $j \in \text{mis}(\boldsymbol{x}_{r_0})$), we have that in an optimal solution $\lambda_j = |x_{ij}^* - x_{aj}|$ ($j \in \text{mis}(\boldsymbol{x}_{r_0})$). Hence, minimizing (10) subject to (8), (9) and the constraint that $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_n}, \boldsymbol{x}_a)$ satisfies all edits leads to the same solution for $\boldsymbol{x}_a$ as minimizing (7) subject to the constraint that $\boldsymbol{x}_{\text{hot}}(\boldsymbol{x}_{r_n}, \boldsymbol{x}_a)$ satisfies al edits.

We illustrate our imputation method by means of the example below.

Example

Suppose there are four variables, T (turnover), P (profit), C (costs), and N (number of employees in fulltime equivalents), and that the edits are given by

$$T = P + C \tag{11}$$
$$P \leq 0.5T \tag{12}$$
$$0.1T \leq P \tag{13}$$
$$T \leq 550N \tag{14}$$
$$T \geq 0 \tag{15}$$
$$N \geq 0 \tag{16}$$
$$C \geq 0 \tag{17}$$

Let us suppose that the weight of variable N in (6) equals 500 and of the other three variables 1. Suppose furthermore that in a certain record $N = 5, T = 2000$ and the values of P and C are missing.

Now suppose that in one of the k nearest (potential) donor records the following values are observed: $N^* = 6, T^* = 2200\ P^* = 900$ and $C^* = 1300$. We then find adjusted values $\hat{P}$ and $\hat{C}$ such that $\hat{P}$ and $\hat{C}$ together with the observed values $N = 5$ and $T = 2000$ satisfy (11) to (17) and

$$|\hat{P} - 900| + |\hat{C} - 1300|$$

is minimized. The problem can easily be formulated as a linear programming problem, which can, for instance, be solved by means of the simplex algorithm. A solution is $\hat{P} = 900$ and $\hat{C} = 1100$.

The total distance (6) is then given by

$$500|5 - 6| + |2000 - 2000| + |900 - 900| + |1300 - 1100| = 700$$

Likewise we calculate this distance for the other $(k-1)$ nearest (potential) donor records. From all k imputed records we select the record with the smallest value for (5).

RESULTS AND DISCUSSION

Evaluation approach

In our evaluation study we have used two data sets. The true values for the data in the two data sets are known. In the completely observed data sets values were deleted, using a Missing At Random (MAR) mechanism. When the missing data mechanism is MAR, there is a relation between the missing data pattern and the values of the observed data, but not between the missing data pattern and the values of the missing data.

For each of our evaluation data sets we thus have two versions available: a version with missing values and a version with complete records. The former version is imputed, without making any use of the complete records. The resulting data set is then compared to the version with complete records.

Methods evaluated

In our evaluation study we have used a weighted and an unweighted version of our adjusted nearest-neighbour approach. In the weighted version the weight u_j for the j-th variable in (6) is set to the reciprocal of the observed mean for this variable in the data set with missing data. In the unweighted version the weight u_j is set to 1 for all variables. We will denote our imputation methods by W 1, W 2, W 5, W 10, NW 1, NW 2, NW 5 and NW 10, where "W" indicates a weighted version and "NW" an unweighted version, and the number indicates the number of nearest-neighbours considered. For instance, NW 5 denotes the imputation method where all weights u_j in (6) have been set to 1, and 5 nearest-neighbours are considered.

We have also examined taking the average over all 8 imputations methods. We consider this as a ninth method, and will be denoted by Mixed. By taking the average over the above-mentioned 8 imputation methods, we have constructed a kind of implicit fractional imputation method (see e.g. Kim and Fuller 2004 and Kim 2011). In fractional imputation, an imputed value is actually a weighted sum of several imputed values, for instance obtained from several donor records. In fractional imputation explicit weights are used. In our case, the weighting is implicit, and depends on how often the same donor value is used to impute a record.

To evaluate the results of our imputation methods we have compared them to nearest-neighbour hot deck imputation, using the sum of the absolute differences (6) as distance function, and random hot deck imputation. We will refer to nearest-neighbour hot deck imputation as NN HD and to random hot deck imputation as Random HD.

Evaluation data

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

The characteristics of the evaluation data sets

Table 1

	data set 1	data set 2
Total number of records	3096	500
Number of records with missing values	287	250
Total number of variables	5	6
Total number of edits	14	17
Number of balance edits	0	1
Total number of inequality edits	14	16
Number of non-negativity edits	5	6

Tables 2 and 3 give the numbers of missing values and (unweighted) means of the variables of our data sets. In brackets 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 is given. The means 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. This variable is only used as auxiliary variable.

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

Table 2

Variable	Number of missing values	Mean
R_1	0 (0.0 %)	37.4
R_2	79 (2.2%)	777.6
R_3	76 (4.2%)	11574.8
R_4	73 (4.8%)	209.9
R_5	67 (2.6%)	169.2

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

Table 3

Variable	Number of missing values	Mean
S_1	61 (12.2%)	97.8
S_2	86 (17.2%)	175018.3
S_3	131 (26.2%)	731.0
S_4	61 (12.2%)	175749.3
S_5	109 (21.8%)	154286.5
S_6	91 (18.2%)	7522.3

When method W 1, W 2, W 5, W 10, NW 1, NW 2, NW 5 or NW 10 is used, sometimes values from a donor record are used directly whereas in other cases these values are adjusted as a result of solving a linear programming problem. Table 4 below reports how many times donor values were used directly (see the columns “Donor”) and how many times donor values were adjusted (see the columns “Adjusted”). As could be expected, the number of times donor values were used directly increases as the number of nearest-neighbour donors considered increases.

Number of times donor respectively simplex is used

Table 4

	Data set 1		Data set 2	
	Donor	Simplex	Donor	Simplex
W 1	266	21	54	196
W 2	273	14	60	190
W 5	280	7	60	190
W 10	281	6	60	190
NW 1	266	21	39	211
NW 2	276	11	52	198
NW 5	279	8	57	193
NW 10	282	5	60	190

Evaluation measures

In our evaluation study we focus on three measures proposed by Chambers (2003). Each of these measures examines a different aspect, namely the preservation of individual values, the preservation of totals and means, and the preservation of univariate statistical distributions.

The preservation of individual values is measured by the d_{L1} measure. The d_{L1} measure for variable x_j is 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, x_{ij}^{true} the corresponding true value, $M(j)$ the set of records for which the value on variable x_j is missing, and w_i is the survey weight of record i. This measure calculates the average distance between the imputed and true values.

The preservation of totals and means is measured by the m_1 measure, which 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 m_1 measure calculates the preservation of the first moment of the empirical distribution of the true values.

The preservation of univariate distributions is measured by the KS Kolmogorov-Smirnov distance. For weighted data, the empirical distribution of the true values is defined as

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

with $|M(j)|$ the number of records with missing values for the variable x_j under consideration 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. The KS compares the empirical distribution of the original values to the empirical distribution of the imputed values.

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

We have also compared the correlations in the (partly) imputed data to the correlations in the true data in order to evaluate to what extent the relationships between different variables are preserved. In particular, we have calculated the average absolute difference between the correlations in the true complete data and in the imputed data, where we have taken the average over all 10 pairs of variables for data set 1 and all 15 pairs for data set 2. We have also calculated the average of the absolute percentage differences, where the percentage is calculated with respect to the correlations in the complete data over all pairs of variables in each of the data sets.

Evaluation results

The evaluation results for data set 1 are presented in Tables 5 to 7. As variable R_1 has no missing values it is not included in Tables 4 to 6. "Average" is the average of the absolute results over all 4 variables mentioned in these tables.

Results for d_{L1} for data set 1

Table 5

	R_2	R_3	R_4	R_5	Average
W 1	689	5803	72	20	1646
W 2	689	5793	72	20	1644
W 5	694	3967	315	19	1249
W 10	694	3.967	315	19	1249
NW 1	767	3843	315	22	1236
NW 2	702	3967	315	19	1250
NW 5	694	3967	315	19	1249
NW 10	694	3967	315	19	1249
Mixed	625	4532	185	20	1340
NN HD	690	5803	79	28	1650
Random HD	1140	8080	81	26	2331

Results for m_1 for data set 1

Table 6

	R_2	R_3	R_4	R_5	Average
W 1	93	865	32	18	253
W 2	98	936	33	18	271
W 5	96	857	259	19	308
W 10	96	857	259	19	308
NW 1	38	1126	258	17	360
NW 2	89	857	259	19	306
NW 5	96	857	259	19	308
NW 10	96	857	259	19	308
Mixed	8	3	153	18	46
NN HD	93	865	26	10	249
Random HD	182	432	9	12	159

Results for KS for data set 1

Table 7

	R_2	R_3	R_4	R_5	Average
W 1	0.10	0.20	0.19	0.92	0.35
W 2	0.10	0.20	0.19	0.98	0.36
W 5	0.18	0.10	0.46	0.97	0.43
W 10	0.18	0.10	0.46	0.97	0.43
NW 1	0.16	0.08	0.46	0.67	0.35
NW 2	0.18	0.10	0.46	0.97	0.43
NW 5	0.18	0.10	0.46	0.97	0.43
NW 10	0.18	0.10	0.46	0.97	0.43
Mixed	0.08	0.15	0.35	0.81	0.35
NN HD	0.10	0.20	0.21	0.92	0.36
Random HD	0.32	0.51	0.47	0.92	0.55

The evaluation results for data set 2 are presented in Tables 8 to 10. "Average" is the average of the absolute results over all 6 variables mentioned in these tables.

Results for d_{L1} for data set 2

Table 8

	S_1	S_2	S_3	S_4	S_5	S_6	Average
W 1	28	13549	537	19470	39322	2132	12507
W 2	32	13584	477	19455	42757	2033	13056
W 5	32	13855	471	19791	44222	2092	13410
W 10	32	9975	504	14347	43081	2403	11724
NW 1	31	11329	400	16133	17668	2617	8030
NW 2	31	7579	442	10968	14943	2703	6111
NW 5	32	7814	450	11229	16718	2703	6491
NW 10	28	11718	430	16707	20294	2692	8645
Mixed	23	8988	342	12775	22272	1909	7718
NN HD	31	52319	553	62618	57207	2316	29174
Random HD	36	118355	548	110335	102786	3338	55899

Results for m_1 for data set 2

Table 9

	S_1	S_2	S_3	S_4	S_5	S_6	Average
W 1	4	8049	153	11677	9134	1135	5025
W 2	4	3107	150	4704	19932	831	4788
W 5	4	10000	100	14312	15037	850	6717
W 10	4	7468	202	10961	19045	1269	6491
NW 1	4	1790	24	2576	600	1526	1087
NW 2	1	1515	134	2425	75	1490	940
NW 5	3	4899	24	6959	4144	1652	2947
NW 10	2	122	110	407	2460	1652	792
Mixed	1	4619	112	6753	7593	1270	3391
NN HD	7	2381	58	17754	8751	951	4983
Random HD	8	156	32	15369	6.044	495	3684

Results for KS for data set 2

Table 10

	S_1	S_2	S_3	S_4	S_5	S_6	Average
W 1	0.03	0.02	0.07	0.02	0.03	0.05	0.04
W 2	0.04	0.03	0.04	0.03	0.07	0.03	0.04
W 5	0.03	0.01	0.08	0.02	0.07	0.03	0.04
W 10	0.02	0.02	0.07	0.02	0.02	0.05	0.03
NW 1	0.05	0.02	0.07	0.02	0.02	0.03	0.04
NW 2	0.03	0.02	0.06	0.02	0.02	0.02	0.03
NW 5	0.04	0.01	0.11	0.02	0.01	0.05	0.04
NW 10	0.03	0.02	0.03	0.02	0.02	0.03	0.02
Mixed	0.02	0.02	0.03	0.02	0.03	0.03	0.02
NN HD	0.04	0.03	0.08	0.03	0.06	0.05	0.05
Random HD	0.04	0.09	0.05	0.06	0.06	0.03	0.05

Examining the results for methods W 1, W2, W 5, W 10, NW 1, NW 2, NW 5 and NW 10, we see that they are rather erratic. For instance, W1 performs best of these methods on the m_1 measure for data set 1, but rather bad for data set 2. For data set 2, W 1 is outperformed by the standard methods NN HD and Random HD. NW 1 performs worst on that measure for data set 1, but best for data set 2.

A general result is that our imputation methods perform better on the d_{L1} measure than NN HD and Random HD.

The results for Mixed are much more stable than for our individual imputation methods. Mixed is among the best performing approaches for all evaluation measures and both data sets. Mixed also performs better than NN HD and Random HD in all examined cases.

Table 11 gives the average (over all pairs of variables) of the absolute deviation of the correlations in the imputed data and the correlations in the true, complete data. Between brackets the average (over all pairs of variables) of the absolute percentage differences is given.

Average absolute deviation from true correlations

Table 11

	Data set 1	Data set 2
W 1	0.0016 (0.50%)	0.0328 (13.00%)
W 2	0.0016 (0.46%)	0.0387 (18.48%)
W 5	0.0033 (0.98%)	0.0438 (20.23%)
W 10	0.0033 (0.98%)	0.0343 (14.90%)
NW 1	0.0034 (0.98%)	0.0381 (20.33%)
NW 2	0.0033 (0.98%)	0.0433 (24.14%)
NW 5	0.0033 (0.98%)	0.0417 (21.39%)
NW 10	0.0033 (0.98%)	0.0293 (14.86%)
Mixed	0.0021 (0.76%)	0.0360 (16.00%)
NN HD	0.0016 (0.49%)	0.0371 (10.72%)
Random HD	0.0035 (0.90%)	0.1457 (37.53%)

Mixed is again performing quite well, although it is slightly outperformed by NN HD for data set 1 on this measure.

By design our imputation methods do not violate any edits in any of the imputed records, including Mixed since an average of values satisfying linear edits

again satisfies these edits. The standard methods NN HD and Random HD do lead to violated edits and violated records, i.e. records in which one or more edits are violated. The number of edit violations for these methods are reported in Table 12 below.

Number of failed edits and records for the standard methods

Table 12

	Data set 1		Data set 2	
	Failed edits	Failed records	Failed edits	Failed records
NN HD	21	21	258	211
Random HD	55	55	250	206

Of the 258 failed edits for NN HD for data set 2, 189 times the balance edits was failed and 69 times an inequality edit. Of the 250 failed edits for Random HD for data set 2, 193 times the balance edits was failed and 57 times an inequality edit.
If edit violations are to be prevented, the use of NN HD or Random HD should therefore be avoided.

CONCLUSIONS

Considering the results on the evaluation measures, Mixed would be our recommended method of all methods examined in the current paper. The Mixed method does not lead to any violated edits, it leads to good results for all evaluation measures examined in this paper, and it performs better than NN HD and Random HD in almost all cases. Apparently, the use of edit restrictions pushes the imputations in the right direction for this method, in any case for the data sets considered in our evaluation study.

The imputations produced by Mixed are averages over 8 other imputation approaches. These 8 imputation methods show some rather erratic evaluation results. A method that performs quite well on a certain measure for one data set can perform rather badly on the same measure for the other data set. Apparently, taking the average reduces the erratic behavior of the results, and brings out the best of the 8 imputation methods.

Another advantage of Mixed is that it is a relatively simple method to develop and implement. The most complicated part is using the simplex algorithm, but for that several software packages, e.g. in R, are nowadays freely available. Mixed is considerably simpler than earlier imputation methods that preserve edits that were mentioned in the Introduction.

In the current paper we have not made an attempt to optimize the Mixed approach, i.e. we simply took the average over all other adjusted nearest-neighbour imputation methods we proposed and evaluated. An interesting topic for future research could be finding an optimal mix for the Mixed approach: should we combine only unweighted adjusted nearest-neighbour imputation methods, only weighted adjusted nearest-neighbour imputation methods, or a combination of both? In the latter two cases: what should these weights be? In this paper we took the reciprocal

of the average of the observed values for a variable as the weight for this variable, but other weights might possibly lead to better results. Finally, also the number of adjusted nearest-neighbour imputation methods should be decided upon as well as the optimal number of nearest neighbours for each of these methods.

A possible extension of Mixed would be to include the preservation of known or previously estimated totals in the imputation process. We leave such an extension to possible future research.

References

1. **Andridge, R.A.** and **R.J.A. Little**, 2010, *A Review of Hot Deck Imputation for Survey Non-response*. International Statistical Review *78*, pp. 40-64.
2. **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.
3. **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.
4. **De Waal, T., W. Coutinho** and *N. Shlomo* (forthcoming). *Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions*.
5. **Pannekoek, J.** and **L.C. Zhang**, 2015, *Optimal Adjustments for Inconsistency in Imputed Data*. Survey Methodology 41, pp. 127-144.
6. **De Waal, T., J. Pannekoek** and **S. Scholtus**, 2011, *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New York.
7. **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.
8. **Kalton, G.** and **D. Kasprzyk**, 1986, *The Treatment of Missing Survey Data*. Survey Methodology *12*, pp. 1-16.
9. **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.
10. **Kim, J.K.**, 2011, *Parametric Fractional Imputation for Missing Data Analysis*. Biometrika 98, pp. 119–132.
11. **Kim, J.K.** and **W. Fuller,** 2004, *Fractional Hot Deck Imputation*. Biometrika 91, pp. 559–578.
12. **Little, R.J.A.** and **D.B. Rubin,** 2002, *Statistical Analysis with Missing Data (second edition).* John Wiley & Sons, New York.
13. **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.
14. **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.
15. **Rubin, D.B.**, 1987, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.
16. **Schafer, J.L.**, 1997, *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
17. **Tempelman, C.,** 2007, *Imputation of Restricted Data*. Doctorate thesis, University of Groningen.
18. **Van Buuren, S.**, 2012, *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, Florida.